Análisis de Sentimiento de Tweets de Aerolíneas Españolas

Análisis de cómo los viajeros expresan sus sentimientos

DATA SCIENCE @ UNIVERSITAT DE BARCELONA

Universitat de Barcelona

Equipo TAOPYPY

Miquel Vives

Project Manager

T-Systems Iberia

Ferran López

Director Fundador

TekneCultura

Félix Hernández

Data Engineer

Cornerjob

Berta Izquierdo

Team Lead - Google Cloud

SELLBYTEL

Tao: *Puede traducirse literalmente por 'el camino', 'la vía', o también por 'el método' o 'la doctrina'*
Tao Pai Pai: *Sicario contratado por el Comandante Red para eliminar a Son Gokū.*

Index

Objetivo

Objetivo

Análisis de sentimientos de tweets sobre líneas aéreas y creación de **modelo** de extracción de sentimiento.

Universo:

Numero de tweets: 7867 tweets.

Tweets creados entre noviembre 2017 y enero 2018.

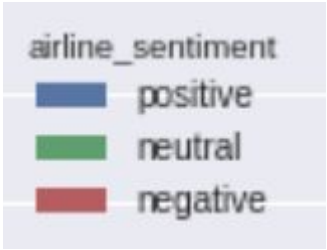
Datos iniciales de cada tweet: sentiment, reply, número de replies, número de retweet, texto del tweet, localización, fecha de creación, tweet_id, timezone del usuario.

Clasificación de sentimientos: positivo , negativo , neutro.

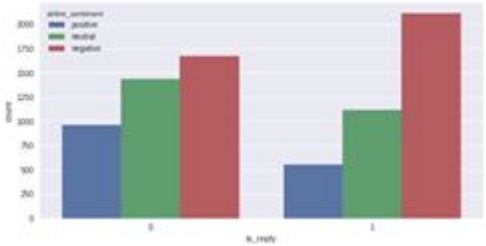
Análisis Inicial

Análisis Inicial

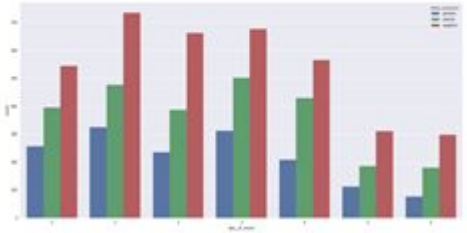
Análisis de los datos originales del set de datos.



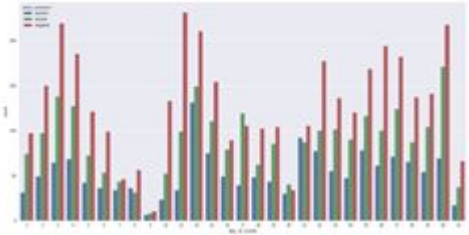
Is_reply



Day of the week

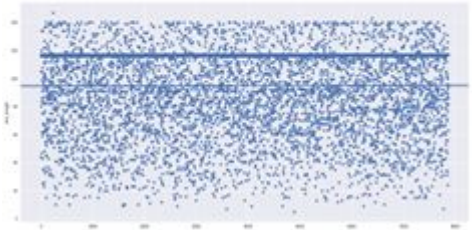


Day of the month



Day of the year

Tweet's length



Preparación de datos y modelos usados

Modelos usados y preparación de datos

Preparación de datos:

- Eliminación de datos incorrectos o tweets sin texto.
- HTML decoding
- Extracción de menciones a líneas aéreas
- Eliminación de URL links

	airline_sentiment	is_reply	text	tweet_created	user_timezone	emojis
1	positive	FALSE	"Los pilotos de Ryanair desconvocan la huelga ...	Mon Dec 18 13:07:04 +0000 2017	Dublin	
2	positive	TRUE	@Iberia @lavecinarubia Si ,por favor las decla...	Sat Nov 04 17:05:11 +0000 2017		
3	neutral	TRUE	@Iberia Me dirías por favor que costo tiene?	Sat Dec 02 15:24:09 +0000 2017		
4	negative	TRUE	@SupermanlopezN @Iberia @giroditalia Champion,...	Thu Dec 21 23:17:43 +0000 2017	Central Time (US & Canada)	
5	negative	TRUE	@SrtaFarrellDM @KLM @Iberia Eso de avianca es ...	Wed Dec 06 00:44:25 +0000 2017	Eastern Time (US & Canada)	😠

Modelos clásicos

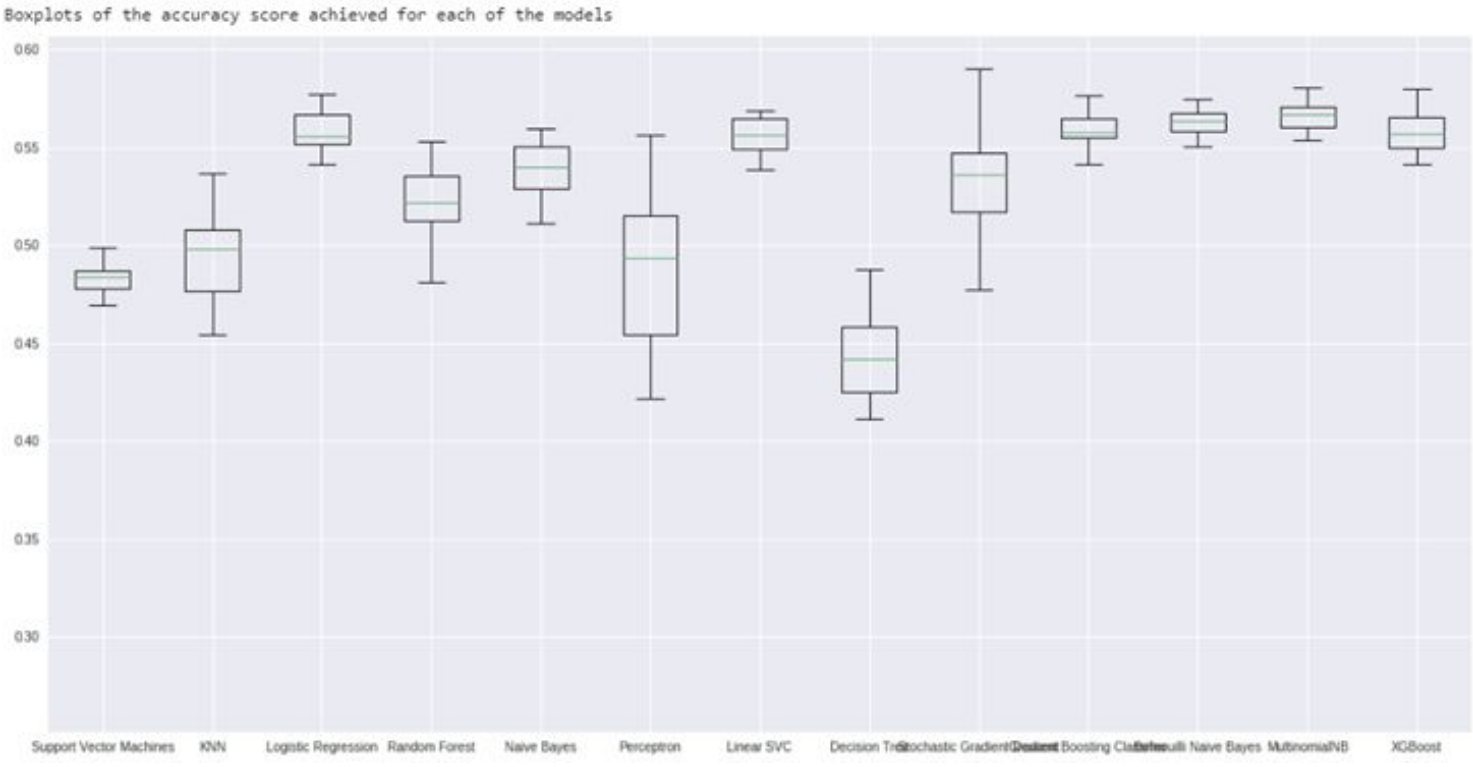
Estudio de 13 modelos clásicos sin ajuste de parametrización, para escoger aquel que inicialmente tiene mejor resultados.

Usamos:

- TfidfVectorizer
- Token = r'([A-Za-z]{3,}|no)
- stopwords de español de nltk.
- Datos de Train = 75%; Datos de Test = 25%
- BoW de 500 palabras

Modelos clásicos

Comparativa de modelos



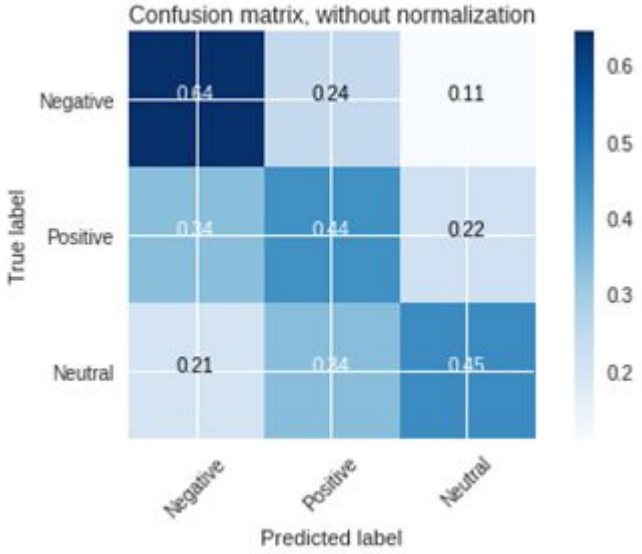
Modelos clasicos: XGBoost

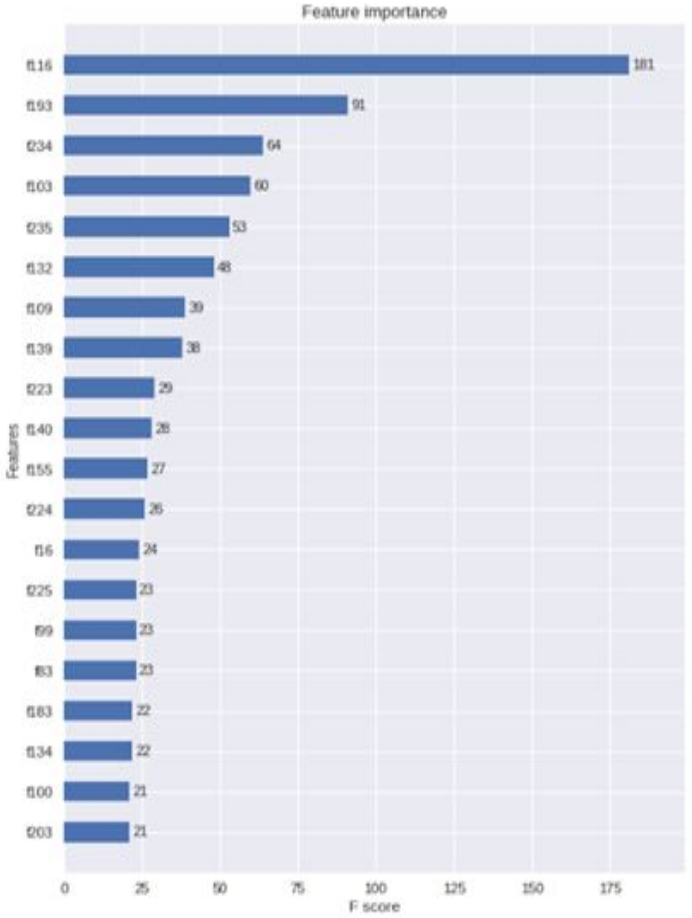
Elección de XGBoost porque inicialmente parece el que tiene mayor accuracy, y porque la bibliografía nos indica que es el de mayor rendimiento en tiempo.
Optimizamos los parámetros de XGBoost:

- 18 parámetros optimizados.
- Dataset con/sin emojis y con/sin bigramas.

Mejor resultado con XGBoost: con emojis y bigramas: **Accuracy 61%**

Modelos clasicos: XGBoost





facturar, necesito, hora/horas, precio/precios/pagar, equipaje, destinos, 'mejor precio‘, 'equipaje mano'

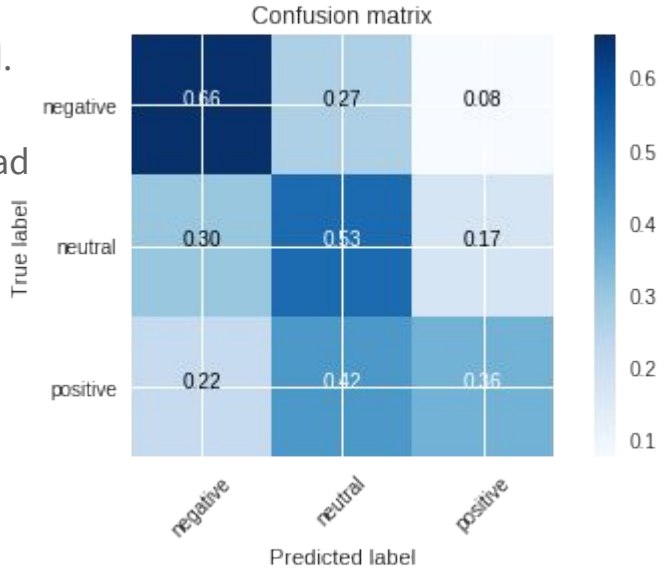
Redes neuronales: RNN-LSTM

Recurrent Neural Network + Long-Short Term Memory:

La bibliografía nos indica que son usadas para este tipo de modelado.

Características usadas:

- Una capa embedding: convierte las palabras en vectores para la red.
- Una capa recurrente de tipo LSTM.
- Una capa densa: convierte la salida de la red a vector de probabilidad de longitud 3.
- Igual pre proceso de datos: Stopwords de nltk y limpieza de links.
- Imput de la RNN: texto y sentimiento.

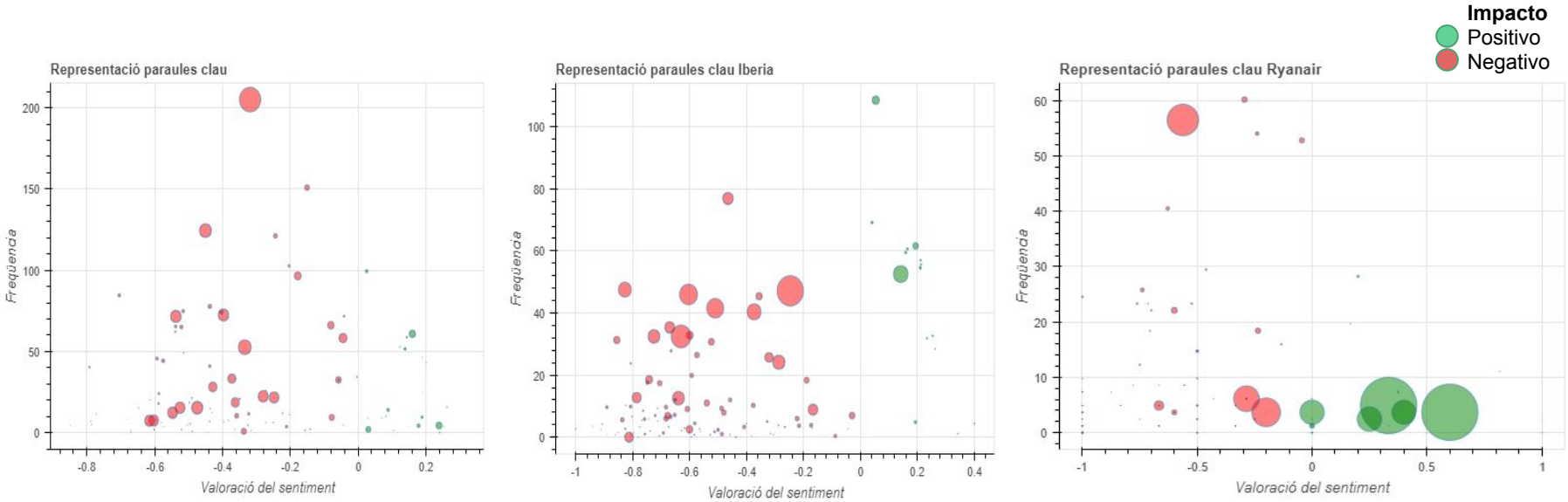


Accuracy: 81%

Resultados

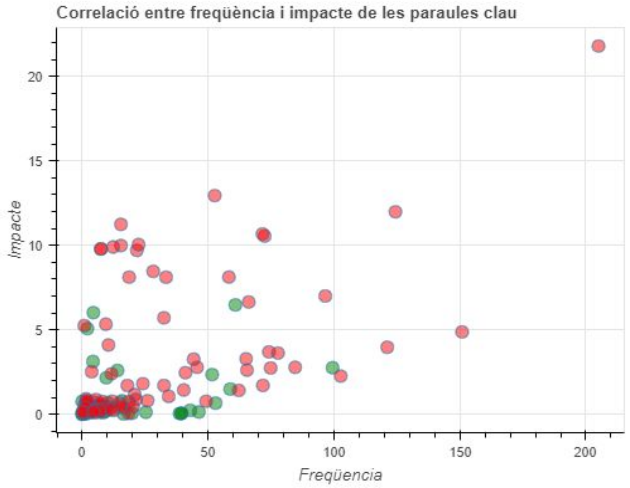
Resultados

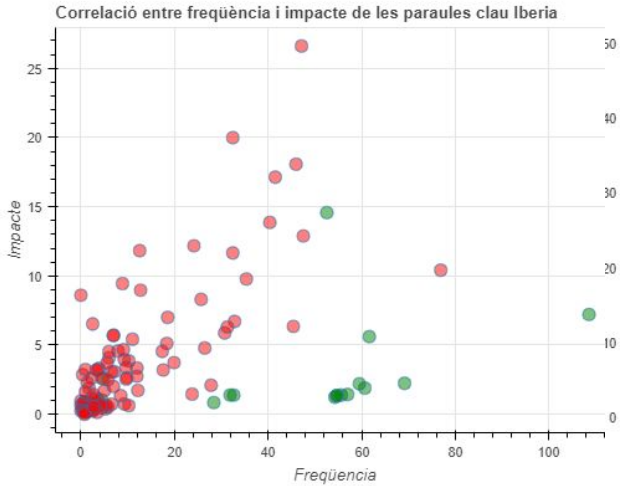
Analizamos el contenido de los tuits clasificados, su sentimiento en función de su frecuencia e impacto (evaluado a partir del volumen de seguidores que potencialmente han visto los tuits)

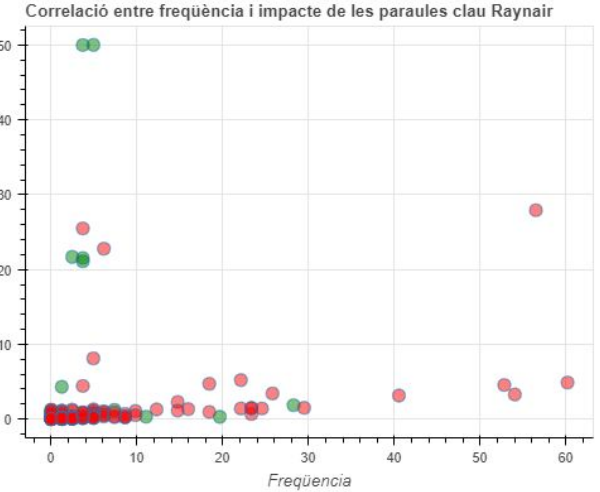


Resultados

Analizamos la correlación entre frecuencia e impacto, observando (en el caso de tuits que mencionan @iberia) que los tuits negativos tienen un impacto potencial mayor. La comunidad más conectada en twitter utiliza la plataforma como vehículo de queja.







Resultados

La obertura de nuevas rutas, las ofertas y concursos o acciones de marketing para “comprando” contenido positivo y la atención al cliente con resultado positivo son las principales, casi únicas, razones por las que las compañías aéreas reciben alabanzas en la red.

El mal servicio, problemas con maletas y la reiteración de estos problemas producen los tuits más negativos.

Las redes sociales van a ser más foco de desprestigio y de mala relación con el cliente si los tuits no son convenientemente clasificados y tratados para mitigar los potenciales efectos negativos

Twitter, un termómetro para pulsar la reputación de IBERIA

El Periódico Taopypy

ECONOMÍA

Iberia aplica tecnologías digitales inteligentes para mejorar la comunicación con sus clientes.

El Periódico Taopypy
Barcelona - Martes, 03/07/2018

Iberia analiza el perfil de los usuarios de Twitter con la finalidad de mejorar su experiencia de usuario. Nuevos destinos y promociones encabezan la lista de satisfación. Al contrario, incidencias con los **equipajes** y los **retrasos** encabezan la lista de quejas y reclamaciones. Este análisis ha permitido a la compañía aérea implementar un servicio para seguir la situación del equipaje y de los vuelos en tiempo real, el cual parece haber tenido buena acogida entre los usuarios.

Mejoras futuras

Mejoras futuras

XGBoost vs RNN+LSTM.

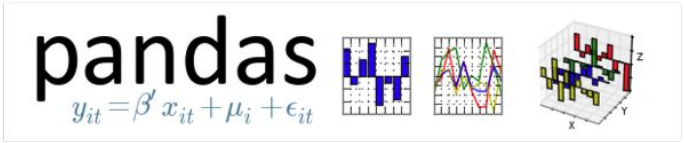
- Igual preparación de datos.
- Accuracy muy diferente entre ambos modelos: 61% vs 81%.

Posibles mejoras:

- Emojis quizá no correctamente usados en el modelo clásico.
- Uso de sinónimos para palabras y verbos.
- Añadir datos al modelo: localización, reply y hora.
- Estudio de palabras para stopwords.

Herramientas utilizadas

Herramientas utilizadas



¡Gracias!

?