TRABAJO FIN DE MASTER
ANALISIS DE SENTIMIENTOS DE TWITTER: @vertiseguros

Javier González Méndez

Tutor: Felipe Ortega

Máster Data Science Curso 2016/2017

1.- Objetivos

El principal objetivo es la aplicación de los conocimientos adquiridos en el máster.

En este caso, se trata de realizar un análisis de sentimiento de una cuenta de Twitter: @vertiseguros.

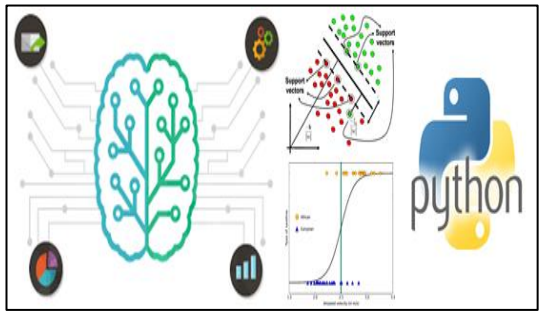
Verti es una compañía de venta directa de seguros. Productos: auto, moto, hogar y mascotas.

2.- Planteamiento inicial: fases y tecnologías

Obtención de datos de Twitter

Tecnologías: Python, Twitter

Objetivos: recuperar los datos iniciales para el trabajo



Modelización estadística

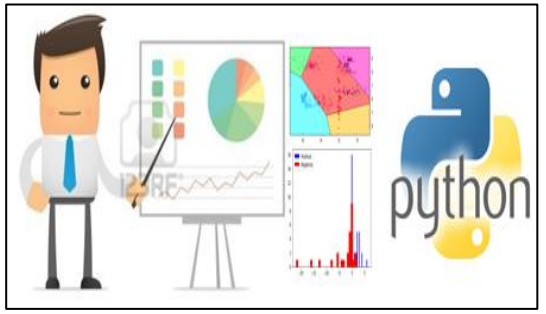
Tecnologías: Python (NLTK, Sklearn)

Objetivos: entrenamiento y validación de modelos estadísticos. Decisión de tomar el mejor o el que más nos interesa

Limpieza y depuración de datos

Tecnologías: Python, Afinn

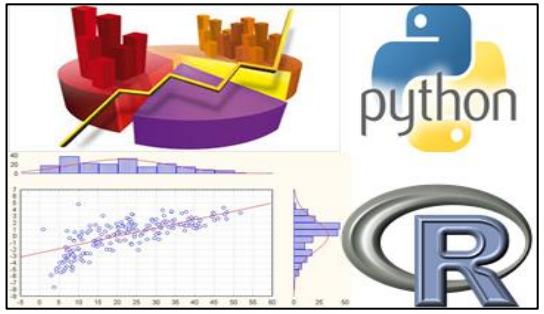
Objetivos: una vez obtenidos los datos, los tratamos y los depuramos para dejarlos óptimos



Presentación de resultados

Tecnologías: Python

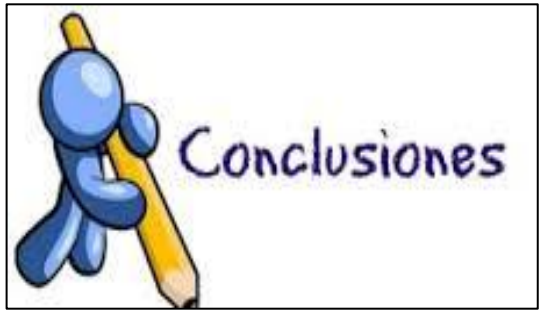
Objetivos: aplicación del modelo ganador a datos actualizados (agosto 2017)



Análisis exploratorio de datos

Tecnologías: Python, R

Objetivos: realizamos un análisis descriptivo y exploratorio de los datos para seguir depurando y creamos los ficheros de entrenamiento



Conclusiones y próximos pasos

Objetivos: resumen de las principales conclusiones e impresiones del proyecto y próximos pasos

3.- Obtención de datos de Twitter

@vertiseguros nace en 2011

PROBLEMA

La api de Twitter sólo recupera datos de una semana de antigüedad

GetOldTweets-Python (Jefferson-Henrique) :

- Scrapper en Python
- Permite recuperar los tweets según un patrón de búsqueda, fechas, usuarios, …
- Exporta los datos a csv

SOLUCIÓN

Principal problema: saca pocos datos (10 campos) y no hay filtro de lenguaje, ni geográfico,…

Primer intento: descargar todos tweets que contengan Verti → Problema idiomático, demasiados datos no útiles

Después de pruebas, se hacen búsquedas más precisas: usuario @vertiseguros, consultas por palabras claves, como son “verti seguro”, “verti spotify”, “verti bien”, “verti mal”, “verti peor”, “verti siniestro”,… → 22.000 tweets

4.- Limpieza y depuración de datos

Tratamiento al texto de los datos:

- Eliminación de urls
- Eliminación de caracteres extraños

Diferenciamos 4 grupos de tweets (sólo usaremos los dos primeros):

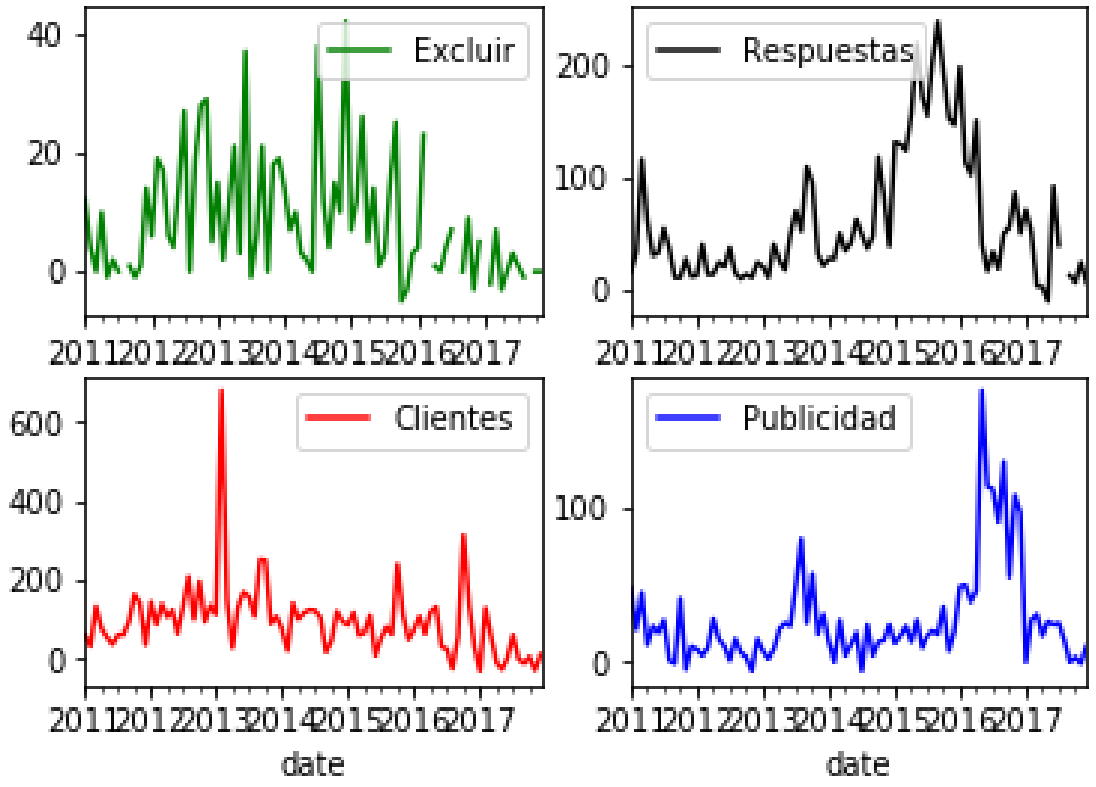
- Respuestas: son tweets emitidos desde la cuenta de Vertiseguros para responder a los clientes
- Clientes: son tweets de clientes o potenciales clientes en los que se refieren a la compañía (pueden ser emitidos a la cuenta @vertiseguros o un tweet genérico que contenga la palabra Verti)
- Publicidad: son tweet emitidos desde la cuenta de Vertiseguros con propósitos exclusivamente comerciales
- Excluidos: estos son tweets a excluir debido a quien los ha escrito (CEO, Otras compañías,…)

Asignamos polaridad a los Tweets:

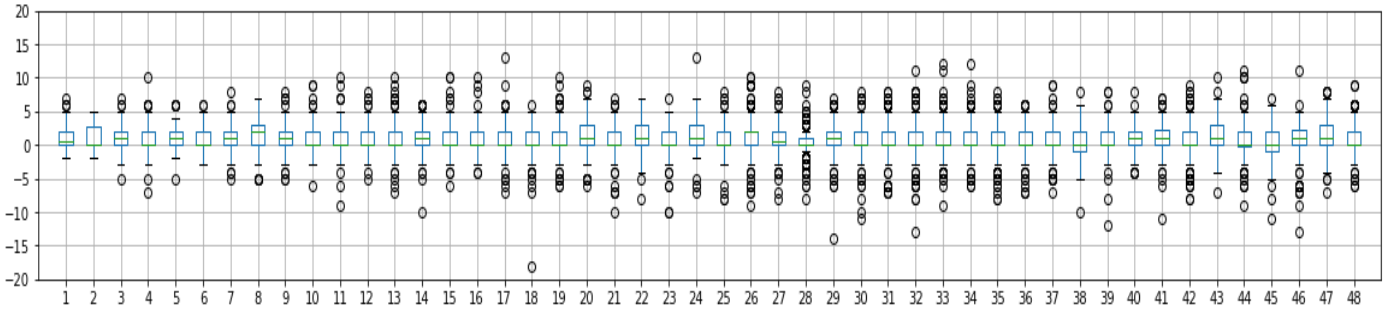
- Diccionario de términos AFFIN-165
- Traducimos a castellano y 'lematizamos' → dada una forma flexionada (plural, femenino, conjugada, etc.), hallar el lema correspondiente
- Según esas palabras asignamos puntuación y calculamos la polaridad → uso como fichero de entrenamiento

5.- Análisis exploratorio de datos (1)

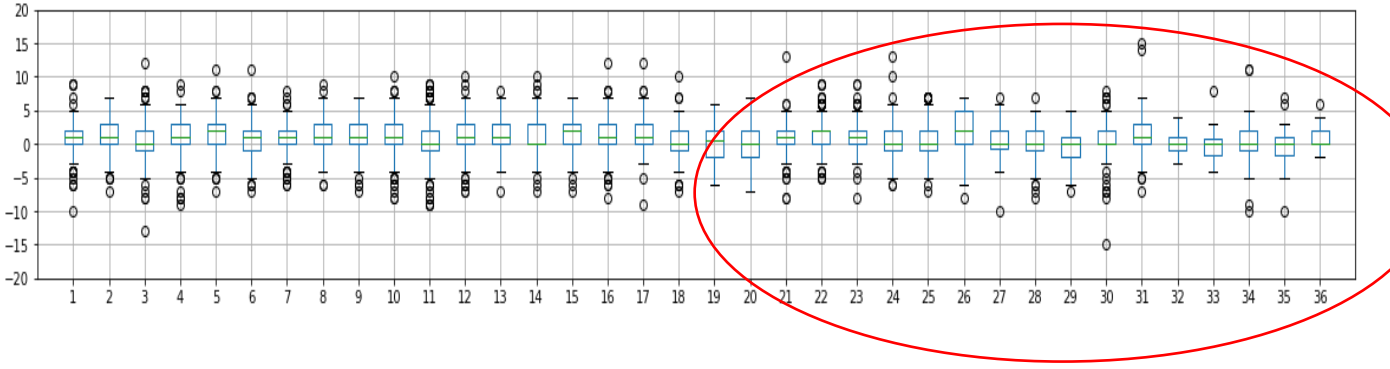
Distribución de tweets



Desde 2011 - 2014



Desde 2015 - 2017



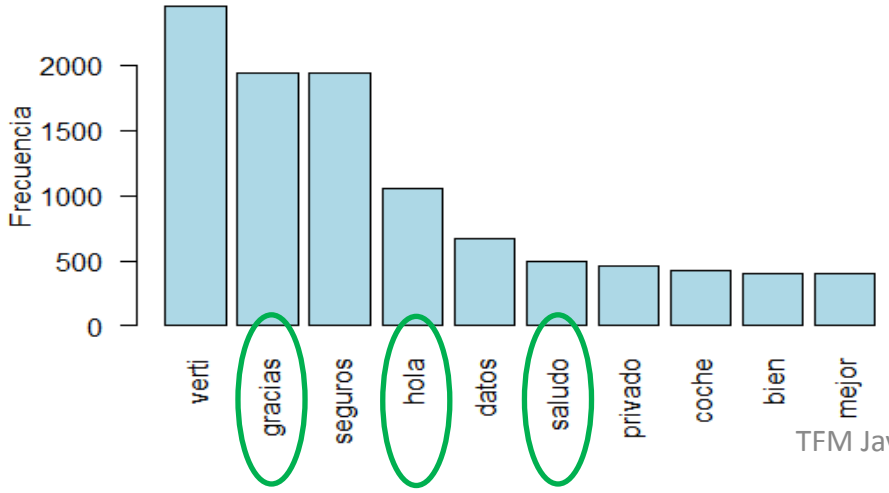
Aumento de tweets negativos

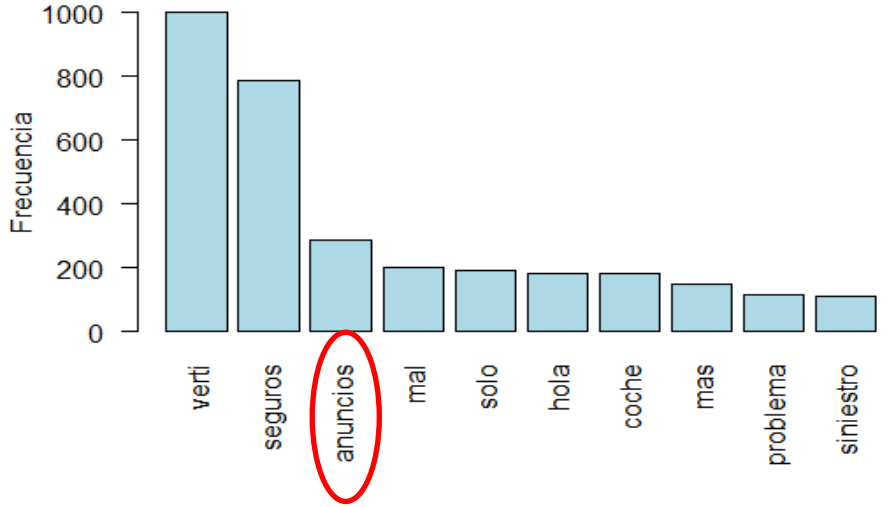
5.- Análisis exploratorio de datos (y 2)

POSITIVOS

NEGATIVOS

Palabras más frecuentes





6.- Modelización estadística (1)

nltk.classify:

NaiveBayesClassifier

MaxentClassifier

sklearn.naive_bayes:

MultinomialNB

BernoulliNB

7 modelos

sklearn.svm:

LinearSVC

sklearn.linear_model:

LogisticRegression

SGDClassifier

estadísticos:

Accuracy - Preccision

Recall - F-measure

entrenamiento:

sin validación cruzada

con validación cruzada

32 combinaciones

texto:

con stop-words

sin stop-words

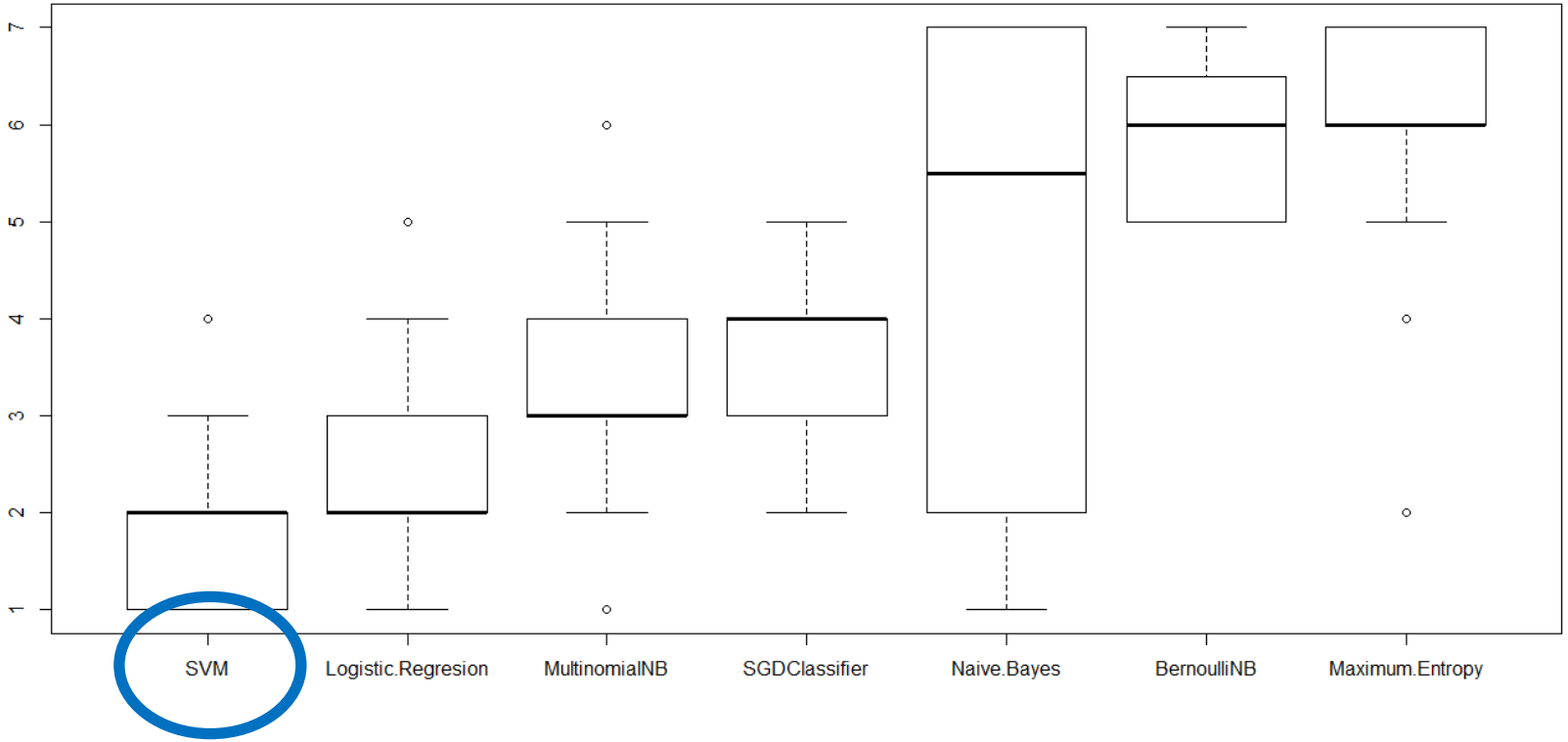
N-gramas:

uni-gramas

bi-gramas

6.- Modelización estadística (2)

Elección del mejor modelo → *'rankineamos'* los valores obtenidos por cada combinación y modelo de 1 a 7



6.- Modelización estadística (y 3)

Viendo las palabras más relevantes según el modelo estadístico de Naive Bayes, obtenemos otro punto a tener en cuenta:

trate = True	neg : pos	=	71.8 : 1.0
puta = True	neg : pos	=	64.7 : 1.0
juro = True	neg : pos	=	57.7 : 1.0
baje = True	neg : pos	=	48.3 : 1.0
odio = True	neg : pos	=	45.9 : 1.0
guardia = True	neg : pos	=	43.5 : 1.0
enfermedad = True	neg : pos	=	31.8 : 1.0
perros = True	neg : pos	=	29.4 : 1.0
saludo = True	pos : neg	=	27.4 : 1.0
duele = True	neg : pos	=	27.1 : 1.0

A vigilar

Palabra pol1 : pol2 = XX.X : YY.Y

Este cuadro indica que si aparece la palabra **Palabra** hay **XX.X** veces sobre **YY.Y** posibilidades de que el tweet tenga una polaridad **pol1**

Ejemplo:

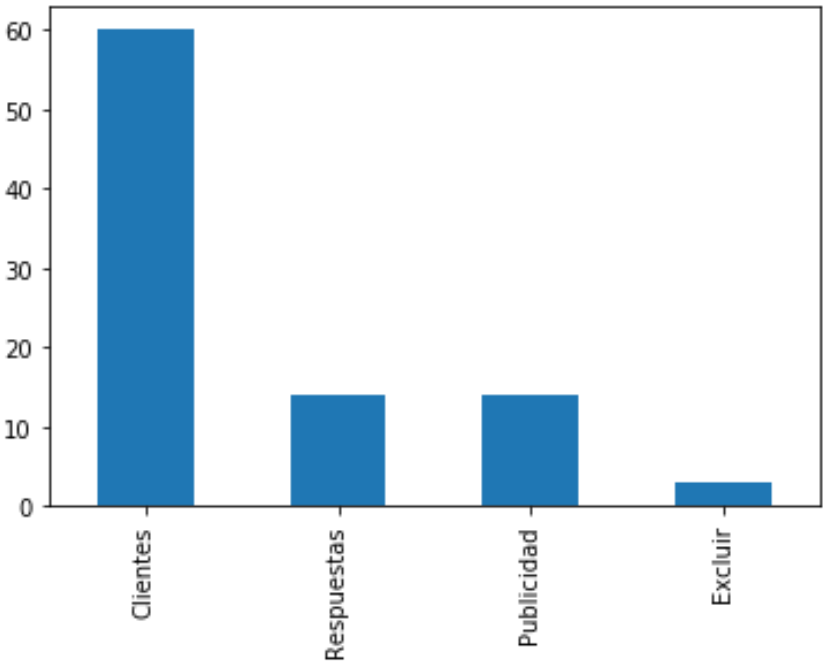
odio = True neg : pos = 45.9 : 1.0

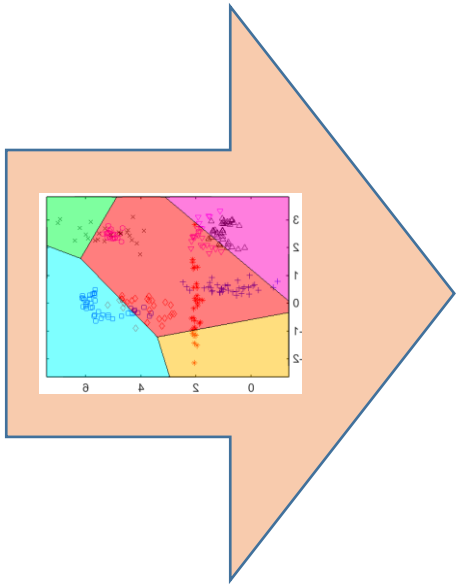
Si aparece odio en el texto, hay 45,9 veces más posibilidades de que el tweets sea catalogado como negativo

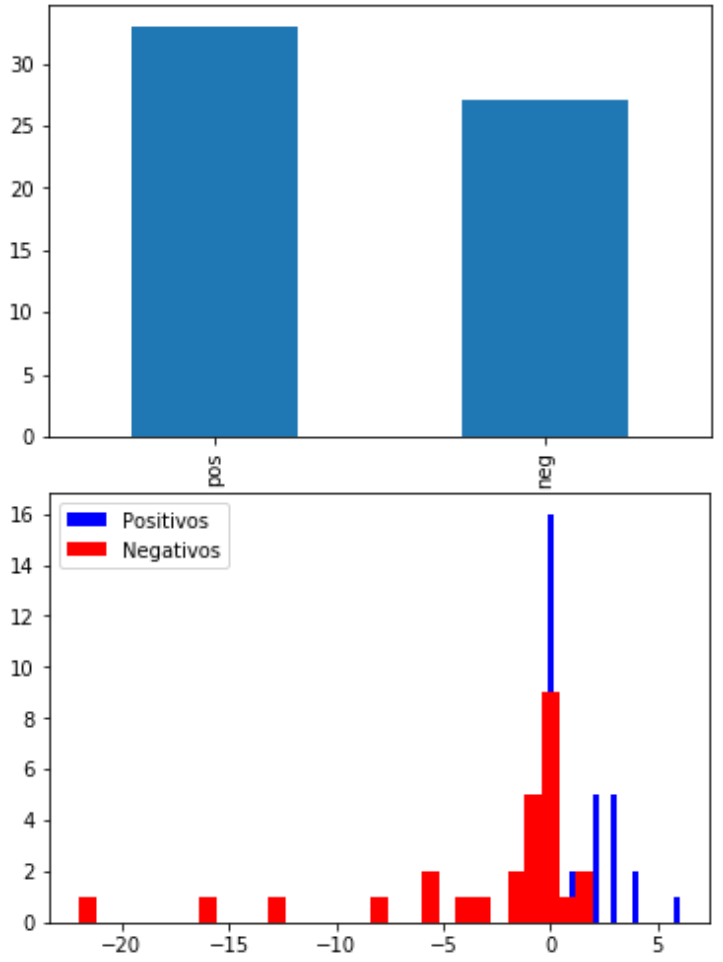
7.- Presentación de resultados (1)

Una vez que tenemos modelo ganador, lo evaluamos con datos reales de agosto 2017

Distribución de tweets de ago17

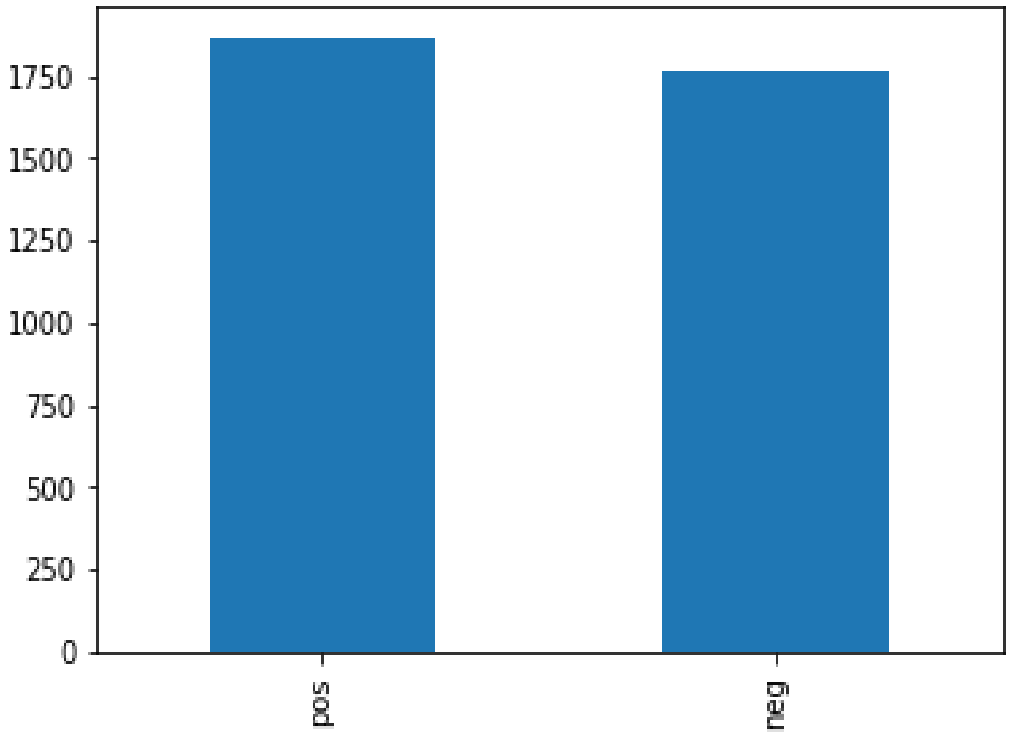


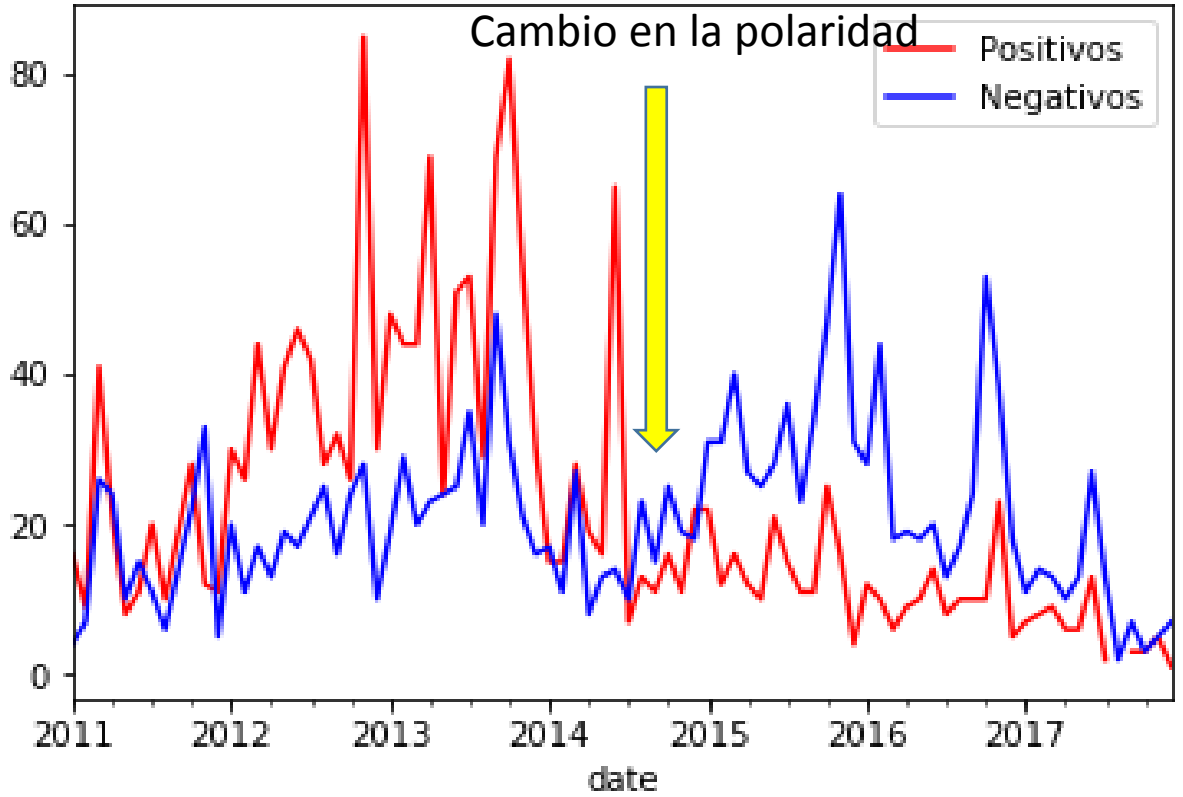




7.- Presentación de resultados (2)

¿Qué pasa con los tweets neutros no usados en el entreno? ¿Cómo son?





8.- Conclusiones

- Problemática en la obtención de datos.
- La evolución anual de la polaridad inicial de los tweets en los últimos tiempos está tendiendo hacia la negatividad.
- En la nube de palabras negativas aparece una que hay que vigilar: anuncio. ¿Los anuncios de Verti son molestos? ¿Es algo puntual en el tiempo o es de manera general?.
- Hemos probado hasta 7 modelos distintos con 32 combinaciones. Modelo ganador → SVM (kernel lineal).
- En las palabras más relevantes según modelo de N-B aparece la palabra perro, ¿Hay algún problema con este producto?.
- Los tweets catalogados inicialmente como neutros, una vez categorizados según el modelo indican una inversión en la polaridad.

- Mejoras posibles:
  - Algoritmos calculados por defecto → tunearlos y optimizarlos
  - Mejorar la captura de datos (más términos clave)
  - ¿Uso de estos algoritmos para detección de fraude según partes de accidente?

Gracias