**Proyecto Data Science**

[**1. Presentación de la empresa, organización o problema específico.**](#_ynrc9xjn3mf1) **2**

[**2. Preguntas y objetivos de la investigación.**](#_gtr7ps313bln) **2**

[**3. Conformación del equipo de trabajo.**](#_pud0a9vkr8js) **3**

[**4. Indicación de la fuente del dataset y los criterios de selección (Data Acquisition).**](#_q619j6hxu4nl) **3**

[**5. Data Wrangling y EDA**](#_1e0zf42hgyeo) **4**

[**5.1 Filtros aplicados a los datos. Distribución. Dataset final para analizar**](#_au3s747w0tsn) **4**

[**5.2 Análisis Univariado**](#_qqz1a69v9ewk) **5**

[5.2.1 Pie Chart porcentajes overall](#_nhtx90ofx1w4) 5

[5.2.2 Cantidad de sentimientos negativos y positivos](#_f6jrgul5f8be) 6

[5.2.3 Ranking formatos musicales más vendidos](#_zgudjdx7frab) 6

[**5.3 Análisis Bivariado**](#_5l5z7my0c131) **7**

[5.3.1 Análisis de Componentes Principales](#_d82s2y4ur5cr) 7

[**5.3 Análisis Multivariado**](#_x5mx0g7odqcj) **10**

[5.4 Elección Modelo Algoritmo Entrenamiento](#_lcw8nv6o9e1t) 11

[5.5 Balanceo/Optimizaciones del modelo](#_20zpq9mzjcg5) 11

[5.6 Futuras Líneas](#_id15aos16wc) 13

[5.7 Conclusiones](#_1mru2mdmx6er) 14

# **1. Presentación de la empresa, organización o problema específico.**

El dataset a analizar es una base de productos de formato musical vendidos en Amazon donde se buscará , a través de los comentarios dejados por los sus compradores, identificar el sentimiento generado. El dataset fue seleccionado con la idea de acercarse y presentar los primeros pasos necesarios para la creación de una herramienta que modere las dificultades de interpretación emocional de textos escritos, de algunas personas dentro del Trastorno del espectro autista (TEA).

Según los criterios definidos por la Asociación Estadounidense de Psiquiatría (2020) en el Manual diagnóstico y estadístico de los trastornos mentales, una persona que vive dentro del espectro: posee deficiencias persistentes en la comunicación social y en la interacción social, lo que incluye: deficiencias en la reciprocidad socioemocional, utilización de conductas comunicativas no verbales utilizadas en la interacción social y el desarrollo, mantenimiento y comprensión de las relaciones sociales.

Si bien el producto final presentado en el proyecto del curso Data Science de CoderHouse va estar centrado en el análisis del dataset de Amazon, la idea es que con este producto se pueda acercar a en un nivel más avanzado poder realizar una aplicación de estos conocimientos para la creación de una herramienta que sea a las personas dentro del espectro.

# **2. Preguntas y objetivos de la investigación.**

**Objetivo:** Crear un modelo de Machine Learning capaz de predecir la emoción del usuario de compra de productos musicales en Amazon, por medio del comentario (review) que escribe en la plataforma.

**Preguntas:**

1. ¿Cuáles son las variables que determinan la categoría emocional en el modelo?
2. ¿Cómo se pueden distinguir emociones a partir de un texto utilizando un modelo de machine learning?

# **3. Conformación del equipo de trabajo.**

**Participantes:** Diana Paola Almeida Luna y Gabriel Eduardo Almeida Luna.

# **4. Indicación de la fuente del dataset y los criterios de selección (Data Acquisition).**

El dataset cuenta con 1.584.082 de registros, los campos seleccionados para realizar el análisis y por tanto el modelo fueron los siguientes:

* **Overall:** Puntuación de 1 a 5 que brinda el usuario según su satisfacción con el producto, siendo 1 el más bajo y 5 el puntaje más alto.
* **Verified:** Campo true/false que indica si la cuenta que realiza el comentario es o no verificada
* **review\_time:** Fecha en la que se generó el comentario y puntaje del producto.
* **reviewer\_id**: Documento hasheado del cliente.
* **review\_text**: Comentario completo sobre el producto adquirido.
* **summary**: Comentario resumido sobre el producto adquirido.
* **style.format:** Formato del contenido musical comprado.
* **vote:** Cantidad de votos sobre el comentario realizado.
* **sentiment:** Campo creado según el overall, donde se determina 1(sentimiento positivo) si es mayor a 3 el puntaje , y 0 (sentimiento negativo) si es menor o igual a 3.
* **review\_length:** Campo creado en donde se cuenta el número de caracteres del comentario a fin de determinar si existe alguna relación desde la longitud del mismo con el sentimiento que genera en el usuario.

Los criterios utilizados para la selección de los sets de datos se basaron en la cantidad de campos, a fin de evaluar diferentes variables y sus relaciones, así como la cantidad de registros. Esto último, a fin de contar con un diligente entrenamiento del modelo y esperar por un nivel de precisión más exacto.

*Enlace dataset Digital Music Amazon modificado:*<https://drive.google.com/file/d/1tNfteg9zZvlcwmOQFDPkxMrBmWtnwN5A/view?usp=sharing>

*Enlace al dataset original sin modificaciones:*

[Amazon review data (nijianmo.github.io)](https://nijianmo.github.io/amazon/index.html)

# **5. Data Wrangling y EDA**

Como se ha explicado con anterioridad, el objetivo primordial es identificar la valencia emocional (positiva o negativa) de una reseña escrita en la plataforma Amazon, lo cual está estrechamente vinculado al análisis a través de las diferentes variables mencionadas en el inciso ([4. Indicación de la fuente del dataset y los criterios de selección (Data Acquisition).](#_q619j6hxu4nl)

***Amazon Product Data- Digital Music:*** El dataset contiene 10 columnas: emoción, overall, verified, reviewTime, reviewerID, summary, style.Format:, vote, sentiment y review\_lenght.

## **5.1 Filtros aplicados a los datos. Distribución. Dataset final para analizar**

**Los cambios realizados en el datasets fueron:**

1. A través del siguiente código se normalizan las columnas, lo que quiere decir que como la columna de styleFormat: poseía varias subdivisiones de columnas dentro, las separa en columnas distintas:

data2 = pd.json\_normalize(data)

1. Como se mencionó anteriormente, se hizo una modificación en el dataset para crear la nueva columna titulada *‘sentiment’* la cual asigna 1 a los comentarios positivos (asumidos como puntajes por encima de tres en el *‘overall’*) y 0 a los comentarios negativos (asumidos como puntajes por debajo de tres en el *‘overall’*).
   1. df['sentiment'] = np.where(df['overall']>3, '1', '0')

df.head()

1. Se elimina las columnas ‘asin’, ‘unixReviewTime’, ‘image’, ‘reviewerName’, ‘style.Color:’ y ‘style.Size:’:
   1. df = df.drop(['asin', 'unixReviewTime', 'image', 'reviewerName', 'style.Color:', 'style.Size:'], axis=1)
2. Se cambia el tipo de dato de reviewTime a datetime
   1. df['reviewTime'] = df['reviewTime'].str.replace(' ', ',')

df['reviewTime'] = df['reviewTime'].str.replace(',,', ',')

df['reviewTime'] = df['reviewTime'].str.replace(',', '-')

df['reviewTime'] = pd.to\_datetime(df['reviewTime'])

1. En la columna ‘vote’ se reemplazan los campos vacíos con 0, para que represente numéricamente la cantidad de votos de cada comentario del producto.
   1. df['vote'] = df['vote'].fillna(0)
2. Se cambia la columna de ‘vote’ que es interpretado por la máquina como tipo string/object por la coma de miles a float y luego a int. de string por la coma de miles a float y luego a int
   1. df['vote'] = df['vote'].str.replace(',', '').astype(float)

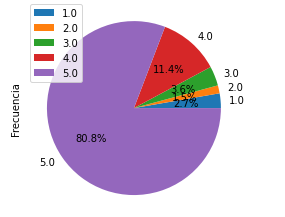
df['vote'] = df['vote'].astype(int)

1. Cambiar review\_length a int
   1. df['review\_length'] = df['review\_length'].astype(int)
2. Cambiar overall de float a int
   1. df['overall'] = df['overall'].astype(int)
3. Se cuentan la cantidad de nulos y se eliminan las filas que poseen estos datos vacíos.
   1. df['reviewTime'].isna().sum()

df = df.dropna()

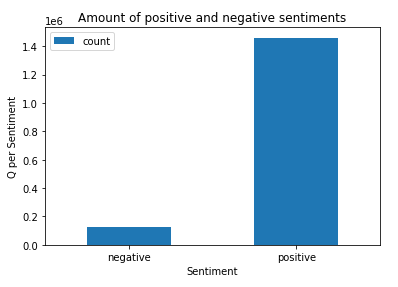
## **5.2 Análisis Univariado**

### **5.2.1 Pie Chart porcentajes overall**

****

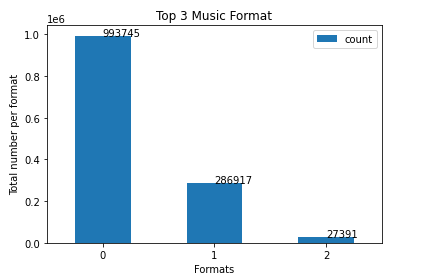
Se encuentra con que alrededor del 80% de los productos de música digital en Amazon tienen puntuaciones de 5.0 estrellas. El 20% restante está dividido en las puntuaciones inferiores; de manera descendente a medida que baja la puntuación: 4.0 (11.4%). 3.0 (3.6%), 2.0 (1.5%), 1.0 (2.7%).

### **5.2.2 Cantidad de sentimientos negativos y positivos**

****

En cuanto a la distribución de la valencia (positiva o negativa) de los comentarios, se muestra que la mayoría de los comentarios son positivos, solo 124.006 de ellos son negativos.

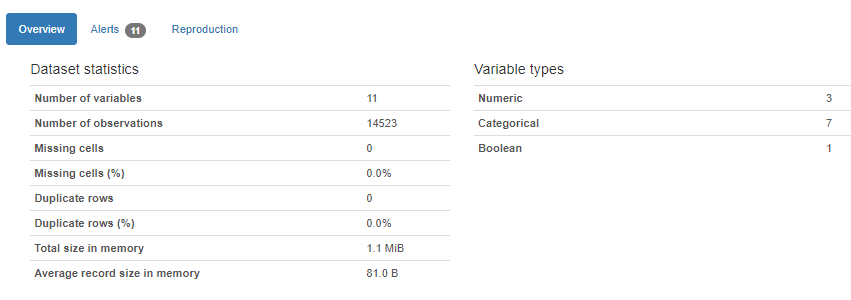
### **5.2.3 Ranking formatos musicales más vendidos**

****

Finalmente, el formato de música con más reseñas es Audio CD, seguido del MP3 y finalmente Vinilo. Los otros 6 formatos tienden a no ser representativos ni comparables con la cantidad de comentarios que presentan los 3 primeros.

## **5.3 Análisis Bivariado**

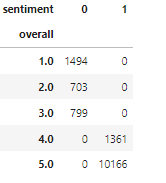
### **5.3.1 Análisis de Componentes Principales**



Se verifica que no se obtengan filas duplicadas como parte del proceso de limpieza, y el tipo de cada variable se observa para contemplar en las futuras evaluaciones de correlaciones y pruebas de hipótesis nula y alternativa.



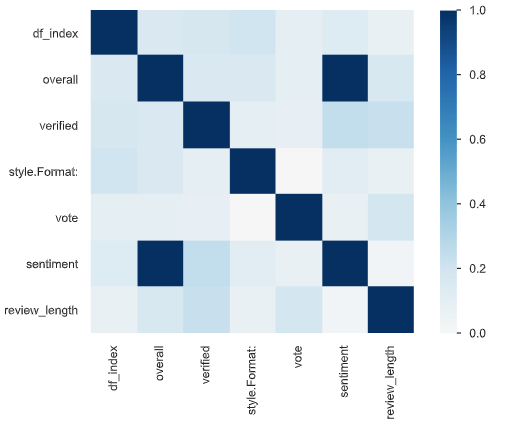
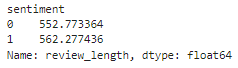
Claramente por la definición que se le dio al campo ‘sentiment’ se observa que es la de mayor correlación con overall.

****

Se confirma que la división de sentimiento positivo(1) negativo (0) fue hecha de manera correcta, ya que se ve una media para el sentimiento negativo de 1.76 y para un sentimiento positivo de 4.8. Adicionalmente, se evidencia la dependencia entre la variable de sentimiento y el overall, por la definición que se le dio al campo *‘sentiment’* desde su creación.

**Anotación importante:** El único *filtro* aplicado se utiliza ya que la biblioteca de *pandas\_profilling* no permitió que se corrieran el total de los datos. Por lo cual, únicamente para fines prácticos, este apartado se filtró por un año, ya que es la cantidad máxima de datos que permite hacer el análisis con la cantidad de RAM a disposición:

df = df[df['reviewTime'] >= '01-01-2016']

****

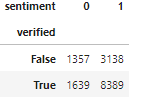
No existe ninguna relación significativa entre la cantidad de caracteres del comentario y el sentimiento generado.

Se crea una tabla para comparar dos variables categóricas (style.Format: y sentiment), de la cual se obtiene el siguiente p-value: 3.6519621324845776e-28. No existe evidencia suficiente para afirmar que hay independencia entre el formato y el sentimiento. Por lo cual, si existe una dependencia entre el formato y el sentimiento, siendo el top 3 : CD.MP3 y Vinyl.

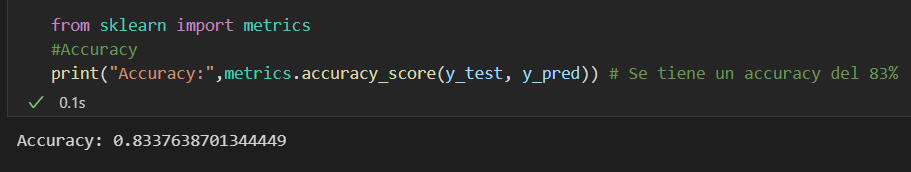


Se prosigue a hacer el mismo proceso con las variables verified y sentiment, que luego de realizar la prueba de χ²chi y obtener un p-value de 8.038324562939505e-81, se rechaza la Ho, por lo cual si existe una dependencia entre el comentario verificado y el sentimiento

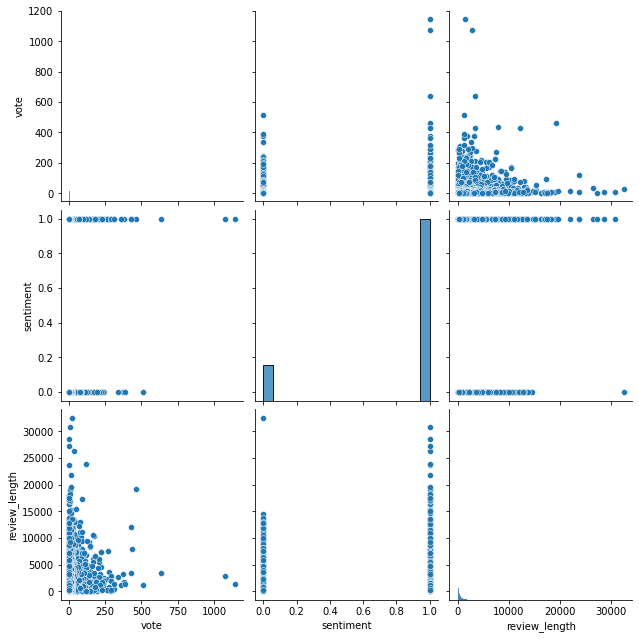




## **5.3 Análisis Multivariado**

Para este análisis se intenta hacer una predicción del sentimiento o overall según verified ,vote, styleformat y reviewTime. Se predice que el modelo puede identificar con un accuracy del 83% si un comentario es positivo o negativo:  


Si se hace un PairPlot se encuentra que entre más carácteres menos votos recibe el producto comprado. Las otras dos gráficas no muestran una correlación significativa para ser analizada.

****

## **5.4 Elección Modelo Algoritmo Entrenamiento**

Se tuvo la intención de probar dos modelos diferentes, el de SVM y el de BERT a fin de comparar los resultados de cada uno después de realizar el balanceo a fin de correr los dos modelos y evaluar su eficiencia.

**5.4.1 Métricas de Desempeño**

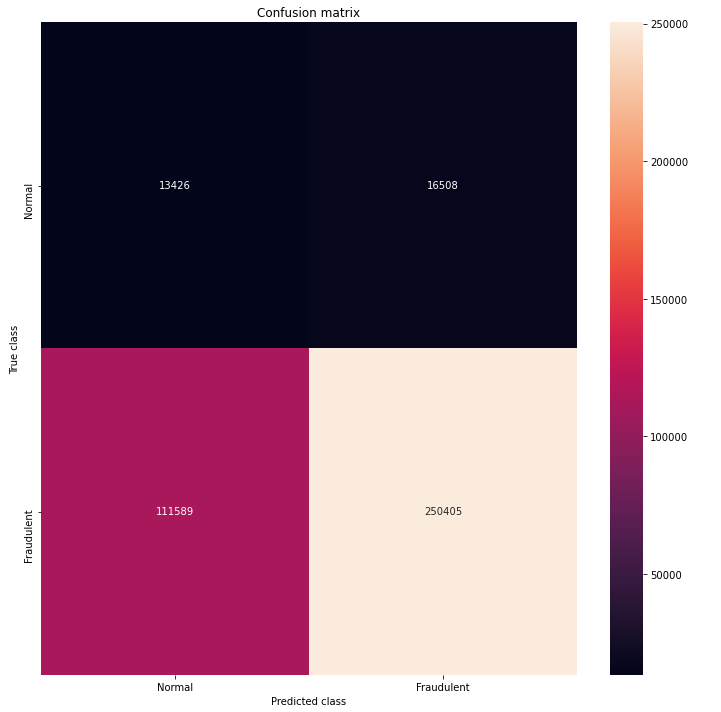
En el caso del modelo SVM se presentó la situación de que luego de realizar el balanceo con la técnica de SMOTE, los datos incrementaron y al correr el modelo no terminaba de realizarse la ejecución pero tampoco enviaba algún tipo de error.

En el caso del modelo BERT sucede la misma situación. A pesar de que se intentó reducir el número de datos con un Undersampling, no se pudo ajustar el modelo con este nuevo balanceo. El modelo corre hasta un punto el cuál está señalado en el código del modelo en mayúsculas, pero no se obtiene un resultado del accuracy o las métricas de desempeño por esta razón.

## **5.5 Balanceo/Optimizaciones del modelo**

En nuestro caso tenemos un desequilibrio de clases , donde el escenario presentado surge al tener una distribución desigual de clases en un conjunto de datos, es decir, el número de puntos de datos en la clase negativa (clase minoritaria) es muy pequeño en comparación con la clase positiva(clase mayoritaria).

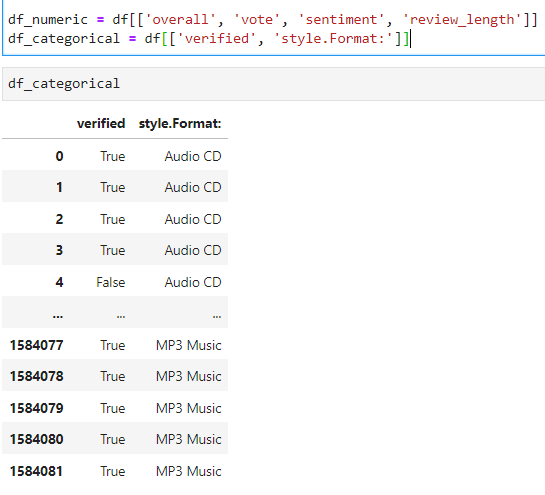
Si los datos desequilibrados no se tratan de antemano, esto degradará el rendimiento del modelo de clasificación y por tanto se caerá en un sesgo en el modelo. Por lo anteriormente explicado, se procede a trabajar el desequilibrio de datos con Smote, la cual permite crear nuevas observaciones sintéticas utilizando las muestras existentes de la clase minoritaria. De este modo, se procede a crear registros de entrenamiento virtual por interpolación lineal para la clase minoritaria, donde los datos se reconstruyen y se aplica el modelo SVM para los datos procesados.



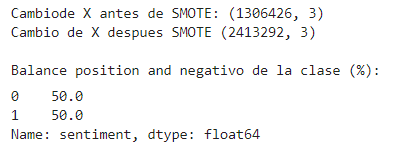
Con esta distribución de la matriz podemos ver lo sesgada que está nuestra muestra al tener tantos valores positivos:

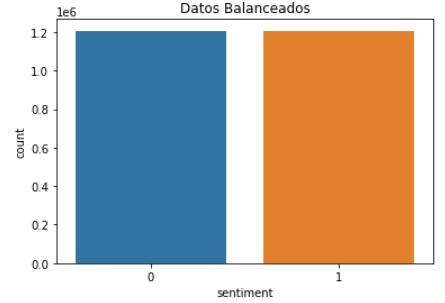


En primera instancia fue necesario hacer una transformación de variables nominales a numéricas.



Se aplica el balanceo resultando en lo siguiente:



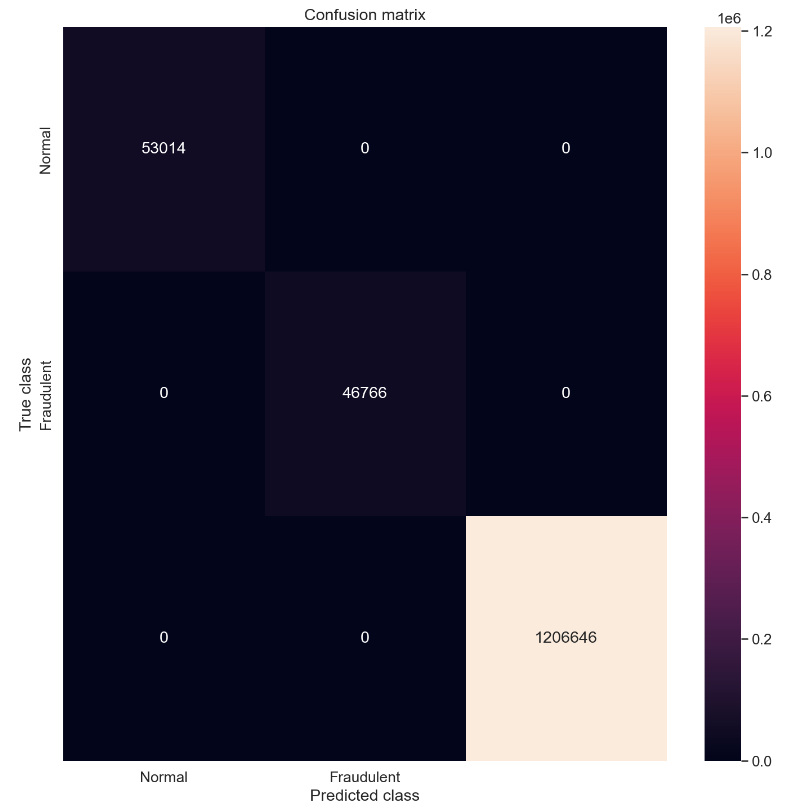


Luego de garantizar una disminución en el sesgo, se procede a ejecutar el modelo de SVM.

## En el caso del modelo BERT se procede a hacer un balanceo pero con un Undersampling para intentar disminuir la cantidad de datos y que el modelo procese mejor. En las siguientes dos líneas se encuentra lo desbalanceada que se encontraba la muestra:



## La nueva matriz de confusión después del undersampling es la siguiente:



No se encuentran datos fraudulentos ya que se pasó la columna sentiment que establece la división clara entre sentimientos: positivos, neutros y negativos. La razón de la realización del undersampling no es para la clasificación, sino para la reducción de los datos para luego utilizarla en el modelo BERT.

## **5.6 Futuras Líneas**

El propósito principal del proyecto era intentar acercarse a lo que podría ser la interpretación de los sentimientos escritos o en audio de diferentes personas por medio de un algoritmo. Esto en búsqueda de encontrar una posibilidad de realizar una aplicación que se le brinde a las personas con autismo con un alto grado de disfuncionalidad. De esta manera, estas personas pueden tener una herramienta que les permite entender el lenguaje pragmático, lo que es en muchos casos la problemática central de adaptación que tienen las personas con autismo. Con este proyecto, se pudo acercar a la posibilidad de la elaboración de un algoritmo que pueda ser aplicado en futuras líneas de investigación o innovación para personas que les interese realizar una iniciativa tecnológica para ayudar a personas con algún grado de dificultad o discapacidad del reconocimiento o el entendimiento del lenguaje pragmático.

## **5.7 Conclusiones**

A través del trabajo realizado una de las conclusiones más interesantes que se obtuvieron fue que no existe ninguna relación significativa entre la cantidad de caracteres del comentario y el sentimiento generado. Se podría pensar que si existiese alguna queja podría tomar mayor cantidad de carácteres y por tanto asumirlo como la generación de un sentimiento negativo, sin embargo, luego de realizar los análisis de correlación se descartó por completo dicha hipótesis.

Una de las grandes conclusiones para este tipo de trabajo es intentar empezar con muestras pequeñas, manejables, que permitan correr los modelos y algoritmos, así se ponga en riesgo la confiabilidad del modelo. Sin esta información, no es posible ofrecer conclusiones claves y cruciales para determinar la eficacia de los modelos y de lo realizado en los primeros pasos de la construcción del algorítmo.