FACULTAD DE INGENIERÍA

Magíster en Gestión de Tecnología de Información y Telecomunicaciones

ESTABLECER UNA SECUENCIA EFECTIVA DE APRENDIZAJE A TRAVÉS DE UN MODELO SISTEMÁTICO DE BÚSQUEDA DE CONSTRUCCIÓN DE DATOS.

Tesis de postgrado para optar al título de MAGÍSTER EN GESTIÓN DE TECNOLOGÍA DE LA INFORMACIÓN Y TELECOMUNICACIONES

Miguel Ángel Contreras Mesa

César Llanos Sánchez

Erwin Vásquez Cariaga

Profesores Guías: Alejandro Figueroa Amenábar.

David Alfredo Ruete Zuñiga.

Santiago de Chile, 2021

DEDICATORIA

Dedicamos este trabajo a nuestras Familias

por su apoyo incondicional y darnos motivación

y la confianza para continuar nuestra formación académica.

Tenemos la dicha de poder decir que hay a nuestro lado gente maravillosa que nos apoyó

sin nosotros decir una sola palabra por eso y más, gracias.

Miguel, César y Erwin.

AGRADECIMIENTO

Agradecemos de manera especial a nuestro profesor guía, el Sr. Alejandro Gastón Figueroa Amenábar, quien nos acompañó y apoyó en el desarrollo de la tesis; a la Universidad Andrés Bello y a todos los docentes que nos entregaron con la mejor disposición todos aquellos conocimientos que nos hacen y harán marcar la diferencia en nuestra formación académica y profesional de aquí en adelante.

Además, agradecer a nuestras Sras. Eliana, Gleenda y Carol, quienes nos han entregado todo el cariño y apoyo para culminar con éxito este escalón más de nuestras vidas. Este logro es gracias a ustedes.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

INDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ABSTRACT

In this thesis, we have established an effective learning sequence through a systematic search model of data construction using the Yahoo! Answers in the prediction context of gender and age.

To obtain this sequence, the training data was segmented according to the gender and year of birth of each user in batches. To select the best batch, a greedy algorithm was implemented considering the highest value of the macro-average metric.

As a result, an effective learning sequence or curriculum was obtained, which was verified by training a Bayes model. To check the learning curriculum, 2 experiments were carried out, in the first experiment the model was trained without the curriculum, that is, with all the data of the users (329.025 examples) and in the second experiment the model was trained with the curriculum 11.176 examples). The final result was that both models had the same prediction performance, however, in the second experiment less data was needed, so their training was faster and more efficient, thus checking the effectiveness of the curriculum.

Key words: community question answering, age prediction, gender prediction.

RESUMEN

En esta tesis, hemos establecido una secuencia efectiva de aprendizaje a través de un modelo sistemático de búsqueda de construcción de datos mediante la plataforma de Yahoo! Answers en el contexto de predicción de género y edad.

Para obtener esta secuencia, se segmentaron los datos de entrenamiento según el género y año de nacimiento de cada usuario en lotes. Para seleccionar el mejor lote se implementó un algoritmo greedy considerando el mayor valor de la métrica macropromedio.

Como resultado se obtuvo un curriculum o secuencia efectiva de aprendizaje, la cual fue verificada mediante el entrenamiento de un modelo de Bayes. Para comprobar el curriculum de aprendizaje se realizaron 2 experimentos, en el primer experimento se entrenó el modelo sin el curriculum, es decir, con todos los datos de los usuarios (329.025) ejemplos y en el segundo experimento se entrenó el modelo con el curriculum (11.176 ejemplos). El resultado final fue que ambos modelos tuvieron el mismo rendimiento de predicción, sin embargo, en el segundo experimento se necesitaron menos datos, por lo que su entrenamiento fue más rápido y eficiente, comprobando de esta manera la efectividad del curriculum.

Palabras Claves: respuesta a preguntas de la comunidad, predicción de edad, predicción de género.

1 DESCRIPCIÓN DEL PROYECTO

1.1 Introducción

En la actualidad nos encontramos en una era digital que gira en torno a las nuevas tecnologías, trayendo consigo cambios profundos y transformaciones de una sociedad que se mueve en un mundo globalizado, produciendo la necesidad de estar activamente interactuando en los canales digitales.

La cantidad de datos creados en todo el mundo en 2018 alcanzó los 33 zettabytes (un zettabyte equivale a 1.000 millones de terabytes), 16,5 veces más que solo hace nueve años. No obstante, gracias a los nuevos desarrollos tecnológicos, como el internet de las cosas, se estima que la cantidad de información digital generada en 2035 ascienda a los 2.142 zettabytes. **(Statist Digital Economy Compass, 2019)**

En tal contexto, la interacción y flujo constante de información entre los usuarios a través de las plataformas, crea oportunidades para estudiar y analizar el comportamiento de estos en el mundo digital, alcanzando a través de soluciones tecnológicas identificar perfiles y tendencias de los usuarios. una de estas soluciones es el aprendizaje automático o más conocido en inglés como Machine Learning (ML). Hemos visto en la Inteligencia Artificial (IA) la capacidad de aprender tareas de forma automática. Por ejemplo, los filtros de spam de correo electrónico utilizan un algoritmo que consideran reglas definidas previamente con el fin de detectar qué mensajes son correo basura y separarlos de aquellos que no lo son.

Dado lo anterior, nuestra tesis busca resolver la hipótesis planteada en base a la investigación de la funcionalidad del aprendizaje automático mediante la respuesta predictiva de los datos género/edad, considerando como fuente de datos la página web de Yahoo! Answers.

1.2 Yahoo! Answers.

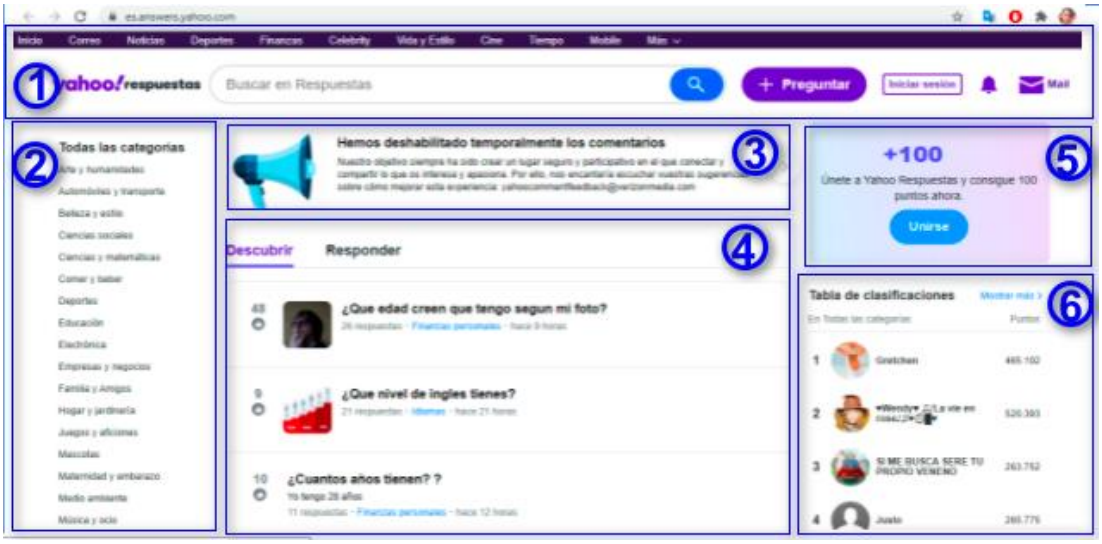
Yahoo! Answers es un sitio web de preguntas y respuestas (a.k.a community question answering; cQA) impulsado por la comunidad, o mercado del conocimiento, de Yahoo! **(Yahoo, 2021)**, que permite a los usuarios interactuar con base en el envío de preguntas y respuestas de otros usuarios.

En la diagramación de la página principal de Yahoo! Answers, Figura 1, se identifican las siguientes áreas:

- Menú principal:
- Categorías: hay 26 en las cuales se clasifican los diversos temas.
- Notificaciones: área en que se informan acciones tomadas en la página web.
- Área de preguntas
- Publicidad: anuncios
- Tabla de clasificaciones: top de la clasificación de los usuarios según su interacción en el sitio web**.**

Figura 1

Diagramación página principal sitio web Yahoo Answers.



Nota. La figura muestra la página principal del sitio web Yahoo Answers nos muestra sus categorías de los temas de discusión y respuestas, además de una tabla de clasificación de los 10 de los usuarios con mayor interacción en el sitio.

Entre sus características, los usuarios pueden considerar una amplia diversidad de temas, desde lo serio hasta lo trivial, y a su vez, existen algunos de ellos, expertos en temas específicos; que colaboran con sus respuestas. En esta dinámica, aquellos que respondan preguntas suelen recibir

puntos al ser seleccionada su respuesta, o al ser calificada positivamente por otros miembros de la comunidad.

En relación a las políticas de convivencia, cada miembro debe comportarse con integridad, decencia y respeto, el uso de Yahoo! Answers, está sujeto a normas establecidas por la comunidad, si estas no son cumplidas puede resultar en la cancelación de su cuenta de Yahoo! Answers sin previo aviso. Por razones de seguridad, los niños menores de 13 años no pueden hacer o responder preguntas en Yahoo Answers. **(Yahoo, 2021)**

1.3 ¿Principales problemas en Yahoo! Answers?

Los usuarios de la web que interactúan en los sitios web de cQA, a menudo se podrían encontrar con algunos de los siguientes problemas:

- Esperar días hasta que otros usuarios de cQA publiquen respuestas a sus preguntas, que incluso podrían ser incorrectas, ofensivas o spam.
- Recibir respuestas breves sin una argumentación que le dé contexto a la pregunta efectuada.
- Recibir respuestas de parte de usuarios que no dominan el área de la pregunta.
- Validar que el perfil del usuario cumpla con las políticas de la comunidad, ya que es, por ejemplo, difícil confirmar cuál es el género y grupo etario de los usuarios al ser una interacción virtual, sincrónica o asincrónica.

En esta tesis, como equipo nos enfocaremos en la última problemática mencionada anteriormente, porque nos permitirá descubrir, comprender y responder a múltiples preguntas sobre el comportamiento y acciones de usuarios que persiguen distintos fines al interactuar en sitios de preguntas, y que en algunos casos no condicen con su género y rango etario.

1.4 ¿Qué es la predicción automática de géneros y grupos etarios?

En la rama de minería de datos que tiene relación con la predicción de las probabilidades y tendencias futuras. Permite extraer conclusiones confiables sobre eventos futuros, a través de la aplicación de métodos estadísticos, matemáticos y de reconocimiento de patrones. Dicho lo

anterior, en este trabajo se buscará predecir automáticamente los géneros y determinar un grupo etario a través de los datos contenidos de la interacción de los usuarios que se registran en cQA de Yahoo Answers.

1.5 ¿Qué busca la predicción de géneros y grupos etarios?

Para la predicción de géneros se busca determinar ciertos comportamientos, actividades y atributos que la sociedad considera característicos para los diferentes géneros, esto resulta un factor de mucha importancia para las plataformas de cQA a la hora de poder mejorar la experiencia y la retroalimentación con los usuarios de la comunidad.

En el caso de la predicción de los grupos etarios, permite relacionar y clasificar a los miembros con los aspectos demográficos de cada usuario, tales como, género, edad, intereses y ubicación. Sin embargo, muchas veces esta información está incompleta o no está disponible, ya que es opcional que los miembros de la comunidad la proporcionen.

1.6 ¿Qué es entrenar un modelo para predecir género y un grupo etario?

Entrenar un modelo consiste en recopilar un gran número de datos e incorporarlos en un algoritmo de clasificación para determinar el género o grupo etario para unas determinadas características de entrada.

El flujo de proceso para el entrenamiento de un modelo es el siguiente:

1. Recopilar datos.
2. Establecer una métrica de cumplimiento de objetivos.
3. Establecer un protocolo de evaluación.
4. Preparar los datos.
5. Desarrollar un modelo de referencia.
6. Desarrollar un mejor modelo y ajustar sus hiper parámetros.
7. Comparar.
8. Análisis de Resultados.
9. Conclusiones.

1.7 ¿Qué dificultades presenta el proceso de la búsqueda de los datos?

a) Conseguir una muestra de datos representativa y completa, de tal manera que generalice bien los datos.
b) Determinar cuáles son las características que ayudan a identificar el género y edad del usuario. Por ejemplo: sentimientos, gustos, temas de interés.
c) Encontrar un algoritmo adecuado que obtenga un buen desempeño en la clasificación de características de los datos.

1.8 Hipótesis

Considerando la problemática se estableció la siguiente hipótesis:

"Mediante la creación de un plan de entrenamiento de datos se puede encontrar un subconjunto que generalice mejor un modelo de datos, más precisamente, mediante un método sistemático de búsqueda de construcción, podremos establecer una secuencia efectiva de aprendizaje según el contexto de predicción de género y/o grupo etario dentro de la comunidad Yahoo Answers."

1.9 Objetivos

El **objetivo general** de este trabajo es cumplir a través de las diferentes etapas de la metodología de proyecto la validación de la hipótesis planteada.

El **objetivo específico** considera que, mediante un método sistemático de búsqueda de construcción, podremos establecer una secuencia efectiva de aprendizaje para la predicción del grupo demográfico dentro de la comunidad Yahoo Answers.

1.10 Metodología de Desarrollo

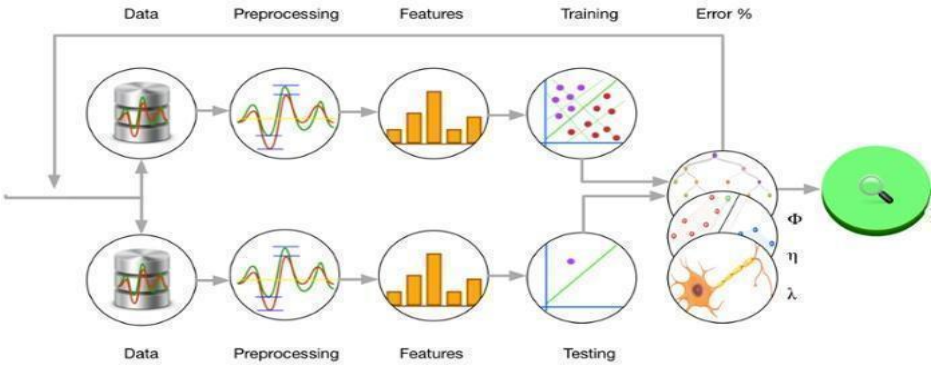
La metodología considerada en el proyecto es la Metodología Exploratoria que pretende darnos una visión general de tipo aproximativo respecto a una determinada realidad.

1.11 Metodología de Proyecto

Para el desarrollo del proyecto se ha considerado utilizar la metodología secuencial de Aprendizaje Automático **(Bonnin, 2017)** como se aprecia en la Figura 2.

Figura 2

Proceso de aprendizaje automático



Nota. En la Figura podemos apreciar el proceso de aprendizaje automático paso a paso, Rodolfo Bonnin (2017).

2 ANÁLISIS EXPLORATORIO DE DATOS

2.1 Bolsa de Palabras

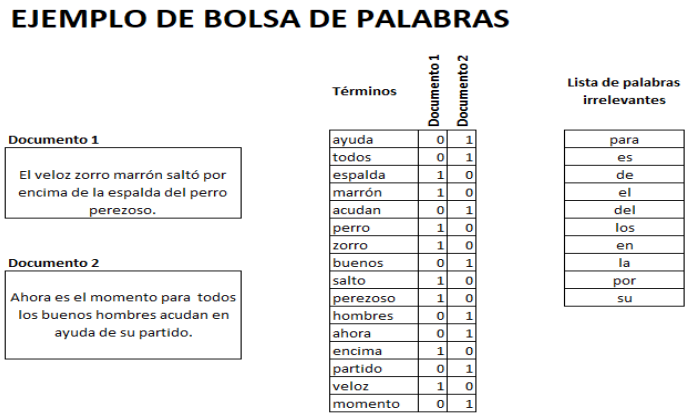
Es una técnica de procesamiento de lenguaje natural de modelado de texto, que permite extraer características de los documentos. Se denomina "bolsa" de palabras porque se descarta cualquier información sobre el orden o la estructura de las palabras en el documento. El modelo solo se ocupa de que las palabras conocidas aparecen en el documento, no en qué lugar del documento.

En otras palabras, es una representación de texto que describe la aparición de palabras dentro de un documento. Simplemente hacemos un seguimiento del recuento de palabras y hacemos caso omiso de los detalles gramaticales y el orden de las palabras.

Uno de los mayores problemas con el texto es que está desordenado y no estructurado, y los algoritmos de aprendizaje automático prefieren entradas de longitud fija estructuradas y bien definidas y, al utilizar la bolsa de palabras (BoW), Ver ejemplo en Figura 3, podemos convertir textos de longitud variable en una de un vector de longitud fija.

Figura 3

Bolsa de Palabras



Nota. En la figura podemos apreciar un ejemplo de una bolsa de palabras, (figura de desarrollo/elaboración propia).

Además, a un nivel mucho más granular, los modelos de aprendizaje automático funcionan con datos numéricos en lugar de datos textuales. Entonces, para ser más específicos, al usar la técnica de la bolsa de palabras (BoW), convertimos un texto en su vector equivalente de números.

Para la extracción de las bolsas de palabras se utilizó la base de preguntas y respuestas de los usuarios de Yahoo Answers, Figura 3**,** y se generaron 3 archivos: uno de entrenamiento con 329.025 líneas, uno de prueba con 109.676 líneas y otro de validación con 109.676 líneas. La exploración de los datos se hará sobre el archivo de entrenamiento, ya que, es el archivo como mayor cantidad de datos.

2.2 Estructura

En la Figura 4 se puede apreciar una línea extraída del archivo de entrenamiento la cual corresponde al conjunto de palabras extraídas para la respuesta de un usuario de Yahoo Answers.

Específicamente esta línea tiene una estructura que es repetitiva y constante para todos los demás registros del archivo. Esta estructura, está formada por un identificador encriptado del usuario, el género (F o M), el año de nacimiento y luego el conjunto de palabras usadas por el usuario con sus respectivas frecuencias. Ver Figura 4.

Figura 4

Frecuencia que la bolsa de palabras utiliza

WGRWISNWTDMOKWQAWQIZX6H7I M 1997 ,:10 -:2

WGRWISNWTDMOKWQAWQIZX6H7I M 1997 ,:10 -:2 -LRB-:2 -RRB-:2 .:14 1000:1 18:1 2x:2 ?:22 ??:4 ANY:2 ASAP:1 And:1 But:2 Can:2 Canada:2 EU:2 Estate:1 How:1 I:32 If:1 Is:1 PC:1 Please:2 Prefferably:1 Real:1 SSD:2 Slovakia:4 So:1 UK:2 USA:4 What:6 Will:1 Windows:2 XP:1 a:4 am:4 amount:2 an:1 and:11 any:4 anything:2 aprehend:1 are:4 bad:1 be:1 because:3 become:1 can:1 citizen:2 closing:1 closings:2 college:8 countries:1 country:3 degree-not:1 different:1 do:16 doe:1 does:1 don:4 dont:3 double:3 escape:4 etc.:1 experience:2 find:1 from:3 go:5 good:2 ground:1 happen:1 have:10 here:1 high:1 hours:2 how:1 i:1 idea:1 if:2 ilnesses:1 imigrate:1 in:6 instead:2 institutions:1 is:1 it:6 jobs:2 just:6 know:2 legal:1 legally:1 life:2 like:2 live:1 living:1 long:1 make:3 might:1 money:3 move:1 my:6 need:1 no:1 nor:1 not:1 of:5 on:2 one:1 only:1 onto:1 other:2 plan:1 please:1 poor:1 possible:2 powers:1 psychiatry:1 put:2 question:1 quit:2 really:1 require:2 required:1 school:1 seems:1 share:1 so:1 some:3 specific:1 state:2 states:1 still:1 t:4 take:1 that:3 the:5 there:2 this:1 to:31 trim:1 want:8 we:1 what:3 where:1 will:2 willing:2 with:1 work:5 worked:2 working:2 would:2 years:1 you:1

Nota. En la figura 4 podemos apreciar la frecuencia de la bolsa de palabras, (Figura de elaboración propia)

2.3 Limpieza de stop words

Se realizó un preprocesamiento donde se filtraron todas aquellas palabras que no son útiles. Estas palabras son llamadas Stop Words y para identificarlas se usó la lista de palabras de la librería nltk de Python. En la Figura 5 se puede visualizar las palabras que conforman el conjunto de stop words según la librería ntlk **(NLTK Project, 2021)**.

Figura5

Detalle de stop words

```
import nltk
from nltk.corpus import stopwords
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves',
'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves',
'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has',
'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for',
'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on',
'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more',
'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't",
'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't",
'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan',
"shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Nota. En la figura se puede visualizar las palabras que conforman el conjunto de stop words según la librería nltk (figura de elaboración propia).

Figura 6

Limpieza a los registros del archivo

```
['MHDH56D2MVV7Q64XQSU4U2HVYU', 'F', '1983', 'I:5', 'I!!:1', 'I!!!:1', '&amp;:2', "'m:1", "'re:1", "'ve:1", ',:5', '.:11', '...:1', '2:1', '2day:1',
'4:1', '::1', '?:9', 'Angelina:1', 'Any:2', 'Does:1', 'For:1', 'Growing:1', 'How:1', 'I:25', 'In:1', 'It:1', 'Jolie:1', 'LADIES:2', 'LADIES!How:1',
'Not:1', 'Now:1', 'ONLY:1', 'Products:1', 'The:1', 'What:1', 'Why:1', 'Yes:1', 'You:1', 'actually:1', 'adorable.I:1', 'advice:3', 'ahead:1',
'always:1', 'answered:1', 'anybody:1', 'anything:2', 'around:1', 'beer:1', 'believe:1', 'bikini:2', 'boyfriend:1', 'bring:1', 'buttons:1', 'buy:1',
'ca:1', 'calories:1', 'care:1', 'cause:5', 'children:1', 'clean:1', 'constantly:1', 'could:2', 'craziest:1', 'crazy:1', 'cute:1', 'da:1', 'day:2',
'deal:1', 'differently:1', 'done:1', 'drink:1', 'drinks:1', 'drive:1', 'drug:1', 'easily:1', 'every:1', 'everything:1', 'ex:1', 'example:1',
'experiences:2', 'far:1', 'feel:2', 'find:1', 'first:1', 'four:1', 'funny:1', 'future:1', 'gaining:1', 'get:3', 'getting:1', 'girls:1', 'go:2',
'going:1', 'got:1', 'grew:1', 'grocery:1', 'guy:4', 'hands:1', 'homemade:1', 'honestly:1', 'it.But:1', 'it.I:1', 'keep:1', 'know:3', 'know.But:1',
'knowing:1', 'legs:1', 'life:1', 'like:2', 'lips:1', 'long:1', 'love:3', 'luck:1', 'makes:1', 'man:2', 'married:1', 'matter:1', 'means:1',
'messages:1', 'mixed:1', 'moderation:1', 'money:1', 'month:1', 'move:2', 'much:1', "n't:6", 'natural:1', 'necessarily:1', 'need:1', 'never:2',
'person:1', 'prison:1', 'products:1', 'public:1', 'push:1', 'put:1', 'questions:1', 'realized:1', 'really:1', 'reflection:1', 'relationship:1',
'saw:1', 'say:2', 'seem:2', 'send:1', 'sent:1', 'serious:1', 'sex:1', 'sexy:1', 'share:1', 'shop:1', 'shower:1', 'sms:1', 'sounds:1', 'still:2',
'store:1', 'stunning:1', 'summer:1', 'sumthing:1', 'swimming:1', 'system:1', 'take:1', 'talk:1', 'term:1', 'test:1', 'text:1', 'thing:1', 'things:2',
'think:2', 'thinking:1', 'thought:1', 'three:1', 'time:2', 'times:2', 'together:1', 'told:1', 'total:1', 'two:1', 'u:1', 'uneasiness:1', 'upcoming:1',
'upset:1', 'upsets:1', 'visit:1', 'wait:1', 'wanting:1', 'waxed:1', 'way:1', 'wearing:1', 'weekdays:1', 'weekends.I:1', 'weight:1', 'window:1',
'woman:2', 'worn:1', 'worried:1', 'would:8', 'write:1', 'wrong:1', 'years:1']
Before Cleaning:  269  After Cleaning:  201
['a:12', 'about:4', 'again:1', 'against:1', 'all:1', 'an:2', 'and:7', 'any:2', 'are:1', 'at:1', 'be:1', 'before:3', 'being:1', 'but:5', 'can:2',
'do:10', 'does:2', 'doing:2', 'down:1', 'each:1', 'for:8', 'had:1', 'have:9', 'having:1', 'he:2', 'here:1', 'him:4', 'his:2', 'how:3', 'if:2', 'in:9',
'is:3', 'it:6', 'just:1', 'me:1', 'most:1', 'my:7', 'myself:1', 'no:1', 'now:1', 'of:5', 'off:1', 'on:3', 'once:1', 'or:3', 'other:2', 'our:3',
'out:2', 'over:3', 'own:1', 'same:1', 'should:3', 'so:1', 'that:5', 'the:8', 'them:1', 'they:6', 'to:10', 'too:1', 'up:3', 'was:5', 'we:4', 'what:1',
'when:1', 'while:2', 'who:1', 'with:8', 'you:5']
```

Nota. En la figura se puede observar la extracción de las palabras removidas (figura de elaboración propia)

Para el primer registro mencionado en la Figura 5, había 269 palabras, luego de la limpieza quedaron 201 palabras. Las palabras removidas se pueden visualizar al final de la Figura 6.

A continuación, se muestra en la Tabla 1, una sumarización de los datos antes y después de remover las stopwords. En total se removieron 89.876 palabras únicas con una frecuencia total de 457.421.152. En promedio por usuario el total de palabras únicas utilizadas es de 700 palabras y después de remover las stopwords el promedio bajó a 624,54. Si se compara el promedio de palabras de perfiles femenino vs perfiles masculinos se puede observar que las mujeres tienen un promedio mayor de palabras y al remover las stopwords este promedio se reduce aproximadamente en un 20% para ambos géneros. Esta diferencia en la cantidad de palabras viene dada a que está demostrado que las mujeres hacen mayor uso de palabras que los hombres.

Tabla 1

Detalle estadístico de la bolsa de palabra.

	Con Stopwords	**Sin Stopwords**	**Diferencia**
Número de Palabras únicas	5.901.196	5.811.320	89.876
Total Palabras	1.178.689.618	721.268.466	457.421.152
Promedio	700,80	624,54	76
Desviación Estándar	1.035,67	1.022,11	14
Promedio Masculino	676,97	604,04	73
Desviación Estándar Masculino	1.162,83	1.150,01	13
Promedio Femenino	714,99	636,75	78
Desviación Estándar Femenino	951,58	937,43	14

Nota. En la tabla es posible visualizar el detalle de los resultados entregados por la Bolsa de Palabras (tabla de elaboración propia).

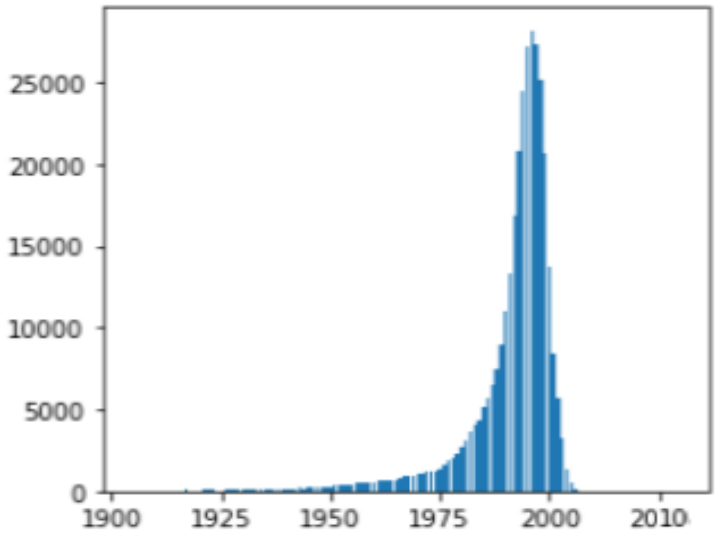
2.4 Definición de Grupos etarios

Para determinar los grupos etarios se realizó un diagrama de barras con la frecuencia de los años de nacimiento de cada perfil, ver Gráfico 1.

En el diagrama de barras se puede observar que la mayor cantidad de perfiles se concentran en los años de nacimiento de 1980 hasta 2010, teniendo el mayor peak en el tramo de 1990 al 2000.

Gráfico 1

Frecuencia de los años de nacimiento de cada perfil por grupo etario



Nota. En el gráfico se puede observar la cantidad de perfiles se concentran en los años de nacimiento desde 1980 hasta 2010 (gráfico de elaboración propia).

Para determinar los rangos de los grupos etarios, se consideró la definición mencionada en la web de kasasa **(KASASA, 2021)** y zion & zion **(Zion & Zion, 2018)**, debido a que los rangos propuestos se ajustan a los resultados de concentración de generación obtenidos en el diagrama de barras del Gráfico 1. Para el número de casos que están fuera de los rangos etarios definidos se les agrupó como Outliers, los cuales no serán considerados en el entrenamiento.

La Tabla 2 da una percepción sobre la distribución de las edades observadas dentro de nuestro dataset de investigación. Para proporcionar este rango de agrupación se estableció bajo la teoría generacional de William Strauss y Neil Howe **(Brouoadmin, 2016)**.

Tabla 2

Detalle de grupo etario.

Nombre del Grupo	Rango de Años	Número de registros en archivo
Generación Silenciosa	<1946	3.087
Baby Boomers	1946-1964	8.409
Generación X	1965-1980	21.076
Generación Y	1981-1996	190.471
Generación Z	1997-2015	105.980
Outliers	2015 (fuera de los rangos definidos)	2

Nota. En la tabla es posible visualizar el detalle del rango etario establecidos en nuestro proyecto (tabla de elaboración propia).

A continuación, se describe cada grupo etario:

a) GENERACIÓN SILENCIOSA

Una de las prioridades de esta generación es mantenerse en contacto con sus seres queridos, en especial los que viven lejos, por lo que toda tecnología que facilite el acercamiento y la comunicación con sus familiares y amigos será bien recibida. Es importante tener en cuenta que esta generación es muy susceptible de sufrir estafas y fraudes a través de internet. **(Brouoadmin, 2016)**

b) BABY BOOMERS

Esta generación se caracteriza por usar mayormente YouTube y Facebook, disfrutan compartiendo vídeos, fotos y artículos con los que se sienten identificados, quieren mantenerse activos, tanto física como mentalmente, por lo que las herramientas digitales que les permitan encuentros y experiencias con personas afines o ejercicios para mantenerse sanos serán bien recibidas. **(Statista, 2019)**

c) GENERACIÓN X

La generación puente entre los Baby Boomers y Generación Y, muchas veces olvidada por las marcas y el mundo del marketing. Sin duda, les gusta usar su teléfono móvil, pero pueden no tener idea de cómo funciona Snapchat. Llegaron a la mayoría de edad durante el boom de la tecnología, pero pasaron más tiempo al aire libre mientras crecían, por lo que se preocupan menos por subir su último selfie a Facebook. La nostalgia de la época de su juventud en los 90s es una buena estrategia para conectar con ellos. Muy orientados a la familia, y sus necesidades, la Generación X valora, además, la autenticidad y la honestidad de las marcas. Demandan aplicaciones móviles que les permitan estar cerca de su familia y que a la vez garantizan un uso seguro de internet por parte de sus hijos. Están abiertos a experimentar con lo digital en el punto de venta y a dejarse sorprender por las marcas. **(ICEMD, 2019)**.

d) GENERACIÓN Y

Para conectar con la Generación Y también denominados "Millennials" es fundamental entender sus estilos de vida, sus gustos e intereses. Estar al día de lo que culturalmente está de moda, los últimos movimientos estéticos o musicales. Más que consumir productos y servicios, los Millennials co-crean sus marcas personales con las empresas. Las redes sociales son sus escaparates, a través de las cuales se expresan y se definen en comunión con las marcas. Experiencias novedosas, toques de humor e ironía para la comunicación y el marketing orientado a esta generación. Productos y servicios novedosos, "on demand" también para los padres Millennials y sus hijos **(ICEMD, 2019)**.

e) GENERACIÓN Z

Esta generación demanda tener control en la manera en la que reciben el contenido. Hay que dejarlos decidir sus preferencias, ya que las interrupciones de marca no son bienvenidas. Están muy comprometidos socialmente con causas como el feminismo, el debate en torno al género o la diversidad, por lo que las marcas que se alineen con estas iniciativas tendrán el visto bueno de esta

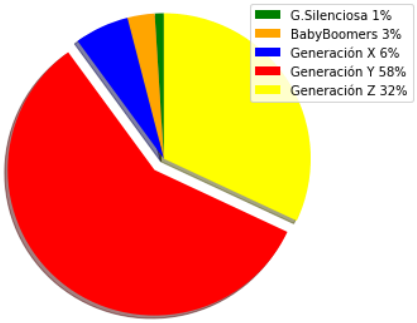
generación. Será importante comunicarse con ellos a través de múltiples plataformas y adaptar el contenido a cada plataforma: el mensaje general puede ser el mismo, pero el contenido debe ser específico de Instagram, YouTube, Facebook, Snapchat, etc. Y sobre todo mantenerse actualizado, saber qué nuevas plataformas utilizan y para qué **(ICEMD, 2019)**.

2.5 Distribución porcentual de grupos etarios

Según el Gráfico 2, el 90% de los perfiles de los usuarios de Yahoo Answers se concentran en los grupos etarios de Generación Y y Generación Z, siendo la generación Y a la que pertenecen la mayor cantidad de perfiles. Esta distribución se puede explicar por las características propias de cada grupo etario en donde los usuarios pertenecientes a la generación Y y Z son usuarios digitales de edades entre 20 a 40 años que son activos en las redes sociales y plataformas como Yahoo Answers, en contraste con usuarios pertenecientes a la Generación Silenciosa, Baby Boomers y Generación X, los cuales hacen poco o ningún uso de las redes.

Gráfico 2

Porcentaje de grupos etarios



Nota. A continuación, en el gráfico se puede visualizar los perfiles que se concentran en los grupos etarios por porcentajes (gráfico de elaboración propia).

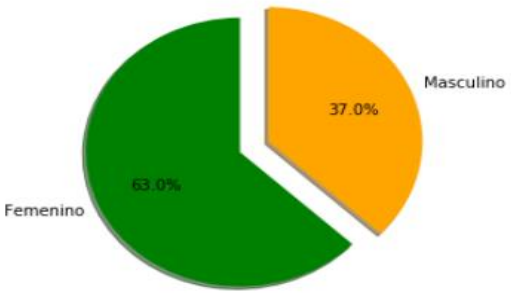
3 EXPLORACIÓN DE LOS DATOS

3.1 Distribución porcentual de género por grupo etario

Tal como se puede observar en el Gráfico 3, el 63% de los perfiles de Yahoo Answers corresponden a perfiles de género femenino y un 37% a perfiles de género masculino. Sin embargo, al determinar la distribución de estos géneros entre los rangos de grupo etarios definidos, se puede observar en el Gráfico 4 y 5 que ambos géneros están distribuidos en proporciones muy similares (diferencia de 1%) dentro de los mismos grupos etarios, específicamente la Generación Y y Z.

Gráfico 3

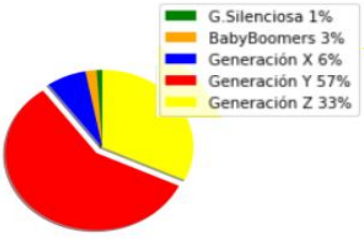
Porcentaje de perfiles de usuarios por género



Nota. En el siguiente gráfico se muestra la distribución de perfiles por género (gráfico de desarrollo propio)

Gráfico 4

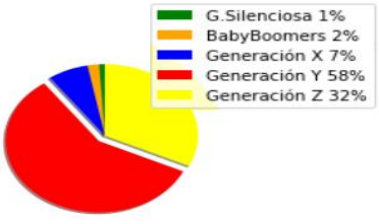
Porcentaje de distribución del género masculino entre los rangos de grupo etarios



Nota. En el siguiente gráfico se muestra la distribución de perfiles por género masculino (gráfico de desarrollo propio).

Gráfico 5

Porcentaje de distribución del género femenino entre los rangos de grupo etarios



Nota. En el siguiente gráfico se muestra la distribución de perfiles por género femenino (gráfico de desarrollo propio).

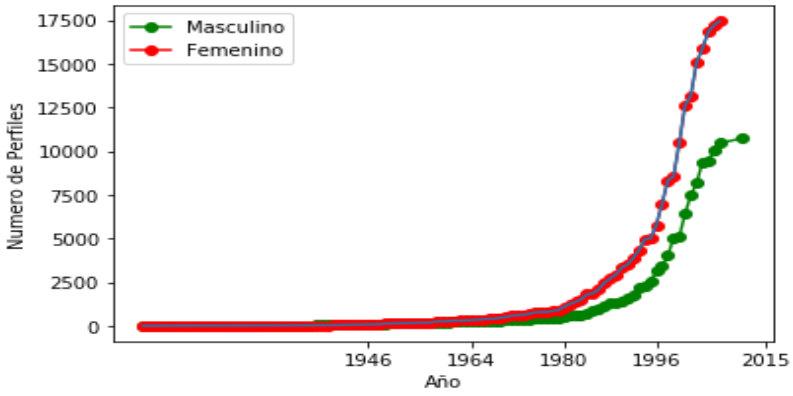
3.2 Gráfico de líneas, por año y género

Tal como se puede observar en el Gráfico 6, se puede ver una similitud entre las curvas del género masculino y femenino, y los puntos que representan los años. Esto complementa lo analizado en los Gráficos 3, 4 y 5.

Podemos apreciar que en el transcurso de los años existen más perfiles de género femenino que masculinos.

Gráfico 6

Se detalla el número de perfiles por año



Nota. Se observa en el gráfico el número de perfiles por año y género

3.3 Palabras más utilizadas por género.

En las Figuras 7 y 8, podemos determinar que la palabra más utilizada por género es “Friend” (Amigo(a)), y la menos utilizada es “Lot” (cantidad). Además, se puede apreciar que entre los géneros las palabras utilizadas “School” (colegio), “God” (Dios) y “Car” (auto) son sólo utilizadas por el género masculino en cambio el género femenino es el único que utiliza las palabras “Love” (amor), y Baby (bebé).

Figura 7

Palabras más utilizadas por el género masculino.

Nota. En la figura podemos visualizar e identificar por su mayor tamaño las palabras más utilizadas por los hombres.

Figura 8

Palabras más utilizadas por el género femenino.

Nota. En la figura podemos visualizar e identificar por su mayor tamaño las palabras más utilizadas por las mujeres.

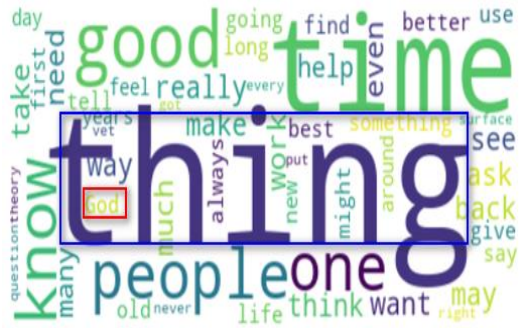
3.4 Palabras más utilizadas por grupo etario.

En la Figura 9 que representa la Generación Silenciosa se destaca la palabra “Thing” (cosa) como la más utilizada, y en el caso de la Figura 10 que representa la Generación Baby Boomers se muestra la palabra “Know” (saber) como la más utilizada, no obstante, podemos destacar que la Generación Silenciosa utiliza la palabra “God” (Dios) con menor frecuencia que los Baby Boomers.

En las Generaciones X (Figura 12) y la Generación Y (Figura 11), se identifica que la palabra más utilizada es “Good” (bueno(a)) en la Generación X, y en cambio en la Generación Y, la más utilizada es la palabra “friend” (amigo(a)). Por el contrario, en la Generación Y, las palabras menos utilizadas son “work” (trabajo), “never” (nunca) y “Thanks” (gracias), A su vez, para la Generación X, las palabras menos utilizadas son “child” (niño(a)) y “always” (siempre).

Figura 9

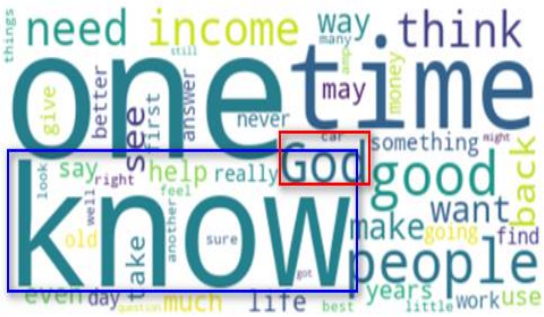
Palabras más utilizadas por Generación Silenciosa.



Nota. En la figura podemos visualizar e identificar por su mayor tamaño las palabras más utilizadas por la Generación Silenciosa.

Figura 10

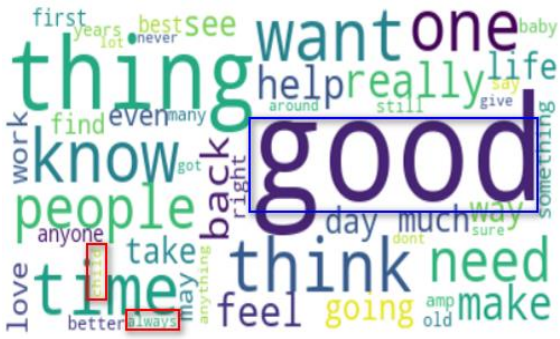
Palabras más utilizadas por los Baby Boomers.



Nota. En la figura podemos visualizar e identificar por su mayor tamaño las palabras más utilizadas por los Baby Boomers.

Figura 11

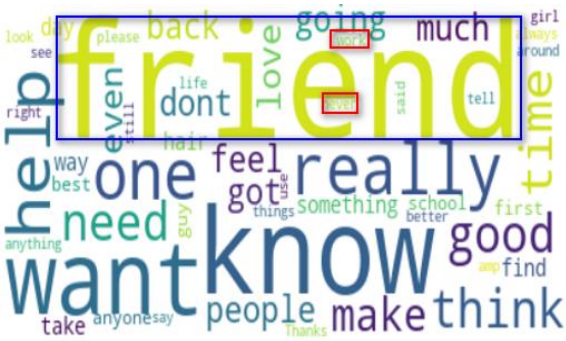
Palabras más utilizadas por la Generación Y.



Nota. En la figura podemos visualizar e identificar por su mayor tamaño las palabras más utilizadas por la Generación Y.

Figura 12

Palabras más utilizadas por la Generación X.



Nota. En la figura podemos visualizar e identificar por su mayor tamaño las palabras más utilizadas por la Generación X.

En el caso de la generación Z (Figura 13), destacan 2 palabra con uso más frecuente, que son “Friend” (Amigo(a)), “Please” (por favor), esto confirma la interconexión que existe entre esta generación y la forma de comunicarse a través de este sitio web de preguntas y respuestas como es Yahoo! Answers.

Figura 13

Palabras más utilizadas por la Generación Z.

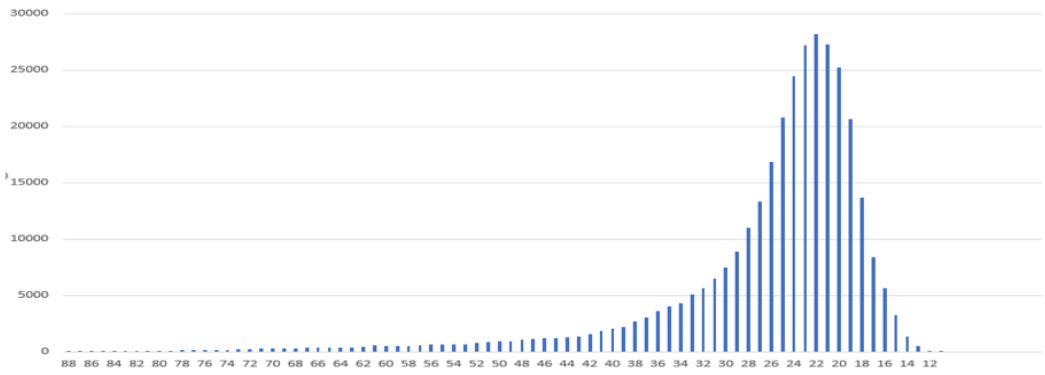
Nota. En la figura podemos visualizar e identificar por su mayor tamaño las palabras más utilizadas por la Generación Z.

3.5 Distribución por edades

El Gráfico 7 muestra la distribución de usuarios entre 12 y 88 años, siendo los usuarios de 22 años que más presencia tienen en el sitio Yahoo Answers.

Gráfico 7

Visualización gráfica de la relación de la cantidad de usuarios y sus edades.



Nota. De acuerdo con lo que podemos observar en el gráfico 7, la distribución de la cantidad de los usuarios versus sus edades.

3.6 Entropía de palabras por Género

A continuación, en la Tabla 3, se detallan las 18 palabras más utilizadas de acuerdo con el género masculino y femenino, ordenadas de mayor a menor entropía.

Tabla 3

Detalle de las 18 palabras más utilizadas en la bolsa de palabras

N°	Palabras	Frecuencia Masculino	Frecuencia Femenino	Entropía
1	girl	31337	36214	0.299897141
2	people	61877	86699	0.294940674
3	good	91847	131286	0.294210339
4	work	25771	37019	0.294023986
5	life	21658	33794	0.29054443
6	Thanks	24571	38523	0.290323516
7	way	28144	45033	0.289358261
8	see	26437	44943	0.286266176
9	need	71318	121347	0.286219121
10	find	24312	41479	0.286072863
11	lot	17744	30483	0.285697584
12	one	101526	176557	0.28502401
13	around	19177	33846	0.284191383
14	best	24887	44981	0.282812819
15	know	160361	293294	0.282108294
16	time	61736	114910	0.281045127
17	back	43242	81832	0.280017465
18	even	33658	63723	0.279989969

Nota. Podemos observar en la tabla, el detalle de la entropía de palabras por género

A continuación, se detalla en Tabla 4, las 21 palabras más utilizadas de acuerdo con el grupo etario, estas se encuentran ordenadas de mayor a menor entropía.

Tabla 4

Detalle de las 21 palabras más utilizadas según los grupos etarios

N°	Words	Silente	Baby Boomer	Generación X	Generación Y	Generación Z	Entropía
1	see	2110	8391	8972	38551	13356	0.54840599
2	people	4417	12830	14505	80950	35874	0.5285997
3	way	2058	6994	8664	42339	13122	0.52211623
4	take	1882	5903	8832	41660	13281	0.5156157
5	time	4880	13262	20022	102853	35629	0.51166641
6	things	1196	3506	5595	29503	11321	0.5058938
7	never	1157	4729	5487	35176	13354	0.49632457
8	best	1413	3903	7117	41243	16192	0.48759055
9	make	2835	7684	13164	84035	38186	0.48520939
10	even	1792	6213	8251	57115	24010	0.48484558
11	first	1211	4324	6898	40270	13348	0.48317843
12	one	3438	18019	25388	161806	69432	0.48280573
13	good	4533	12530	16496	130302	59272	0.47758417
14	day	1247	4586	8384	51816	22086	0.47639734
15	something	1642	4912	7053	53500	22464	0.47223063
16	need	2761	9981	14211	114855	50857	0.46314925
17	think	3099	9978	20027	143313	61560	0.45731089
18	much	1797	5407	9051	71871	28384	0.45486213
19	feel	1255	3803	9345	73932	37198	0.4419283
20	going	1557	4775	9092	79673	33415	0.43856822
21	know	3806	13492	30803	269345	136209	0.43345256

Nota. Se aprecia que en la tabla el detalle de la entropía de palabras por grupos etarios

4 ESTADO DEL ARTE

4.1 Trabajo Relacionado

El artículo **(Onur Kucuktunc, 2012)** trata de la extracción de los sentimientos desde los documentos de la web que pertenecen a un sitio de respuestas a preguntas en línea de uso masivo como es Yahoo! Answers. Esto permite observar el estado de ánimo de los usuarios, el cual podría ser utilizado en aplicaciones comerciales.

Este trabajo comenzó el análisis mirando las correlaciones directas, por ejemplo, observando sentimientos más positivos los fines de semana, muy neutrales en el tema de Ciencias y Matemáticas, una tendencia de los jóvenes a expresar sentimientos más fuertes, o de las personas en las bases militares a hacer las preguntas más neutrales.

Posteriormente se analizó lo básico investigando cómo las propiedades del par, los cuales son (preguntador, respondedor) afectan el sentimiento presente en la respuesta. Entre otras cosas, se observa una dependencia del emparejamiento de algunos atributos inferidos de acuerdo con la localidad donde el usuario vive (zona geográfica).

Las mejores respuestas provienen de la divergencia de sus sentimientos, las cuales se diferencian notoriamente de otras respuestas donde no existen los sentimientos como, por ejemplo, en el tema Negocios y Finanzas, las mejores respuestas tienden a tener un sentimiento más neutral que respuestas en otras categorías. Finalmente, se hace el estudio en el contexto de predecir la actitud que una pregunta provocará en las respuestas. En general, se cree que comprender los factores que influyen en el estado de ánimo de los usuarios no solo es interesante desde un punto de vista sociológico, sino que también tiene aplicaciones en publicidad.

En **(Bayot, 2016)** se estudió el problema específico asociado con la tarea de elaboración de perfiles de autores de PAN 2016 **(Webis Group, 2021)**. Esto implicó que el perfilado se realizó en dos dimensiones diferentes: clasificación por edad y género. Se encontró que las SVMs (Support Vector Machines) que mejor clasifican utilizan un kernel radial o polinomial

Por otro lado, el trabajo **(Young-Sik, 2014)** abordó la clasificación de rostros basada en la edad, la búsqueda de niños perdidos, el monitoreo de vigilancia, y reconocimiento facial, el cual es afectado por el envejecimiento se ha convertido en un factor nuevo que hace que la tarea de reconocer automáticamente los rostros sea aún más desafiante. Es por ello que existen muchos factores que afectan la precisión de la estimación de la edad, el género y la expresión facial. La expresión puede tener efectos negativos. Esta investigación, investiga los efectos del género y la expresión facial en la estimación de la edad mediante el método de regresión de vectores de soporte (RVS).

El artículo **(Chen J., 2017)** estudia la predicción de género en el conjunto de datos de las redes sociales. Propone un enfoque que combine el método de indexación semántica latente (LSI) con algoritmo k-Nearest Neighbor (KNN) para predecir el género en función de una colección de publicaciones de la vida real en páginas de blogs reales. La eficacia en el procesamiento de datos a gran escala y de gran dimensión se demuestra mediante resultados experimentales. Los resultados experimentales muestran la eficacia del método propuesto KNN + LSI. Y también presenta un vasto detalle de implementación en un problema clásico de clasificación de texto.

En resumen, **(Morrison, 2013)** es una investigación sobre el conocimiento que desarrolla un programador en relación con la cantidad de años que tiene. Ya que muchos parecen pensar que el envejecimiento conlleva una disminución en la adopción y absorción de nuevos conocimientos de la programación. Desarrollaron varias preguntas de investigación sobre este tema y se basó en datos de StackOverflow (SO) **(Stack Overflow, 2021)** para abordar estas preguntas. Se observaron que los puntajes de reputación de los programadores aumentan en relación con la edad hasta bien entrados los 50, que los programadores de 30 años tienden a centrarse en menos áreas en relación con los más jóvenes o mayores, y que no existe una fuerte correlación entre la edad y los puntajes en conocimientos específicos áreas.

En la actualidad las comunidades de preguntas y respuestas como Yahoo Answers, necesitan mantener un intercambio entre los diferentes usuarios, esto se realiza mediante la asignación de preguntas a usuarios que pudieran responderlas. Esta asignación o vinculación debe realizarse de tal manera de garantizar en lo posible que exista una respuesta a la pregunta que luego será revisada por la comunidad y calificada, por tal motivo los datos demográficos del usuario juegan un

importante rol para el proceso de asignación, sin embargo, estos datos no siempre están disponibles porque los usuarios no siempre los informan. Por este motivo, el trabajo **(Figueroa, 2017)** propone un método para la inferencia de estos atributos a través del aprendizaje automático.

Este estudio no solo se centró en atributos lingüísticos, sino que también consideró atributos extraídos del análisis de sentimientos, de árboles de dependencia y árboles de constituyentes, así como también del análisis de HPSG. También se consideró otras características no lingüísticas extraídas de los datos de clics de búsqueda en la web, información demográfica, social y de metadatos que pueden ser encontrados en algunas plataformas CQA.

Entre los resultados de este trabajo, encontramos que los datos demográficos como la ocupación, la industria y la edad de quien pregunta, en combinación con una categorización detallada de la pregunta, son vitales para identificar correctamente el género. Los rasgos lingüísticos extraídos de la pregunta y la autodescripción del autor también fueron útiles. Por otro lado, los rasgos no lingüísticos tenían mayor poder discriminativo que los rasgos lingüísticos. Aunque la mitad de las selecciones estaban orientadas lingüísticamente, ocho de las diez más efectivas eran no lingüísticas.

En el artículo **(Schwartz HA, 2013)** se presenta un estudio de la personalidad y su relación con el lenguaje realizado en el 2013, para el cual convocaron a 75.000 voluntarios para aplicarles encuestas de personalidad y analizar su perfil de Facebook **(Facebook, 2021)**, con el propósito de desarrollar un vocabulario abierto que permitiera entender los vínculos entre la personalidad y el lenguaje.

La técnica implementada permitió explorar las expresiones y rasgos de las personas en las redes sociales para encontrar la relación entre las palabras, frases y tópicos como funciones de atributos conocidos como género, edad, ubicación y características psicológicas. Por ejemplo, la técnica permitió detectar que las personas con personalidad extrovertida tendían a mencionar palabras sociales como “fiesta”, “amor”, “niños” y “mujeres”, mientras que las personas introvertidas mencionan palabras relacionadas con actividades solitarias como "computadora", "Internet" y "lectura". Otro hallazgo del estudio fue que las personas emocionalmente estables escribían sobre actividades sociales que fomentan una mayor estabilidad emocional, como “deportes”,

“vacaciones”, "playa", "iglesia", "equipo". En cambio, los introvertidos, se interesan en medios japoneses como: "anime", "manga", "japonés", emoticones de estilo japonés: ^_^, y temas de anime.

Para inferir datos demográficos, existen diversas técnicas con diferentes resultados y rendimientos, sin embargo, todas estas coinciden en que requieren generar un etiquetado manual de los perfiles para la data que será usada en el entrenamiento de la máquina, lo cual es costoso tanto de tiempo como recursos económicos. En **(Emmery, 2017)** se proponen una técnica económica y con rendimiento similar, para inferir el género (masculino o femenino) de un perfil en la red social Twitter haciendo uso de consultas simples como etiquetas distantes (supervisión distante) para obtener data para el entrenamiento.

Los sistemas de recomendación son aquellos sistemas que se encargan de hacer sugerencias al usuario en relación con algún producto o servicio. Actualmente existen numerosas propuestas de sistemas de recomendación cada uno con diferentes % de fiabilidad, sin embargo, se hace compleja la comparación de estas propuestas debido a que los investigadores usan diferentes métricas, métodos y conjunto de datos. En **(Beel, 2013)** proponen que los datos demográficos y otras características del usuario deben ser consideradas para evaluar la fiabilidad de estos sistemas de recomendación.

En este trabajo, se evalúa el impacto del género edad y si es usuario registrado o no, en los CTR (Click Through Rates) clics de papers recomendados por el sistema. De un total de 1.028 usuarios que recibieron recomendaciones, el 38,62% no se registró y 61,38% registrados. 21,79% se registró, pero no proporcionó información sobre su género, 33,17% registrados eran hombres y 6,42% eran mujeres. Entre los géneros hay solo una marginal diferencia en el CTR con 6.88% para hombres y 6.67% para mujeres. Hay una diferencia significativa entre los usuarios registrados (6,95%) y los usuarios no registrados (4,97%). Curiosamente, aquellos usuarios que se registraron y no especificaron su género tenían el CTR más alto con 7.14%.

Además, en [14] se mostró que, del universo de usuarios masculinos, 38,09% activó recomendaciones mientras que solo el 34,74% de las mujeres lo hicieron e incluso mucho menos los usuarios que no especificaron su género (28,72%), lo cual podría indicar que estos usuarios

están preocupados por cuestiones de privacidad. Mientras un usuario tenga más edad (mayor de 60 años) mayor será el CTR (promedio 9,92%), mientras que los usuarios más jóvenes (20-24 años) tienen el CTR más bajo (promedio 2.73%).

El artículo, [15] trata de un experimento realizado por un número de personas que tenía que adivinar el género y edad de varios perfiles de Twitter, solamente leyendo algunos tweets. Con este experimento querían comprobar que tan fiable es el resultado de la inferencia de los programas automáticos vs inferencia humana. Según los autores de este artículo inferir el género y edad es algo muy difícil, que actualmente se hace desde una perspectiva biológica y estática, sin embargo, en el artículo proponen que ambos atributos se deben considerar desde una perspectiva social y fluida.

5 MARCO TEÓRICO

En el ámbito de los sitios web de preguntas y respuestas (a.k.a community question answering; cQA), los cuales han tomado protagonismo en el último tiempo, guiando a los usuarios a interactuar en estos sitios buscando respuestas a sus preguntas.

La libertad que entregan estas plataformas permite el anonimato, que puede conllevar a ocultar intenciones que se alejan de la génesis de estos sitios.

Algunos autores buscan utilizar en sus estudios los datos que se generan en los sitios web de preguntas y respuestas, tal como:

"Figueroa, A. (2017). en su artículo Hombre o Mujer: Qué rasgos caracterizan las preguntas planteadas por cada género en la respuesta a preguntas de la comunidad"

En este trabajo se busca utilizar y procesar los datos que se generan en los sitios web como el de Yahoo Answers, a través de la creación de una máquina de entrenamiento, en esta misma línea de búsqueda se selecciona la técnica de bolsa de palabra (BOW), que permite un análisis granulado de los datos y que convierte un texto en su vector equivalente de números.

Con la extracción de las bolsas de palabras se generan 3 archivos orientados al entrenamiento de la máquina: uno para entrenamiento, uno para prueba y otro de validación. Tal como lo define:

"Ripley, B. (1996). página 354, Reconocimiento de patrones y redes neuronales".

La exploración de los datos se hará sobre el archivo de entrenamiento, el cual contiene la mayor cantidad de datos.

El paso siguiente es filtrar los datos inútiles, en el procesamiento del lenguaje natural, las palabras inútiles (datos) se denominan palabras vacías. Para ello, podemos eliminarlos fácilmente, utilizando la **librería de ntlk.corpus (Steven Bird, 2008),** ya que este cuenta con un diccionario de palabras vacías para dar limpieza a nuestros datos de entrenamiento.

Al estar limpio nuestra base de datos de entrenamiento debemos crear un diccionario con una frecuencia de palabra mayor a 5, todas estas en minúsculas y sin signo de puntuación, la cual será la base de datos que utilizaremos para la construcción de una matriz, donde las columnas van a corresponder a cada una de las palabras del diccionario y la fila cada una de los usuarios de entrenamiento, una vez construida la matriz se crea un vector de salida para cada uno de los usuarios de entrenamiento, estas salidas corresponde a una dupla de grupo etario y género.

Con la matriz y el vector de salida creados, se procede a realizar el entrenamiento del modelo mediante el uso de las redes **Naive Bayes (Brownlee, 2020)** y la **librería de Scikit-learn de Python (scikit-learng, 2021).** Para comprobar la hipótesis planteada, se realizaron 2 experimentos. En el primer experimento se realiza el entrenamiento utilizando un solo lote de datos conformado por los datos del archivo de entrenamiento. Para el segundo experimento se segmentan el lote de datos del archivo de entrenamiento en pequeños lotes, clasificados según año y género (curriculums), y se refina mediante la implementación de un **algoritmo greedy (Creative Commons Atribucion Compartir, 2021)** en donde el mejor resultado que selecciona es el que tiene mejor métrica f-macro.

Una vez finalizado ambos experimentos se procede a probar los modelos generados haciendo uso del archivo de test, para generar las métricas f-macro, f-micro, f-weighted y matriz de confusión. Al comparar los resultados, se puede comprobar que se generó una secuencia efectiva de aprendizaje o plan de entrenamiento para el modelo de Naive Bayes, el cual, permite entrenar de manera eficiente con el menor grupo de datos y en el menor tiempo. El análisis de los resultados será explicado en la sección de resultados parciales del informe.

6 RESULTADOS

Para probar la hipótesis de esta investigación se desarrollaron dos experimentos. En ambos hemos utilizado el mismo modelo de clasificación para la predicción de datos de género/edad según el análisis predictivo, mediante la utilización de un archivo de entrenamiento que contiene la información de género, años (desde 1946 al 2015), y una bolsa de palabras (329.025 usuarios). Para poder avanzar con el entrenamiento, se utilizó un método estadístico o de clasificación de Naive Bayes, donde las variables utilizadas contenían un número finito de categorías o grupos distintos, los cuales no tenían un orden lógico, en nuestro caso las categorías que analizamos fueron la clasificación de grupos etarios (edades) y su género; Asumimos que cada uno de los atributos eran independientes, y sin correlación de un factor con el otro.

Para poder implementar este modelo de clasificación, fue necesario crear y utilizar un diccionario de palabras del archivo de entrenamiento, el cual cumplió con las siguientes condiciones:

- Las palabras tienen una frecuencia mayor o igual que cinco (solo se tomaron las palabras que se repetían cinco o más veces)
- Se eliminaron las stop-words o palabras de poco valor según lo definido en la librería nltkcorpus.
- Se eliminaron aquellas palabras que representaran un signo de puntuación.

Como resultado se obtuvo un diccionario con un vocabulario con 435.964 palabras. Este diccionario fue utilizado para construir una matriz X, donde cada fila corresponde a un usuario y cada columna corresponde a una palabra del diccionario, la cual tendría como valor la frecuencia de uso de la palabra usada. Para aquellos casos donde la palabra no fue usada, se le asignó el número cero, debido a que el diccionario contenía todo el universo de palabras del archivo y no necesariamente el usuario las utilizó toda en su totalidad.

Adicionalmente, para cada fila en la última columna se le asignó un número entre el -1 y el 9, el cual corresponde a un valor que clasifica al usuario según su género y grupo etario, como se aprecia en la Tabla 5.

Tabla 5

Detalle de los vectores de clases y géneros.

Vector de Clases		**Género**	
Rango	**Grupo Etario**	**M**	**F**
< 1946	SILENTES	0	5
1946-1964	BABY BOOMERS	1	6
1965-1980	GEN X	2	7
1981-1996	GEN Y	3	8
1997 a 2015	GEN Z	4	9
Outliner	Out	-1	-1

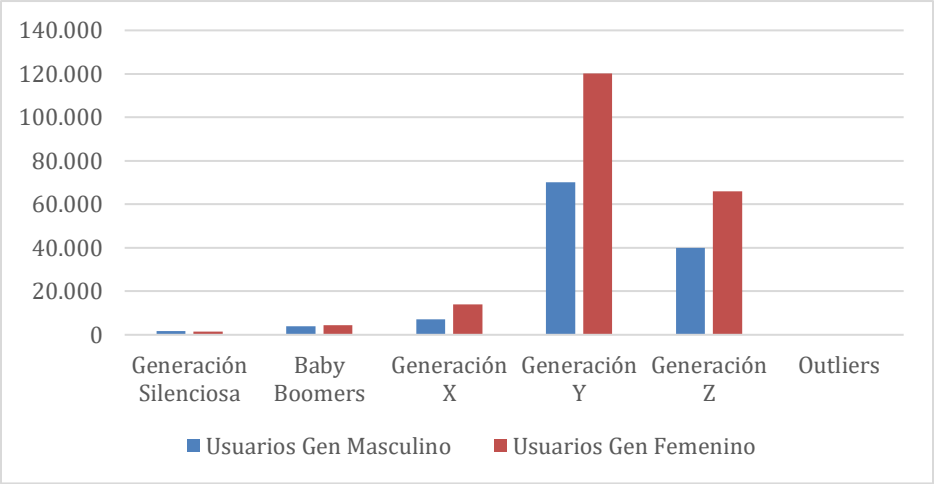
Nota. Se aprecia que en la tabla el detalle del rango y el grupo etario con relación a los géneros masculinos y femeninos (tabla de elaboración propia)

El resultado fue una matriz de 329.025 x 435.954, donde las filas son los usuarios (329.025) y las columnas el total de palabras del diccionario (435.954). Adicionalmente, se agregó una columna correspondiente a la clase, a la cual pertenecía el usuario según la tabla anterior.

Con la definición de las clases antes mencionadas, se generó un vector de clases llamado "Vector Y", el cual tenía la clasificación de cada usuario según grupo etario y género. Este vector tiene un tamaño de 329.025 elementos correspondiente a cada usuario, el cual tiene la siguiente distribución representada en el Gráfico 8:

Gráfico 8

Visualización gráfica de la distribución de usuario por género, según grupo etario.



Nota. De acuerdo con lo que podemos observar en el gráfico 8, la distribución de la cantidad de usuario por género, según grupo etario.

Luego que se generó la Matriz X y Vector Y, se procedió a realizar el entrenamiento del modelo de Bayes para el primer experimento, haciendo uso de la Matriz X y Vector Y mencionados. Con el modelo entrenado se procedió a probarlo con los datos del archivo de validación y los datos del archivo de prueba obteniendo los resultados de la Tabla 6:

Tabla 6

Detalle del primer experimento

	Primer Experimento	
Métricas	Data Validación	Data de Prueba
F macro	0,19467	0,17609
F micro	0,32074	0,32180
F weighted	0,32365	0,32412

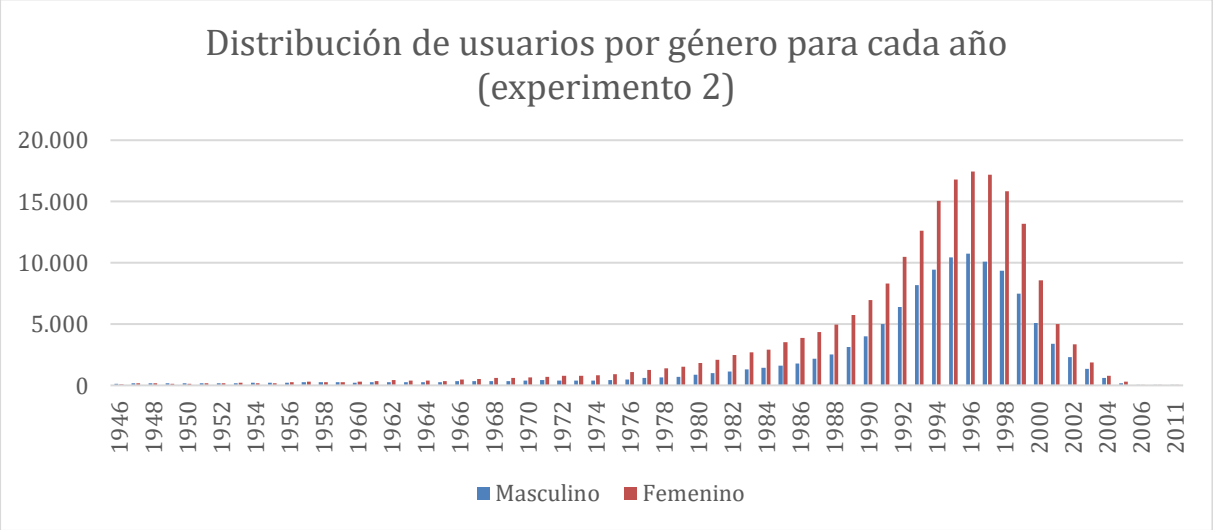
Nota. Se aprecia que en la tabla el detalle del experimento 1 según los datos de validación vs los datos de prueba. (tabla de elaboración propia).

En Tabla 6, para la evaluación del modelo se utilizó el F macro o macropromedio, ya que estamos evaluando identificar a qué clase de año/genero pertenece un usuario, por consiguiente, necesitamos calcular la métrica de forma independiente para cada clase. Para la data de validación el macropromedio fue de 0.1946 y para los datos de prueba fue de 0.176 indicando que el modelo tuvo mejor predicción para los datos de validación que para los datos de prueba.

En el segundo experimento, para el entrenamiento del modelo se generaron batches o lotes de datos correspondiente a la segmentación de los datos según grupo etario y género que se aprecia en la siguiente distribución del Gráfico 9:

Gráfico 9

Distribución de usuario por género para cada año según experimento 2.



Nota. De acuerdo con lo que podemos observar en el gráfico 9, la distribución del número de usuarios (eje y) para cada año de nacimiento agrupado por género (eje x) (gráfico de elaboración propia).

En total se generaron 126 lotes de datos correspondiente a cada año de nacimiento con su respectivo género.

Durante el entrenamiento del modelo se aplicó el algoritmo greedy, en donde se evaluaba cada lote de datos con el archivo de validación, y el que obtuviese mejor F macro se seleccionaba como mejor batch, el cual para una segunda iteración se volvía a evaluar en conjunto con cada lote de datos y se volvía a obtener otro mejor batch que tenía el mejor resultado de la primera iteración más el mejor resultado de la segunda iteración y así sucesivamente hasta alcanzar la convergencia.

Al final, con la iteración número veintiuno se alcanzó la convergencia y se obtuvo una lista de mejores batch o curriculum de entrenamiento, conformado por los siguientes batch y distribución de usuarios, que se representa en la Tabla 7:

Tabla 7

Listado de Batches obtenidos que conforman el curriculum de aprendizaje.

Iteración	Batch	Número de Usuarios	Grupo Etario
1	1981_F	2074	Generación Y
2	1984_M	1413	Generación Y
3	2004_F	776	Generación Z
4	1957_F	305	BabyBoomers
5	1961_M	255	BabyBoomers
6	1980_F	1810	Generación X
7	1980_M	881	Generación X
8	1968_M	356	Generación X
9	1975_M	446	Generación X
10	1965_F	355	Generación X
11	1965_M	263	Generación X
12	1965_M	263	Generación X
13	1966_M	339	Generación X
14	1966_F	471	Generación X
15	1968_M	356	Generación X
16	1970_M	405	Generación X
17	1968_M	356	Generación X
18	2011_M	1	Generación Z
19	2011_M	1	Generación Z
20	2006_M	39	Generación Z
21	2011_M	1	Generación Z

Nota. En la figura podemos visualizar e identificar cada uno de los mejores batch año-genero obtenidos y el grupo etario al que pertenece (figura de elaboración propia).

Tal como se muestra en la Tabla 7, el curriculum obtenido está formado por 21 batches donde el batch con mayores datos de usuario corresponde al batch 1981_F y el batch con el menor datos de usuario corresponde al 2011_M.

El número de batches que pertenecen al género masculino son 15 y al género femenino 6, sin embargo, cuando se totalizan la cantidad de usuarios para los batches, el resultado es 5791 usuarios femeninos y 5375 usuarios masculinos, esta discrepancia se puede explicar a que según los resultados exploratorios realizado sobre los datos inicialmente, la mayor cantidad de usuarios y perfiles que empleaban más palabras pertenecían al género femenino.

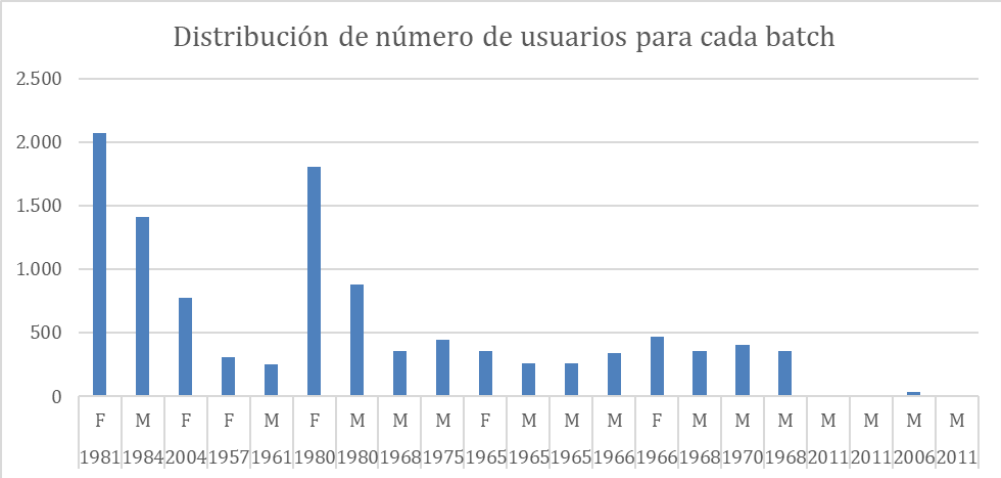
Al analizar los batch por los años y grupo etario, podemos observar que la generación más predominante corresponde a la generación X con 12 batch, luego la generación Z con 5 batch y por último la generación babyboomers y generación Y con 2 batch cada uno. También es posible observar, que para cada grupo etario del batch podemos encontrar batch de ejemplos masculinos y femeninos, es decir, no existe ningún batch perteneciente a un grupo etario que solo tenga un género.

Con respecto a la secuencia de aprendizaje obtenida, el curriculum comienza con ejemplos de la generación Y, generación Z, babyboomers, generación X y luego generación Z nuevamente. Esta secuencia pudiera atribuirse a que, según el análisis previo de los datos, la mayor concentración de los perfiles se encontraba en la generación Y y Z, por consiguiente, el algoritmo inicialmente identificó dichos batch como la mejor opción.

El total de número de usuarios que forman el curriculum es de 11.176, distribuidos, representado en el Gráfico 10:

Gráfico 10

Distribución de número de usuarios para cada batch que conforma el curriculum



Nota. Distribución del número de usuarios para los 21 batch según el año y el género, los cuales corresponden a un total de 11.176 ejemplos o número de usuarios (gráfico de elaboración propia).

Para validar el segundo experimento, al igual que en el primer experimento se utilizó el archivo de validación y el archivo de prueba para generar las métricas. Al comparar ambos resultados, se puede observar en la Tabla 8, que la predicción fue ligeramente mejor en el primer experimento (0,17 vs 0,16) y la cantidad de datos o ejemplos que se necesitó para entrenar cada modelo fue diferente, debido a que para obtener la métrica F macro de 0,17 del primer experimento se necesitó entrenar el modelo con los 329.025 datos de los usuarios, a diferencia del segundo experimento en donde solo se necesitó 11.176 datos de usuarios, es decir, solo con un 3% de los usuarios fue posible obtener casi el mismo F macro de 0,16 , por lo cual queda demostrada la hipótesis planteada, en el cual mediante un plan de entrenamiento de datos o curriculum de aprendizaje se pudo encontrar un subconjunto que generalizó mejor el modelo de datos, lo que permitió establecer una secuencia efectiva de aprendizaje en el contexto de predicción de género y/o grupo etario para los datos de la comunidad de usuarios de Yahoo Answers.

Tabla 8

Comparativa entre el primer y segundo experimento

	Primer Experimento		Segundo Experimento	
Métricas	**Data Validación**	**Data de Prueba**	**Data Validación**	**Data de Prueba**
F macro	0,19467	0,17609	0,17028	0,16088
F micro	0,32074	0,32180	0,35655	0,35567
F weighted	0,23657	0,32412	0,33098	0,33199

Nota. Se aprecia que en la tabla el detalle del primer y segundo experimento, según los datos de validación vs los datos de prueba de cada experimento. (tabla de elaboración propia)

7 CONCLUSIONES

7.1 Basado en los resultados obtenidos en el análisis de muestra de datos exploratorios:

a) El 63% de los perfiles de Yahoo Answers corresponden a perfiles de género femenino y un 37% a perfiles de género masculino, es decir, la mayoría de los usuarios registrados activos en la plataforma corresponde al género femenino.

b) Existe una similitud entre la curva del género masculino y femenino con los puntos representados para los años; donde se evidencia que existe una similitud en la distribución del género.

c) Las palabras más utilizadas por ambos géneros son: “Girl”, “People”, “Good”, “Work”, “Way”, “See”, “Need”, “Thanks”, sin embargo, cuando se observa la distribución de estas palabras se consiguen diferencias, por ejemplo, en el grupo etario “Silente”, los años menores al 1946 usa más la palabra “Dios” en comparación con los otros grupos etarios, esto puede deberse a que es un grupo con bases y valores religiosas diferentes que otros grupos de edades.

7.2 Basado en la metodología de trabajo

Considerando la fuente de datos entregada por el profesor guía, datos de usuarios de Yahoo Answers, se alcanzaron todas las etapas metodológicas para el proyecto:

a) **La obtención de los datos:** En esta etapa del proyecto se han obtenido directamente del profesor guía y que han sido considerado para el desarrollo conceptual de lo que se busca lograr con el proyecto. Del conjunto de datos se aplicó la norma 60, 20, 20, en el cual se generó un archivo de entrenamiento con el 60 % de los datos y un archivo de validación de datos y un archivo de prueba con el 20% de los datos cada uno. En total se utilizaron

los datos de 548.377 usuarios, 329.025 para el archivo de entrenamiento y 109.676 para cada uno de los archivos restantes.

b) **El preprocesamiento:** utilizando el lenguaje Python se ha logrado analizar los BoW de cada archivo permitiendo obtener para efectos del entrenamiento un diccionario de datos con solo la frecuencia de palabras mayores o iguales a 5 y limpieza de signos de puntuación y stopwords.

c) **Características:** Como resultado del análisis, se logró identificar las características de los datos, en este caso, los usuarios por géneros y rango etario, clasificándolos en grupos etarios conformados por generación "Silenciosa" los años menores al 1946, "Baby boomers" entre los años 1946 al 1964, la "Generación X" entre los años 1965 a 1980, la "Generación Y" entre los años 1981 a 1996 y la "Generación Z" entre los años 1997 al 2015.

d) **Entrenamiento:** Se realizó el entrenamiento de un modelo haciendo uso de las redes de Bayes. Este se realizó mediante 2 experimentos diferentes, el experimento 1, se utilizó la data de todos los usuarios del archivo de entrenamiento (329.025 usuarios) y en el experimento 2, se hizo uso de un subconjunto de datos (11.176 usuarios) correspondiente al curriculum de entrenamiento generado durante el experimento 2.

e) **Evaluación:** Para evaluar y determinar la efectividad de los modelos se utilizó como métrica el F macro o macropromedio, ya que, se necesitaba identificar a qué clase de año/genero pertenece un usuario, por consiguiente, se requería calcular la métrica de forma independiente para cada clase. Los datos que se utilizaron para la evaluación corresponden al archivo de validación y archivo de prueba

f) **Resultados:** Para el experimento 1 el modelo obtuvo un F macro de 0,17 usando 329.025 datos de usuario, para el experimento 2 el modelo se obtuvo un F macro de 0,16 usando 11.176 datos de usuario. El modelo del experimento 1 tuvo un resultado similar al modelo

del experimento 2, sin embargo, el modelo del experimento 2 solo requirió usar un 3% de los datos por lo que fue más eficiente que el modelo del experimento 1.

7.3 Demostración de la hipótesis

En el experimento 2, se generó un plan de entrenamiento de datos o curriculum de aprendizaje que permitió generalizó mejor el modelo de datos, lo que permitió establecer una secuencia efectiva de aprendizaje en el contexto de predicción de género y/o grupo etario para los datos de la comunidad de usuarios de Yahoo Answers. Para la generación de este curriculum se utilizó el algoritmo greedy durante la fase de entrenamiento y mediante 21 interacciones se alcanzó la convergencia, teniendo como resultado un curriculum formado por 21 batch o lotes de datos que representan un 3% (11.176) del total de usuarios de entrenamiento (329.025).

8 BIBLIOGRAFÍA

Statist Digital Economy Compass. (17 de Abril de 2019). Statista. Obtenido de Statista: **https://es.statista.com/grafico/17734/cantidad-real-y-prevista-de-datos-generados-en-todo-el-mundo/**

Bayot, R. a. (2016). Author Profiling using SVMs and Word Embedding Averages. *CLEF 2016* (págs. 1-9). Portugal: CLEF.

Beel, J. L. (2013). The Impact of Demographics (Age and Gender) and Other User-Characteristics on Evaluating Recommender Systems. *Springer-Verlag Berling Heidelberg*, 400-404.

Bonnin, R. (11 de Juliio de 2017). *Machine Learning Projects: A Step by Step Approach*. Obtenido de Machine Learning: **https://machinelearning.technicacuriosa.com/2017/07/11/machine-learning-projects-step-step-approach/**

Brouoadmin. (11 de Abril de 2016). *El Instituto de investigaciones Sociales*. Obtenido de Comportamiento Humano: Williamm Strauss y Neil Howe: **https://www.iisociales.mx/comportamiento-humano-william-strauss-y-neil-howe-son-grandes-influenciadores-de-esta-decada/**

Brownlee, J. (15 de Agosto de 2020). *Machine Learning Mastery.* Obtenido de Machine Learning Mastery: **https://machinelearningmastery.com/naive-bayes-for-machine-learning/**

Chen J., X. T. (2017). Gender Prediction on a real life blog data set using LSI and KNN. *2017 IEEE 7th Annual Computing and Communication Workshop and Conferene (CCWC)* (págs. 1-10). Las Vegas, NV, USA: IEEE.

Creative Commons Atribucion Compartir. (23 de Febrero de 2021). *wikipedia.org/Algoritmo_voraz.* Obtenido de wikipedia.org/Algoritmo_voraz: **https://es.wikipedia.org/wiki/Algoritmo_voraz#:~:text=En%20ciencias%20de%20la**

%20computaci%C3%B3n%2C%20un%20algoritmo%20voraz,la%20hora%20de%20dise%C3%B1ar%20y%20comprobar%20su%20funcionamiento.

Emmery, C. C. (2017). Simpel Queries as Distant Labels for Predicting Gender on Twitter. *Association for Computational Linguistics*, 50-55.

Facebook. (04 de Febrero de 2021). *Facebook*. Obtenido de Facebook: **https://www.facebook.com**

Figueroa, A. (2017). Male o Female: What traits characterize questions prompted by each gender in community question answering? *El Savier, Expert Systems with Applications*, 405-413.

ICEMD. (01 de 01 de 2019). *Robotica: Una revolución que impacta en todos los sectores e industrias*. Obtenido de Robotica: Una revolución que impacta en todos los sectores e industrias: Percentage of U.S. Baby Boomers who use selected social networks as of February 2019

KASASA. (13 de Enero de 2021). *Community Rising Blog*. Obtenido de Boomers, Gen X, Gen Y, and Gen Z Explained: **https://www.kasasa.com/articles/generations/gen-x-gen-y-gen-z**

Morrison, P. a.-H. (2013). Is Programming Knowledge Related to Age? *Tenth International Workshop on Mining Software Repositories (MSR)* (págs. 69-72). San Francisco, USA: IEEE.

NLTK Project. (20 de Abril de 2021). *NLTK 3.6 Documentation*. Obtenido de NLTK: **https://www.nltk.org/**

Onur Kucuktunc, I. W. (2012). A Large-Scale Sentiment Analysis for Yahoo! Answers. *Fifth International Conference on Web Search and Web Data Mining, WSDM2012.* (págs. 1-10). Seatle, WA. USA: DBLP.

Schwartz HA, E. J. (25 de Septiembre de 2013). *Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach.* Obtenido de Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791

scikit-learng. (Abril de 2021). *Scikit learn.* Obtenido de Scikit learn: **https://scikit-learn.org/stable/index.html**

Stack Overflow. (4 de Febrero de 2021). *stackoverflow/Questions*. Obtenido de stackoverflow/Questions: **https://stackoverflow.com/questions**

Statista. (7 de Febrero de 2019). *Percentage of U.S. Baby Boomers who use selected social networks as of February 2019*. Obtenido de Percentage of U.S. Baby Boomers who use selected social networks as of February 2019: **https://www.statista.com/statistics/436417/us-baby-boomer-selected-social-networks/**

Steven Bird, E. K. (2008). Multidisciplinary Instruction with the Natural Language Toolkit. *Association for Computational Linguistics*, 62-70.

Webis Group. (2021). *PAN*. Obtenido de PAN: https://pan.webis.de/

Yahoo. (4 de Enero de 2021). *Yahoo Answers*. Obtenido de Yahoo Answers: **https://answers.yahoo.com**

Young-Sik, J. T. (2014). Comparative Study of Human Age Estimation with or without Prelassifiation of Gender and Facial Expression. *Publishing Corporation The Scientific World Journal*, 15.

Zion & Zion. (19 de Junio de 2018). *From Boomers To Generation Z: The Challenge of Generational Labelling*. Obtenido de From Boomers To Generation Z: The Challenge of Generational Labelling: **https://www.zionandzion.com/from-boomers-to-generation-z-the-challenge-of-generational-labeling/**