Proyecto Integrador

Estudio de relacionamiento de usuarios digitales con la marca "Avianca" mediante el uso de analítica de texto y análisis de comunidades utilizando como fuente de información Twitter.

Integrantes:

Jesús Alberto Arcia Hernández, jaarciah@eafit.edu.co.

Edgar Leandro Jiménez Jaimes, eljimenezj@eafit.edu.co.

Jorge Luis Rentería Roa, jlrenteria@eafit.edu.co.

Maestría en ciencia de los datos y analítica.

UNIVERSIDAD EAFIT.

Octubre de 2019.

Tabla de contenido

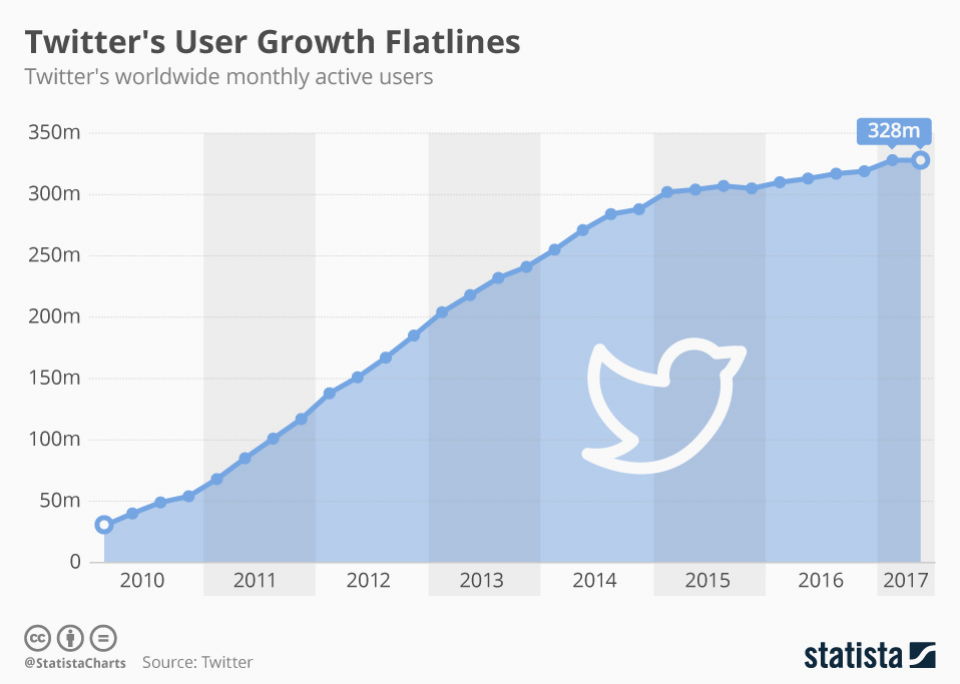
CONTEXTO Y DESCRIPCIÓN DEL PROBLEMA

El estudio del Social Media Data Mining y el estudio de comunidades ha venido tomando cada vez más fuerza gracias a la creciente cantidad de datos que hay en internet y la cantidad de usuarios que interactúan en ella día a día.

Por medio del Social Data Mining se abordan varios conceptos importantes hoy por hoy para las organizaciones, como lo son la **publicidad y el marketing,** donde por medio minería de datos a estas plataformas puede ser de gran ayuda para determinar que estrategia de publicidad funciona mejor para los usuarios y asegurarse que el público objetivo esté satisfecho, otro concepto a resaltar es la **lealtad de marca,** por medio del constante monitoreo a las plataformas de la compañía, esta brinda a sus usuarios y clientes una sensación de ser escuchado y entendido, lo cual es un factor importante que permite la construcción de una base sólida de clientes leales y satisfechos, por otro lado, el **desarrollo de productos,** donde el negocio podrá hacer frente al ritmo cambiante de los usuarios hacia otros productos u marcar de manera rápida mediante el desarrollo de nuevos productos o servicios que permitan capturar más audiencia o consolidar la que tiene. la **comunicación,** a través de la minería de datos permite acercar más a los clientes y los proveedores del negocio, monitoreando las redes sociales en busca de clientes insatisfechos o que no han sido atendidos de manera satisfactoria y dirigir la fuerza de atención al servicio hacia estas personas, logrando una mejor comunicación entre el negocio y sus stakeholders (Dataworks, 2018).

Es así como en este trabajo se abordan los conceptos del Social Media Data Mining y el estudio de comunidades. En primer lugar, por medio del estudio y procesamiento del lenguaje natural en las redes sociales, particularmente en Twitter. Esta red que fue creada en el 2006 inicialmente solo permitía escribir hasta 180

caracteres, hoy en día es permitido hasta 280 caracteres. Esta particularidad de restricción en la extensión de la escritura hace que los usuarios sinteticen sus ideas y utilicen un lenguaje más compacto para poder comunicarse con los otros actores. Twitter se encuentra presenten en los cinco continentes y tienen millones de usuarios que día a día comparten sus ideas, proyectos, pensamientos, críticas y opiniones frente a un tema particular o una marca.



Grafica 1 Evolución de usuarios activos año a año.

La grafica 1 presenta la evolución creciente que ha tenido el número de usuarios activos en Twitter hasta el año 2017 que se encuentra actualizada esta gráfica, se observa que es un crecimiento brusco en los años 2011 a 2015 y que en los últimos 2 a 3 años es un crecimiento un poco más lento y estable. A continuación, se presentan algunos datos obtenidos por Twitter en cuanto a su uso por parte de los usuarios y algunas estadísticas:

Gráfico 2: Estadística y datos de uso de Twitter

Como se puede observar en el grafico 2, aunque la información se encuentre actualizada hasta 2016, la cantidad de usuarios a hoy se estima en valores superiores a los 400 millones, con un volumen inmenso de información principalmente texto e imágenes que contienen las opiniones de los usuarios frente a diversos temas de índole político, social, cultural, etc. Toda esta información puede ser utilizada por las compañías, mediante procesos de minería de datos y analítica de texto, se pueden limpiar los textos, estudiarlos mediante algunas técnicas de PNL y así obtener información valiosa día a día, incluso en tiempo real de lo que está pasando en el medio y lo que la gente está opinando respecto a la empresa.

Así entonces, se pretende estudiar la relación que tienen los usuarios con una marca en particular mediante la manera en que interactúan en su red social y lo que escriben acerca de esta. Paralelamente se estudia los perfiles de las personas para determinar cuan influyentes pueden llegar a ser sus publicaciones con base a una medida propia que considera variables alrededor del conocimiento de la marca, el número de seguidores que posee y el volumen total de usuarios que interactúan con

la marca. Este estudio pretende comprender la relación de los usuarios más allá de solo las frases que se expresan en Twitter y entender su perfil como otro factor de información en la interacción con la marca en la red social.

Por otra parte, el análisis de comunidades pretende identificar las personas con quienes interactúa aquella persona que realiza una publicación, esto mediante el monitoreo de los Id de los usuarios y sus seguidores. Esto se pretende abordar utilizando un análisis de grafos utilizando el software de visualización Power Bi. Por otro lado, el estudio de comunidades permite conocer las personas con mayor número de seguidores, a quien se denomina influencer y cómo su comunidad reacciona a las opiniones generadas, buscando identificar comunidades negativas, mixtas o positivas.

METODOLOGIA

La normalización del proceso de hallar conocimiento en los datos ha convocado esfuerzo desde finales de los años 90, durante este tiempo se han planteado metodologías que buscan alinear la ejecución de los proyectos analíticos con los objetivos del negocio sin descuidar los elementos técnicos que deben tenerse en cuenta. Dentro de esas metodologías se encuentra ASUM – DM planteada por IBM la cuál por medio de cinco categorías busca abordar de manera holística la salida a producción de un sistema analítico partiendo desde la necesidad del negocio (Angée, Lozano-Argel, Montoya-Munera, Ospina-Arango, & Tabares-Betancur, 2018).

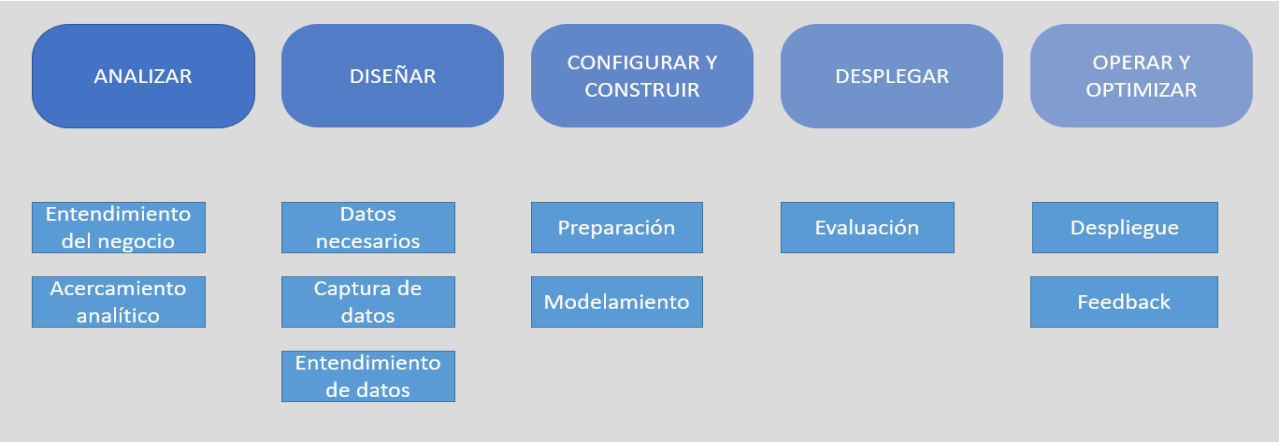


Gráfico 3: metodología ASUM – DM.

ENTENDIMIENTO DEL NEGOCIO:

En un ambiente de competencia abierta y mercados dinámicos estar a la vanguardia de las tendencias que rigen el mercado es una obligación para todas aquellas empresas que buscan sobrevivir, por ellos, conocer a sus clientes de manera integral, sus sentimientos y las comunidades con quienes interactúan son herramientas útiles de decisión para las marcas, esto junto con el uso masivo que han obtenido actualmente las redes sociales para expresar opiniones y sensaciones hacia

una empresa llevan a que redes sociales como Twitter deban ser analizadas empleando métodos analíticos que permitan conocer constantemente el sentimiento que tienen las personas sobre una marca específica y el impacto que tiene el punto de vista de su comunidad digital sobre su opinión virtual.

ACERCAMIENTO ANALÍTICO:

A través de un modelo analítico se logra asignar una clasificación al sentimiento expresado en el tweet de cada usuario con base a cuán positivas, neutras o negativas son las palabras que se hayan empleado, además de identificar los clústers de relaciones donde se indican con quienes interactúan en el día a día.

DATOS NECESARIOS:

Durante el desarrollo del modelo, se identificaron las fuentes integrales que conectan de principio a fin la opinión virtual de un usuario con sus partes relacionadas. Para esto, se requiere:

- **Fecha de creación**: Fecha de creación del tweet.
- **Texto tweets**: Texto crudo del tweet.
- **Cantidad de seguidores**: número de personas que siguen a quien publicó el tweet.
- **Cantidad de seguidos**: número de personas seguidas por quien publicó el tweet.
- **Sentimiento del tweet:** Calificación del tweet como positivo, neutro o negativo.
- **Palabras claves**: lista de palabras categorizadas por su grado de “positividad”, “neutralidad” o “negatividad”.

- **Ranking influencia**: valor de influencia del usuario frente a la marca:

$$APP = \frac{Seguidores\ unicos * conocimiento\ de\ marca}{total\ unico\ de\ seguidores}$$

CAPTURA DE DATOS:

Para obtener la información de Tweets se utiliza el servicio expuesto por Twitter inc. (https://developer.twitter.com/) utilizando el lenguaje de programación Python. Por otro lado, para nutrir el conjunto de palabras positivas, neutras o negativas se obtiene información del primer nivel, esto quiere decir, que se hará uso de la librería "***tweepy***" para descargar la información con la siguiente estructura:

Fecha Creación	Id tweet	Id usuario	Nombre usuario	Numero seguidores	Numero Seguidos	Texto tweet	Ubicación
…	…	…	…	…	…	…	…

Actualmente, la mayoría de modelos que permiten analizar sentimientos se encuentran correctamente entrenados para otras lenguas diferentes al español, por ellos, se contactaron 25 personas entre edades de 23 a 32 años, con un nivel de estudio universitario para que con base a su criterio calificaran el texto del tweet en una escala de 1 a 10, donde de 1 a 4 pertenece a un sentimiento negativo, de 5 a 6 un sentimiento neutro y de 7 a 10 un sentimiento positivo, esto para una base de alrededor de 7.000 tweets crudos (300 tweets por persona aproximadamente).

ENTENDIMIENTO DE DATOS:

En el proyecto se busca utilizar las funciones disponibles para obtener información líquida susceptible de analizar para la base de datos que fue extraída de Twitter de la marca "***Avianca***", esta marca recibe diariamente la opinión de aproximadamente 1.032 usuarios reflejada en un promedio de 4.000 tweets y aproximadamente el 41% de ellos son negativos.

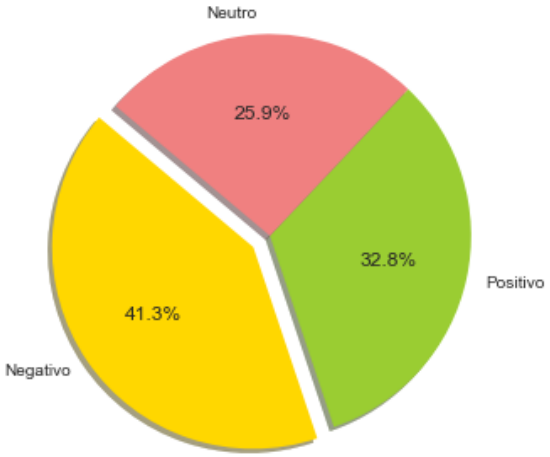


Gráfico 4: Proporción de sentimientos - Avianca

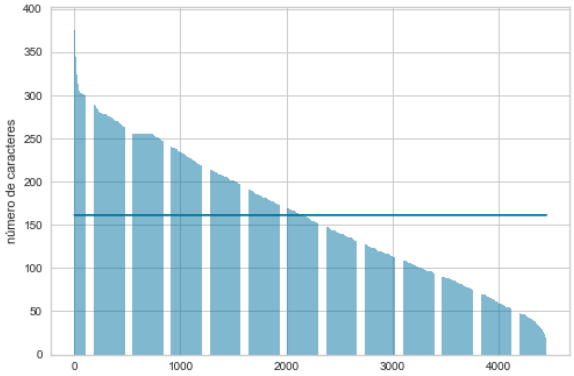


Gráfico 5: Número de caracteres

PREPARACIÓN DE DATOS:

El análisis de texto es un campo creciente. Las grandes cantidades de texto representan una oportunidad para las organizaciones que desean encontrar patrones, conocer las opiniones de sus consumidores, analizar posibles concurrencias, e inclusive, generar predicciones sobre las palabras que el usuario digitará, o traducir en tiempo real un texto entre distintos idiomas.

Se presenta el procedimiento que se aplicó en la limpieza de datos:

1. Librerías utilizadas:

```
import codecs
import spacy
from tokenize import tokenize, untokenize, NUMBER, STRING, NAME, OP
import pandas as pd
import numpy as np
import random
import re
import string
import nltk
from nltk.collocations import *
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
import matplotlib.pyplot as plt
import heapq
from nltk.stem import WordNetLemmatizer
from nltk.stem.snowball import SnowballStemmer

#nltk.download('stopwords')
nlp=spacy.load('es_core_news_sm')
```

2. Carga de los datos:

```
df = pd.read_excel('TweetsEntrenamiento.xlsx')

# Extraemos el texto
corpus= list(df['Texto'])
MiListaStops=list(set(stopwords.words('spanish')))
```

3. Limpieza del texto del tweet:

```
for i,tweet in enumerate(corpus):
    tweet=re.sub('@([\w.]+ )','',re.sub('https.*','',tweet))
    tweet=re.sub('#[\w]*','',tweet)
    tweet = tweet.lower()
    tweet = re.sub(r'\W',' ',re.sub(r'\s+',' ',tweet))
    tweet = re.sub('¡','',re.sub('º','',tweet))
    tweet=tweet.translate(str.maketrans('', '', string.punctuation))
    for word in tweet.split():
        if str(word.lower()) in MiListaStops:
            tweet=re.sub(r'\b'+str(word)+r'\b','',tweet)
    tweet=re.sub(' +',' ',tweet)
    corpus[i] = tweet
```

Dado el “ruido” en la información que se obtiene al cargar los tweets, se eliminan aquellos caracteres o conjunto de ellos que podrían afectar al correcto entrenamiento del modelo pues podrían sesgar el modelo hacía un sentimiento especifico o simplemente aportar mayor dimensionalidad a la bolsa de palabras. Dichos elementos son:

1) **Usuarios** etiquetados en cada tweet: Se identifican por medio de “@” precediendo cada palabra.
2) **Hashtags** o **Etiquetas** en cada tweet: Se identifican por medio de “#” precediendo cada palabra
3) **URLs** en cada tweet: Se identifican por medio de “http” precediendo cada palabra
4) Los **signos de puntuación**
5) **Stopwords**: Además de los conectores usuales, en twitter existen otro conjunto de palabras que no hacen referencia a ningún sentimiento y simplemente tratan acerca de formalismos.

4. Tokenización y lematización:

```
wordfreq = {}
for sentence in corpus:
    tokens = nltk.word_tokenize(sentence)
    tokens=nlp(sentence)
    for token in tokens:
        if not str(token) in list([' ']):
            token=token.lemma_
            if token not in wordfreq.keys():
                wordfreq[token] = 1
            else:
                wordfreq[token] += 1
```

5. IDF

```
for document in corpus:
    t =[]
    for token in nlp(document):
        if not str(token) in list([' ']):
            t.append(token.lemma_)
    milista.append(t)

word_idf_values = {}
for token in most_freq:
    doc_containing_word = 0
    for document in milista:
        if token in document:
            doc_containing_word += 1
    word_idf_values[token] = np.log(len(corpus)/(1 + doc_containing_word))
```

6. TF

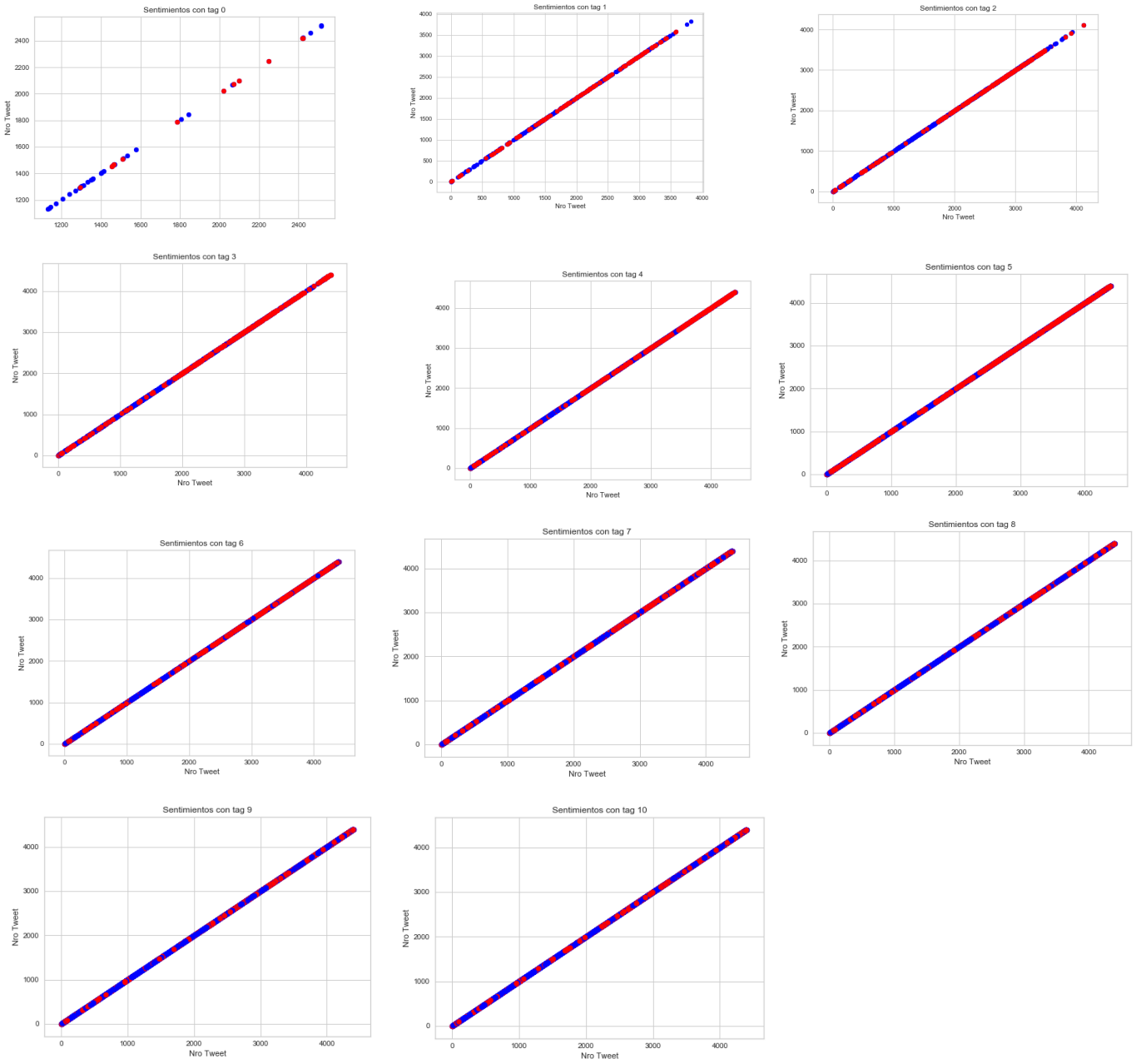
```
tfidf_values = []

for token in word_tf_values.keys():
    tfidf_sentences = []
    for tf_sentence in word_tf_values[token]:
        tf_idf_score = tf_sentence * word_idf_values[token]
        tfidf_sentences.append(tf_idf_score)
    tfidf_values.append(tfidf_sentences)


tf_idf_model = np.asarray(tfidf_values)
tf_idf_model = np.transpose(tf_idf_model)
```

MODELAMIENTO

Los **outliers** en el problema de análisis de sentimiento empleando técnicas de text mining corresponden a aquellas palabras clasificadas con un sentimiento específico pero que dado el contexto o posición en la que aparecen en la oración no corresponden con el sentimiento previamente asignado. Para identificarlos se utiliza la distancia de cooks (Sharyn O'Halloran,2007) pues permite medir el nivel de influencia de un punto considerando su posición en el conjunto de datos (Bag of words idf para el caso). Ad hoc, se eliminan aquellos tweets cuyos sentimientos se encuentren a más de 1 del texto mediano bajo la distancia de cooks (aproximadamente 17% del total de los datos), esta depuración hace sentido debido a la diversidad de interpretaciones que se puede tener para palabras con sentimiento negativo/positivo al ser analizadas por un alto volumen de personas.



Para reducir la dimensionalidad de la matriz de covarianza, se empleó “randomizedpca” de sklearn debido a lo dispersa que era esta.

Entrenamiento del modelo

Dado un problema de clasificación, se utilizan métodos de discriminación buscando lograr una predicción mayor a la obtenida en las herramientas licenciadas actuales (AWS y Azure, 54% y 52% respectivamente).

Buscando disminuir el error de testeo y garantizar que el modelo logre generalizar la mayoría de los datos, se utiliza la validación cruzada para cambiar el conjunto de datos de testeo en paquetes de 20% del total de la información,

Además. se utiliza la optimización de parámetros bajo gridsearch de sklearn, obteniendo los siguientes resultados:

K vecinos cercanos: Se consideran 4 escenarios iterando el número de vecinos hasta 4.

Máquinas de soporte vectorial: Kernel lineal

Arbol de decisión: Profundidad máxima: 124

Random Forest: Profundidad máxima: 233

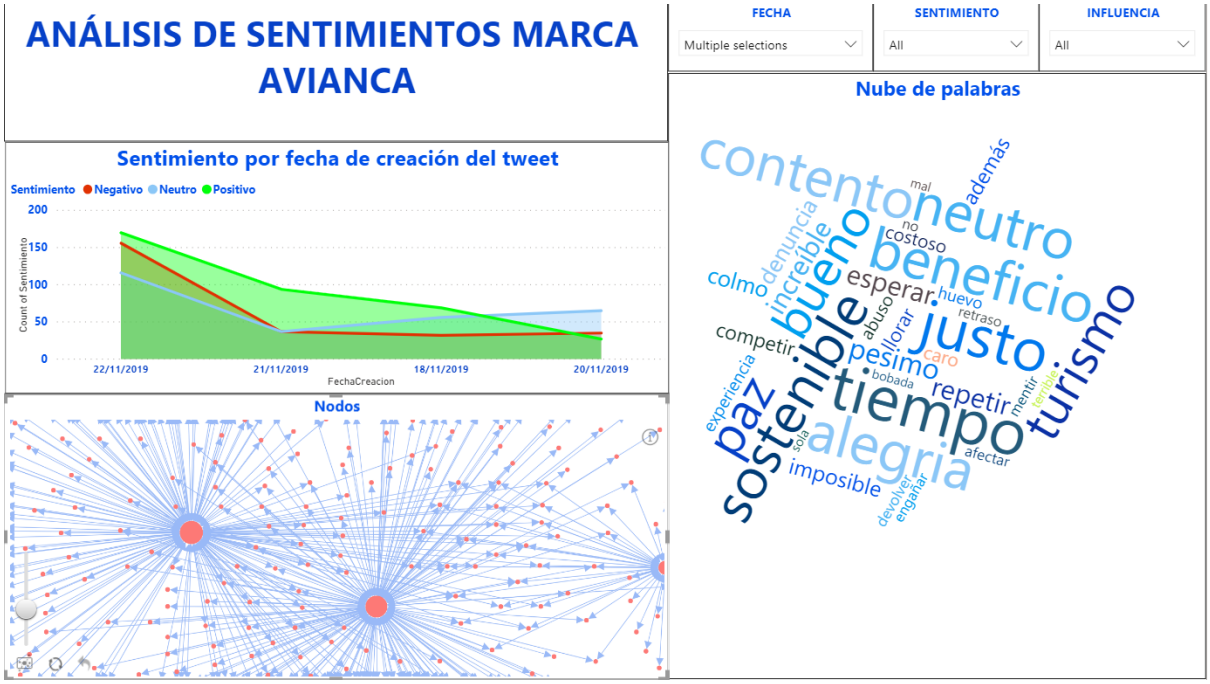
Gradient Boosting: Profundidad máxima: 5

Evaluación

Métricas de evaluación:

1. Precisión (Positivos verdaderos + Negativos verdaderos / Total de registros)

Despliegue



Gráfica 6. Power Bi.

La representación gráfica de los nodos se presentará haciendo uso de PowerBi, siendo cada nodo principal el sentimiento de la marca, y los nodos secundarios el código único (id) que representa un usuario en la red social y sus aristas la relación influyente entre su comunidad y el sentimiento de sus publicaciones.

Las funciones empleadas dentro del software son:

- **ZoomChart Advanced Graph Visual:** Creación de nodos y relaciones en Bi.
- **Area Chart:** Cobertura de sentimientos.
- **WordCloud:** Nube de palabras con tamaño relativo al peso.

FeedBack

Desde el punto de vista de negocio, este feedback brindará información cuantitativa para comprender la relación que tienen sus campañas o contenido digital con sus usuarios y comunidades, con el fin de poder analizar la efectividad de sus campañas

y poder tomar decisiones rápidas frente al relacionamiento con sus usuarios y desarrollo de prototipos y nuevos productos

Repositorios:

Github	https://github.com/eljimenezj2/Proyecto-Integrador
Power BI publicado	https://powerbi.microsoft.com/es-es/landing/signin/?ru=https%3A%2F%2Fapp.powerbi.com%2F%3Froute%3Dgroups%252f568c7310-143f-447e-8880-1ae684704a8f%252freports%252fe7578a8e-4b17-4a72-bc19-acd7ec6e83ce%253fctid%253d99f7b55e-9cbe-467b-8143-919782918afb%26ctid%3D99f7b55e-9cbe-467b-8143-919782918afb%26noSignUpCheck%3D1
Bucket (para PowerBi)	https://proyecto-integrador-2019-2-analisis-sentimientos-marca-avianca.s3.amazonaws.com/Data/DatasetFinal.xlsx

REFERENCIAS

1. Macho stadler, m.

(2009). Topologia de espacios metricos. Recuperado de

Http://www.piffard.ch/

2. Why is social media data

Mining important in today's market scenario. (s. F.). Recuperado 5 de

Diciembre de 2019, de

Https://www.outsourcedataworks.com/why-is-social-media-data-mining-important-in-today-market-scenario.html

3. Eljimenezj2/proyecto-integrador:

Este repositorio contiene el desarrollo y los resultados del proyecto

Integrador para el primer semestre de la maestria en ciencia de los datos y

Analítica de la universidad eafit. (s. F.). Recuperado 5 de diciembre de

2019, de https://github.com/eljimenezj2/proyecto-integrador

3. Text data management and analysis, chengxiang zhai & sean massung, university of

Illinois at urbana–champaign, acm books, 2016.

4. Text mining: concepts, implementation, and big data challenge, taeho jo, springer, 2019.

5.data, information, knowledge, wisdom (dikw): a semiotic theoretical and empirical exploration of the hierarchy and its quality dimension , saša baškarada , andy koronios , australasian journal of information systems , volume 18 number1 2013.

6. Peña, d. (2002) analisis de datos multivariantes. Mcgraw-hill

7. Bailo, A., Grané A. (2008) Problemas resueltos de estadistica multivariada implementados en matlab. Delta Publicaciones.