UNED Trabajo fin de Máster

Universidad Nacional de Educación a Distancia

Desarrollo de un sistema para la detección del machismo en redes sociales

Autor: Francisco Miguel Rodríguez Sánchez

Junio 2019

Directores: Jorge Amando Carrillo de Albornoz
Laura Plaza Morales

Índice

1. INTRODUCCIÓN Y OBJETIVOS

1. INTRODUCCIÓN

- Crecimiento de las redes sociales y “ciber” conflictos.
- Las mujeres tienen más del doble de probabilidades de sufrir acoso debido a su género (Duggan, 2017).
- Según un estudio (Amnistía internacional, 2017), Twitter es una red social tóxica para las mujeres.

1. OBJETIVOS

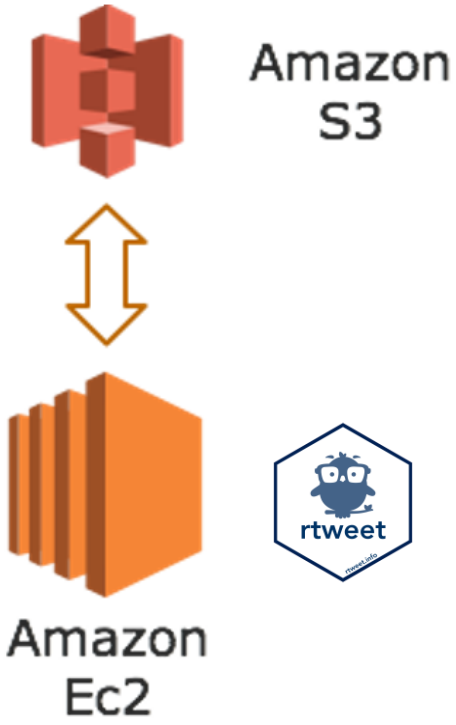
- Detectar señales textuales en castellano que conllevan lenguaje y actitudes machistas en redes sociales.
- Crear un corpus etiquetado con texto machista.
- Desarrollo de un sistema de clasificación supervisada.

2. HERRAMIENTAS UTILIZADAS

2. HERRAMIENTAS UTILIZADAS

Creación del corpus:

- *Amazon Web Services*
- *API REST Twitter*



Sistema de clasificación:

- NLTK (*Natural Language Toolkit*)
- Scikit-learn

3. MeTwo DATASET

3.1 MACHISMO EN TWITTER

- Se han estudiado diversas referencias para recopilar las expresiones más comunes.
- Expresiones o términos que minusvaloran el papel de las mujeres en la sociedad, incentivan el abuso o acoso hacia ellas o no les permiten expresarse libremente.

3.1 MACHISMO EN TWITTER

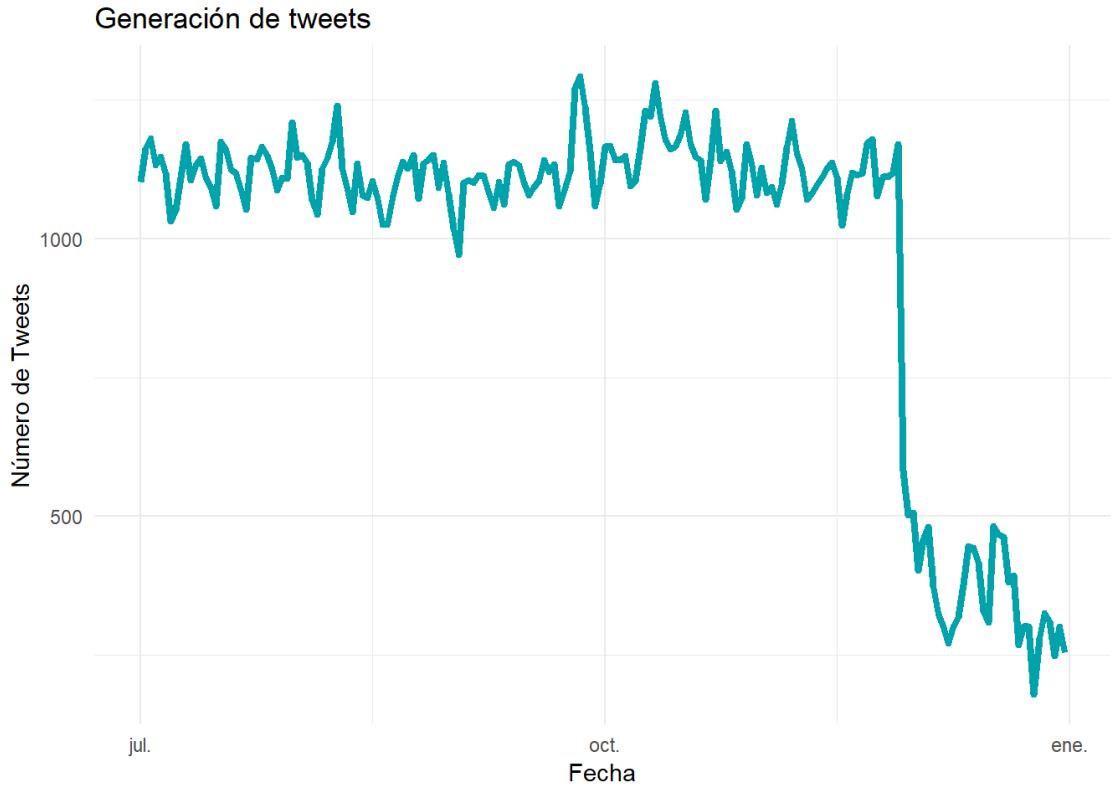
- Refranes y dichos populares: “Mujer al volante peligro constante” o “¡Mujer tenía que ser!”.
- Expresiones recurrentes en Twitter: “feminazi” o “nenaza”.
- Se han seleccionado un total de 29 términos.

3.1 MACHISMO EN TWITTER

- *"A mi me ha pasado igual. Los lobos vestidos de corderos que lo mismo quieren matar fachas que lloran como una **nenaza** no puedo con ellos"*
- *"Pero que violento!!! De cinco billetes, solo en uno aparece una mujer, esto es inaceptable!!! #**feminazi**"*

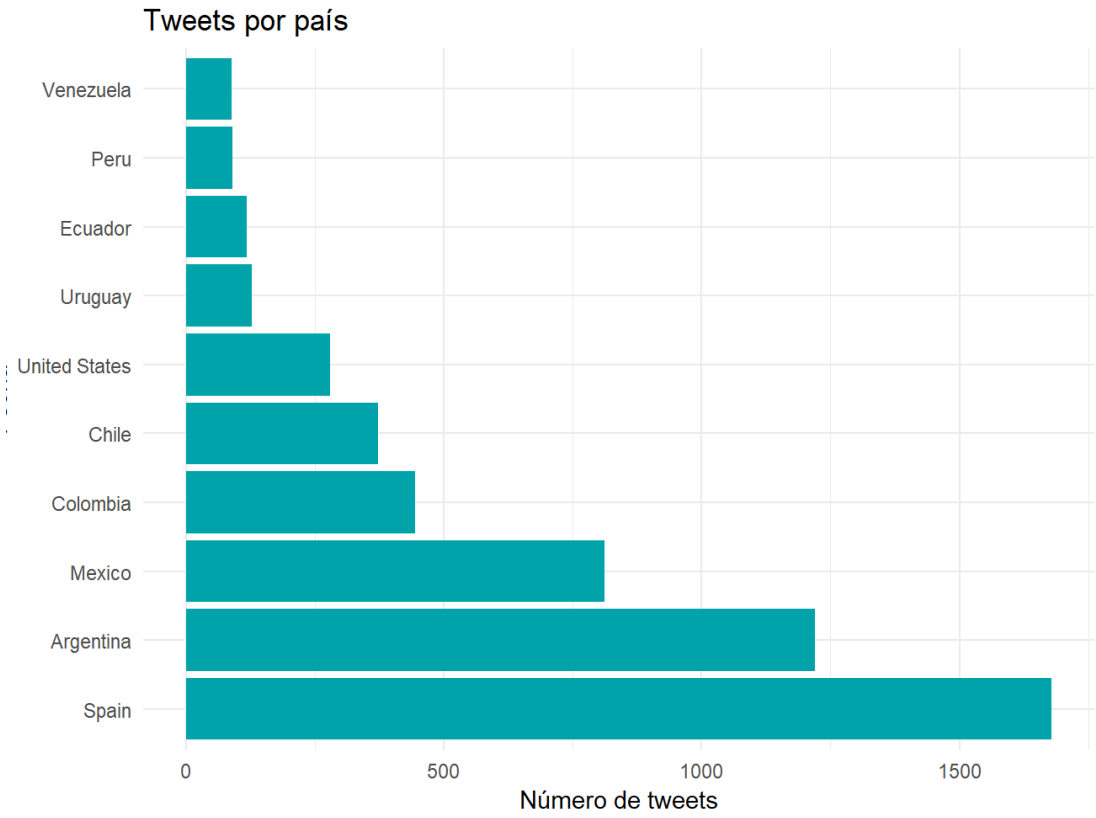
3.2 GENERACIÓN DEL CORPUS

- Proceso de *crawling* durante 6 meses (01/07/2018 - 31/12/2018).
- Se almacenaron un total de 181.792 tweets para todos los términos.
- Límite diario de 100 tweets por término hasta llegar a 15.000 tweets.



3.2 GENERACIÓN DEL CORPUS

- Para cada término se seleccionan 150 tweets aleatoriamente (24 términos).
- Se conforma un corpus final de 3.600 tweets a etiquetar.



3.3 ETIQUETADO DEL CORPUS

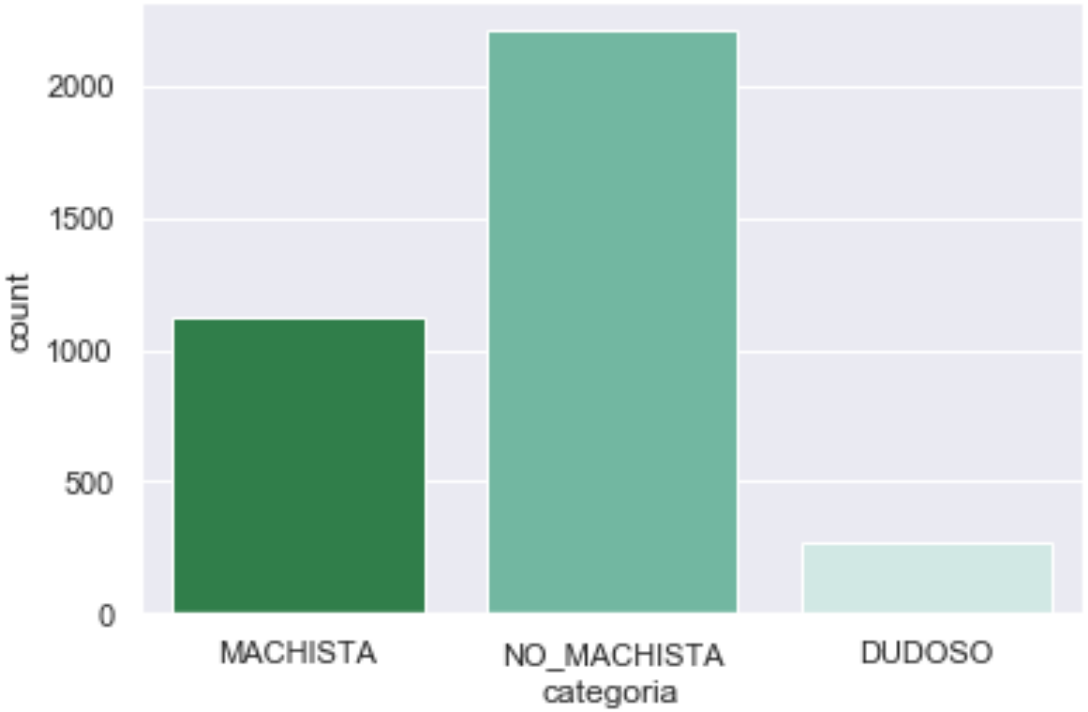
- Tres etiquetas distintas: MACHISTA, NO_MACHISTA y DUDOSO.
  - *“@EmanuelGPA Lo irónico es que lo dice una mujer, que “naturalmente" debería callarse y dedicarse a la cocina, limpiar y criar hijos”*
  - *“@kenia773 @LuisCarlos POR CIERTO, EN TU FOTO DE PERFIL SE PUEDE OBSERVAR QUE ERES BASTANTE VARONIL, ASÍ QUE SI NO ERES MARIMACHO, EMPIEZA A SERLO”*
  - *“@hazteoir @PSOE Más vale que se marche a fregar!”*

3.4 RESULTADOS DEL ETIQUETADO

- 3600 mensajes etiquetados por tres anotadores.

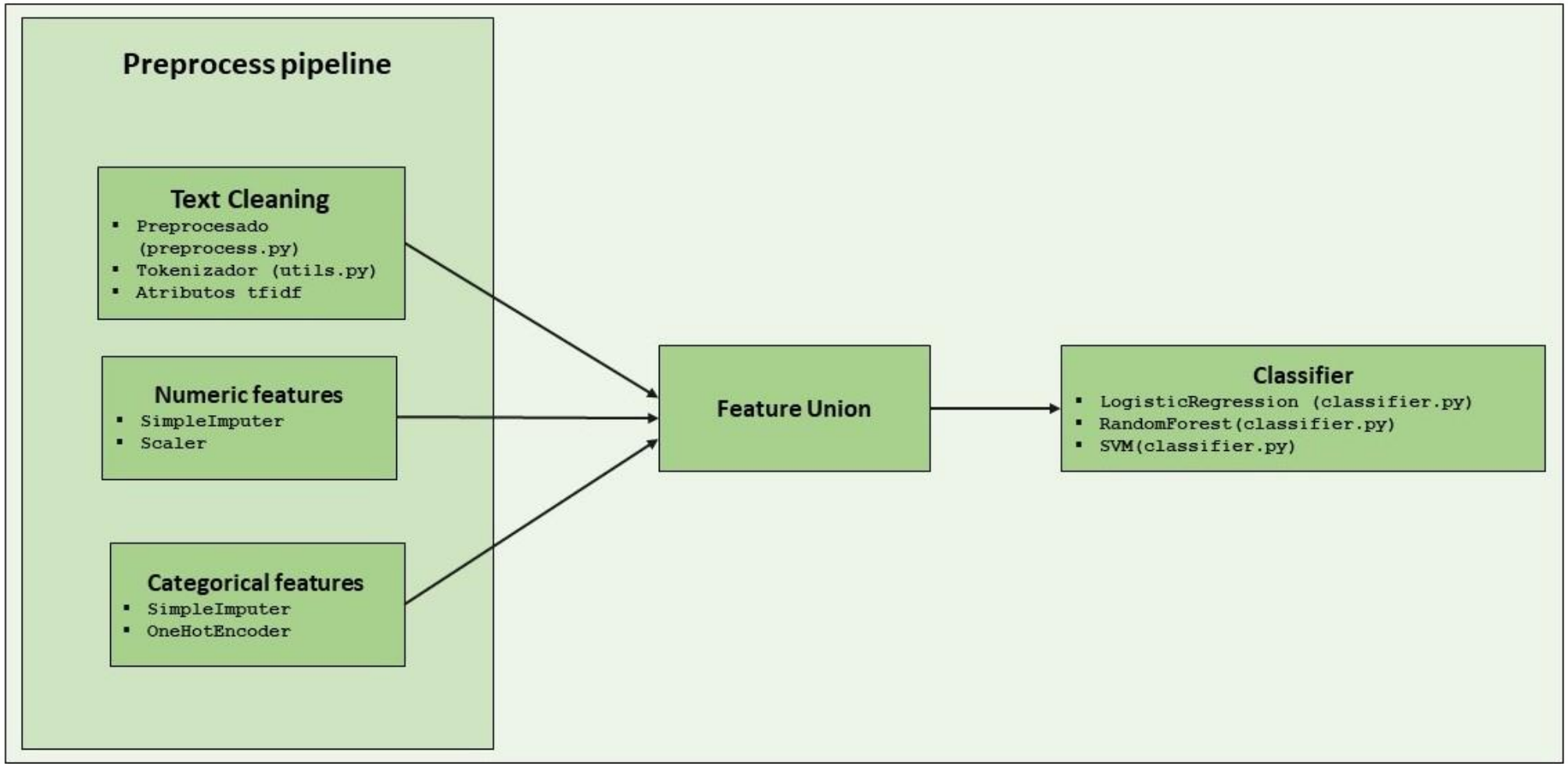
	Kappa
Etiquetador 1-2	0,68
Etiquetador 1-3	0,68
Etiquetador 2-3	0,88
Media etiquetadores	**0,75**

Etiqueta	Veces asignada
NO_MACHISTA	2181 (60,58 %)
MACHISTA	1152 (32 %)
DUDOSO	267 (7,42 %)



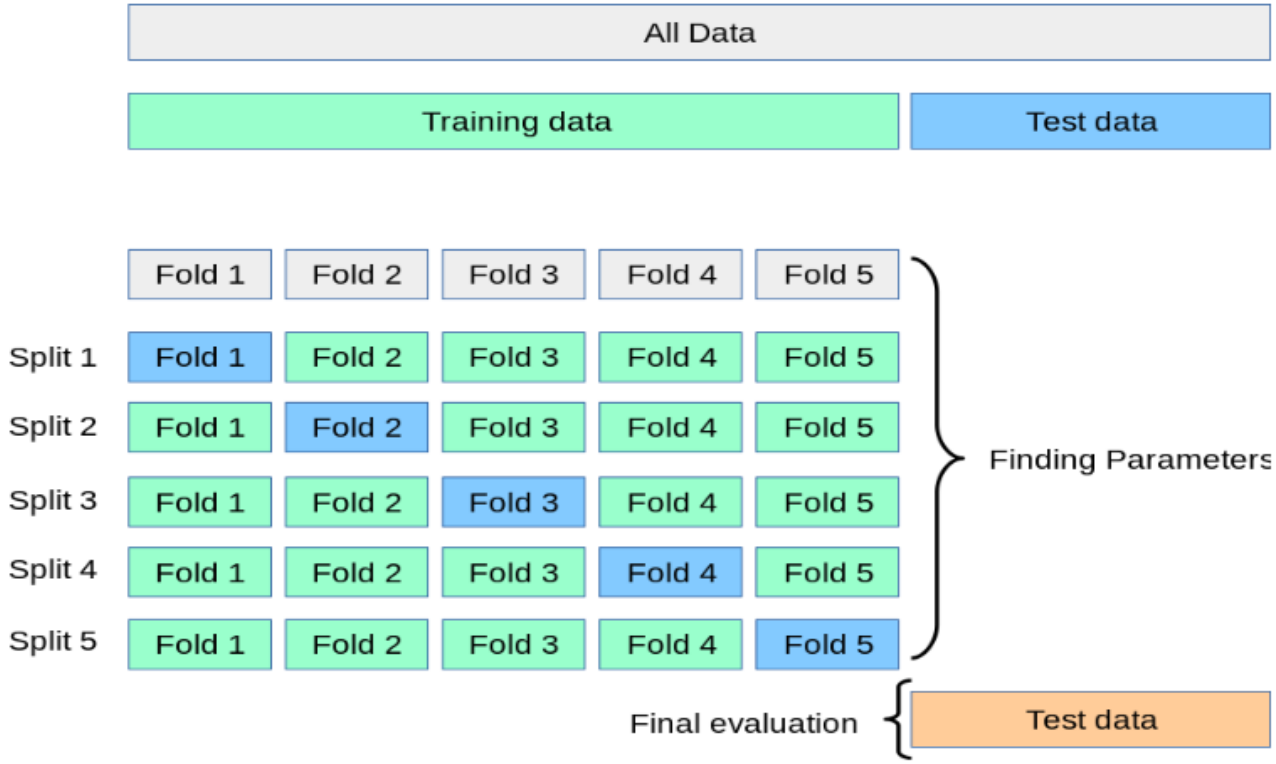
4. SISTEMA PROPUESTO Y EVALUACIÓN

4.1 ARQUITECTURA DEL SISTEMA



4.2 EVALUACIÓN

- **Experimento 1:** Búsqueda de hiperparámetros mediante la optimización de la medida F1.
- **Experimento 2:** Validación cruzada con parámetros por defecto.
- **Desbalanceo de la clase:** Muestreo de las clases mayoritarias para balancear la clase.



5. RESULTADOS Y ANÁLISIS DE ERRORES

5. RESULTADOS

- **EXPERIMENTO 1:**

	Accuracy	F1	Recall	Precision
Baseline (tf-idf)	0,68	0,59	0,62	0,59
Baseline	0,61	0,2	0,3	0,24
LR	0,7	**0,62**	**0,64**	0,62
RF	**0,72**	0,6	0,57	**0,67**
SVM	0,7	0,61	0,63	0,61

- **EXPERIMENTO 2:**

	Accuracy	F1	Recall	Precision
Baseline (tf-idf)	0,69	0,58	0,56	0,62
Baseline	0,61	0,2	0,3	0,24
LR	0,72	**0,63**	0,62	0,65
RF	**0,73**	0,61	0,58	**0,68**
SVM	0,72	**0,63**	**0,64**	0,63

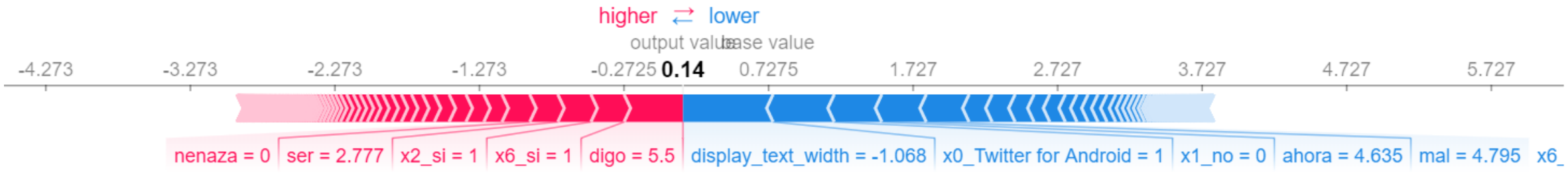
5. RESULTADOS

- **EXPERIMENTO DESBALANCEO DE LA CLASE:**

	Accuracy	F1	Recall	Precision
LR	0,61	0,61	0,61	0,62
RF	**0,68**	**0,68**	**0,68**	**0,68**
SVM	0,63	0,63	0,66	0,63

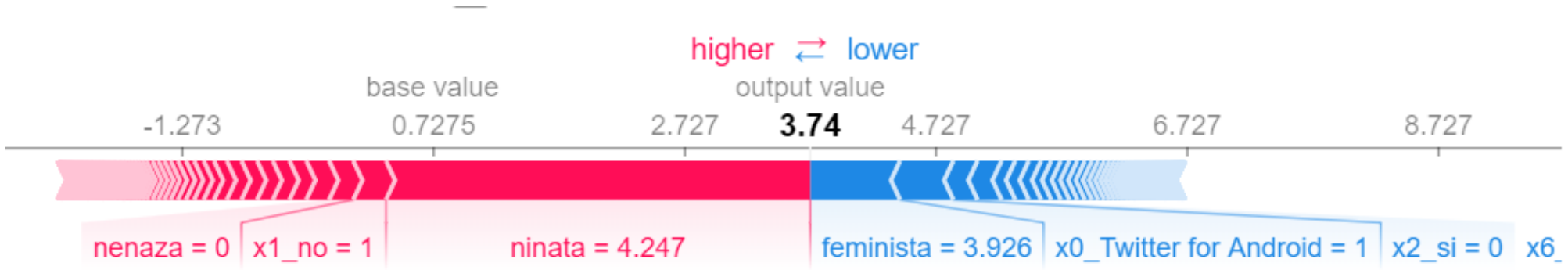
5. ANÁLISIS DE ERRORES

"@CopitoDeSnow_ Ahora es cuando digo "no está mal para ser mujer""



5. ANÁLISIS DE ERRORES

"Buscad mujeres con valores. No prestéis atención a ninguna niñata feminista. No os relacionéis con ellas, salvo para educarlas. No dejemos que nos coma el NOM"



6. CONCLUSIONES Y TRABAJOS FUTUROS

6. CONCLUSIONES

- Línea de investigación de interés actual.
- > 70% de tasa de acierto con método basado en frecuencia.
- Sesgo para términos concretos.
- Problemas para tener en cuenta el contexto.

6. TRABAJOS FUTUROS

- Investigación de los tipos de machismo.
- Ampliación para distintas fuentes de datos.
- Nuevas técnicas para clasificación y representación textual.
- Adaptación al inglés.
- Creación de un léxico machista específico (Wahyu y Patti, 2018).
- Portal web para explorar los tipos de machismo.

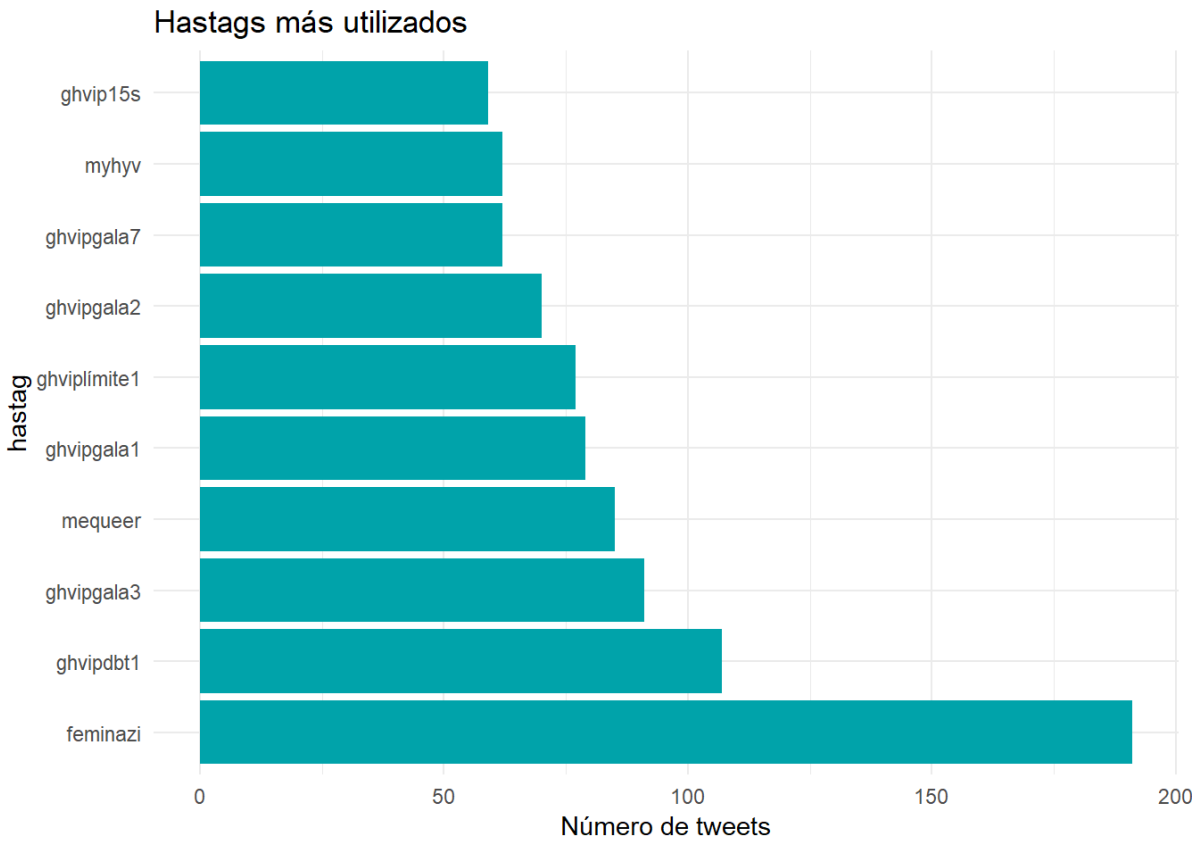
¿Preguntas?

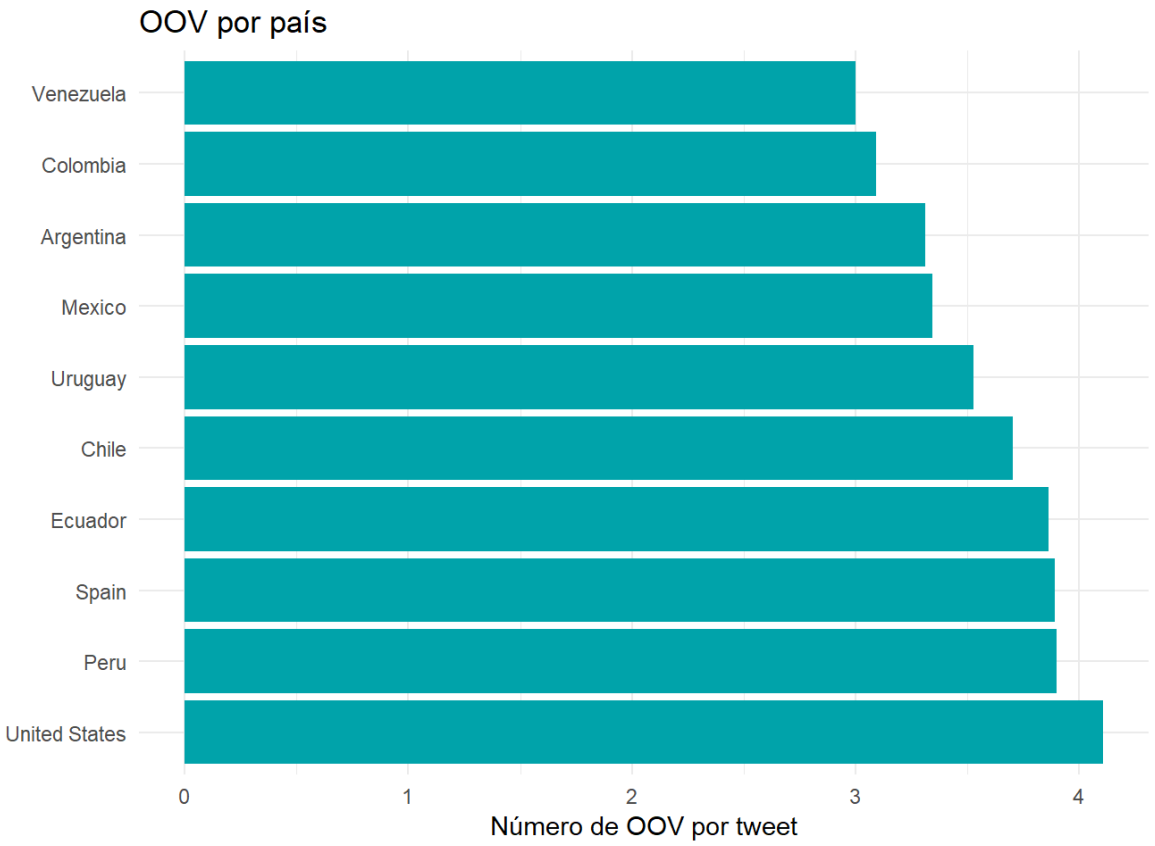
A. DIAPOSITIVAS DE APOYO

Número de término	Texto
1	“feminazi”
2	“a la cocina”
3	“a fregar”
4	“marimacho”
5	“ninata”
6	“mujer tenias que ser”
7	“las feministas”
8	“en tus dias”
9	“zorra”
10	“como una mujer”
11	“como una nina”
12	“pareces una fulana”
13	“pareces una puta”
14	“no ha probado un hombre”
15	“loca del”
16	“obsesionada con el machismo”
17	“para ser mujer”
18	“para ser chica”
19	“hombre que te aguante”
20	“acabaras sola”
21	“mojigata”
22	“mucho feminismo pero”
23	“mujer al volante”
24	“las mujeres no deberian”
25	“A las mujeres hay que”
26	“odio a las mujeres”
27	“las mujeres de hoy en dia”
28	“nenaza”
29	“lagartona”

Término	N
como una mujer	15094
feminazi	15093
a la cocina	15087
zorra	15086
loca del	15084
como una nina	15080
las feministas	15076
ninata	15032
en tus dias	14190
a fregar	14013
mojigata	6008
marimacho	5770
para ser mujer	4693
nenaza	4358
odio a las mujeres	2749
lagartona	2006
A las mujeres hay que	1845
las mujeres no deberian	1285
las mujeres de hoy en dia	991
mujer al volante	962
mucho feminismo pero	852
mujer tenias que ser	683
pareces una puta	474
para ser chica	180
acabaras sola	50
hombre que te aguante	37
obsesionada con el machismo	8
pareces una fulana	5
no ha probado un hombre	1

3.2 GENERACIÓN DEL CORPUS





3.3 DIFICULTADES EN EL ETIQUETADO

- Clasificación de tweets en los que el usuario que escribe el mensaje cita un contenido machista, en ocasiones con el que está en desacuerdo.
  - *"Pareces una puta con ese pantalón. - Mi hermano de 13 cuando me vio con un pantalón de cuero."*
  - *"Cada vez más a menudo (todos lo días) mi padre me dice que las mujeres no deberían recibir premios, trabajar en puestos superiores, que son putas, y que deben quedarse en casa y servir al hombre y criar hijos.*

3.3 DIFICULTADES EN EL ETIQUETADO

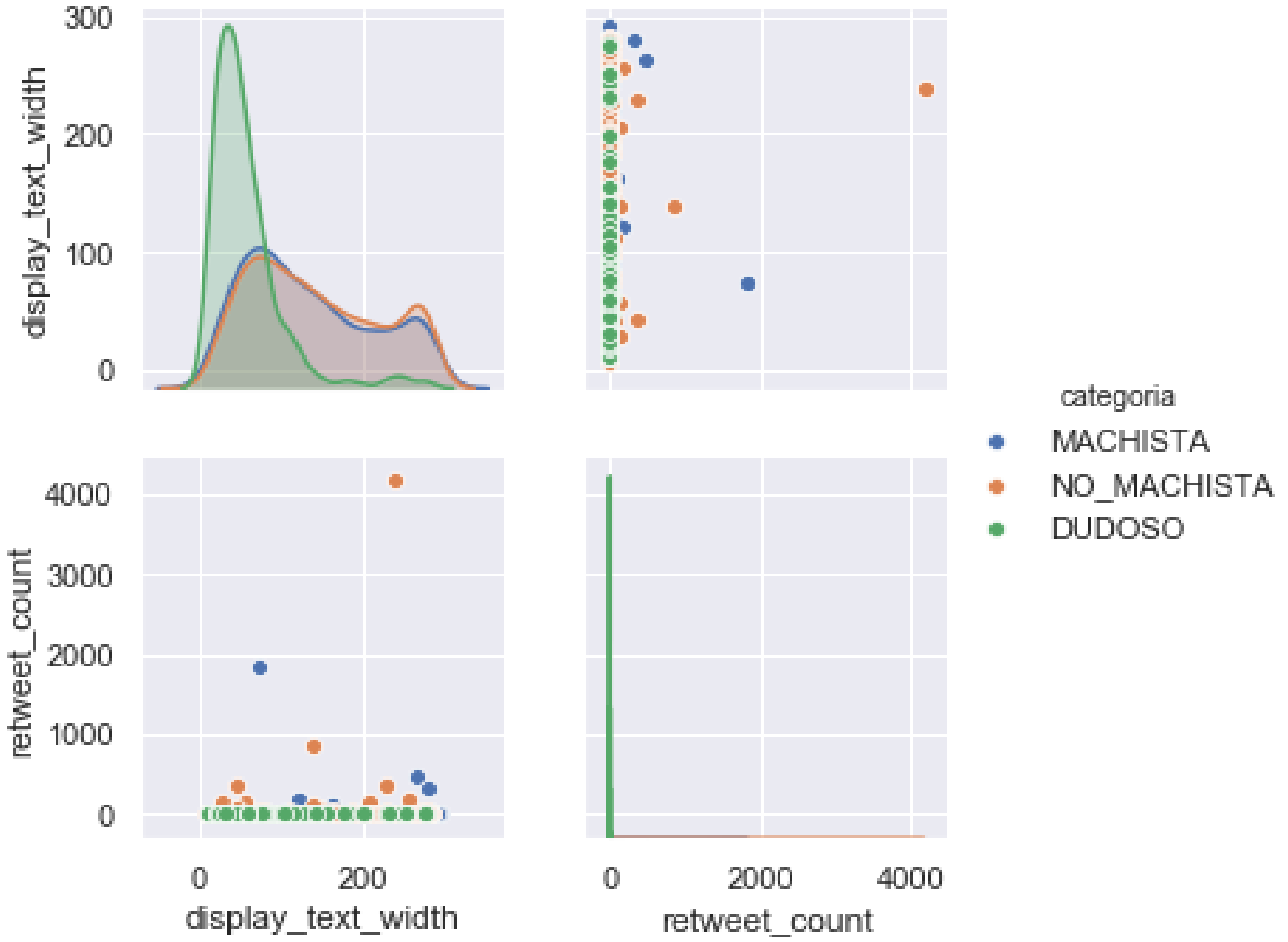
- Ambigüedad, sarcasmo e ironía:
  - *"@tonifreixa La zorra guardando las gallinas. ¡¡ Que se encargue Rosell ¡¡ Bueno..., cuando salga de la cárcel. Cinismo en grado máximo."*
  - *"@marijopellicer @radchiaru @_lxuli Jajajaja si no sabes cuanto odio a las mujeres tengo por favor no veas enemigos donde no los hay."*
  - *Mucho feminismo pero a la primera de cambio..... https://t.co/y2McecsgcT*
  - *Cuando subes a tu amiga la lagartona al Uber porque ya andaba malacopeando https://t.co/DcnK5ZGuL4*

	Kappa
Etiquetador 1-2	0,5
Etiquetador 1-3	0,44
Etiquetador 2-3	0,58
Media Etiquetadores	**0,51**

Tabla 4.4: Kappa obtenido con el 20 % del etiquetado

	Kappa
Etiquetador 1-2	0,76
Etiquetador 1-3	0,78
Etiquetador 2-3	0,74
Media Etiquetadores	**0,76**

Tabla 4.5: Kappa obtenido con el 20 % del etiquetado tras la corrección


display_text_width
300
200
100
0
retweet_count
4000
3000
2000
1000
0
0
200
display_text_width
0
2000
4000
retweet_count
categoria
MACHISTA
NO_MACHISTA
DUDOSO

Tokenizer for tweets.

```
>>> from nltk.tokenize import TweetTokenizer
>>> tknzr = TweetTokenizer()
>>> s0 = "This is a cooool #dummysmiley: :-) :-P <3 and
some arrows < > -> <--"
>>> tknzr.tokenize(s0)
['This', 'is', 'a', 'cooool', '#dummysmiley', ':', ':-)',
':-P', '<3', 'and', 'some', 'arrows', '<', '>', '->', '<--
']
```

Examples using *strip_handles* and *reduce_len parameters*:

```
>>> tknzr = TweetTokenizer(strip_handles=True,
reduce_len=True)
>>> s1 = '@remy: This is waaaaayyyy too much for you!!!!!!'
>>> tknzr.tokenize(s1)
[':', 'This', 'is', 'waaayyy', 'too', 'much', 'for', 'you',
'!', '!', '!']
```

4. TEXTO

- Reemplazo de emojis.
- Filtrado de URLs.
- Filtrado de usuarios.
- Convertidor de hastgs: #FelizDía → Feliz Día
- Convertidor a minúsculas
- Reemplazo de interrogaciones, exclamaciones y signos de puntuación.
- *TweetTokenizer → Stopwords, normalización y stemming.*

4. ATRIBUTOS NUMÉRICOS

- display text width: número de caracteres del tweet.
- favorite count: número de veces que el tweet ha sido marcado como favorito.
- retweet count: número de veces que el tweet ha sido retwiteado.
- followers count: número de seguidores del usuario que publica el tweet.
- friends count: número de personas seguidas por el usuario que publica el tweet.
- listed count: número de listas en las que está inscrito el usuario que publica el tweet.
- statuses count: número de tweets publicados por el usuario que publicó el tweet.
- favourites count: número de tweets que el usuario que publicó el tweet marcó como favorito.

4. ATRIBUTOS CATEGÓRICOS

- source: tipo de dispositivo con el que se publica el tweet.
- respuesta: indica si el tweet es una respuesta a otro.
- respuesta screen name: nombre del usuario al que se responde.
- hastag presence: indica la presencia de hastags en el tweet.
- url presence: indica la presencia de URLs en el tweet.
- media type: indica si el tweet contiene imagenes o videos.
- mentions presence: indica la presencia de la mención a algún usuario en el tweet.
- verified: indica si el usuario que publica el tweet es verificado por Twitter.

4. CLASIFICACIÓN

- **Línea base 1:** Se clasifican todos los registros del test con la categoría mayoritaria.
- **Línea base 2:** Se utilizan solo los atributos tf-idf con una LR con la siguiente búsqueda de parámetros: *C = [1, 10], class weight'= [None, 'balanced']*.
- **Regresión logística:** Se utiliza LR con todos los atributos con la siguiente búsqueda de parámetros: *C = [1, 10], class weight'= [None, 'balanced']*.
- **Random Forest:** Se utiliza RF con todos los atributos con la siguiente búsqueda de parámetros: *n estimators- [250, 450], bootstrap'= (True,False), max depth'= [None, 30].*
- **SVM:** Se utiliza SVM con todos los atributos con la siguiente búsqueda de parámetros: *C = [1, 10, 100, 10000], gamma = [0.001, 0.1, 0.6, 'auto', kernel = ['rbf', 'linear'].*

5. RESULTADOS

- **RF (EXPERIMENTO 1):**

RF	DUDOSO	MACHISTA	NO_MACHISTA
DUDOSO	**70**	32	80
MACHISTA	31	**407**	387
NO_MACHISTA	39	124	**1350**

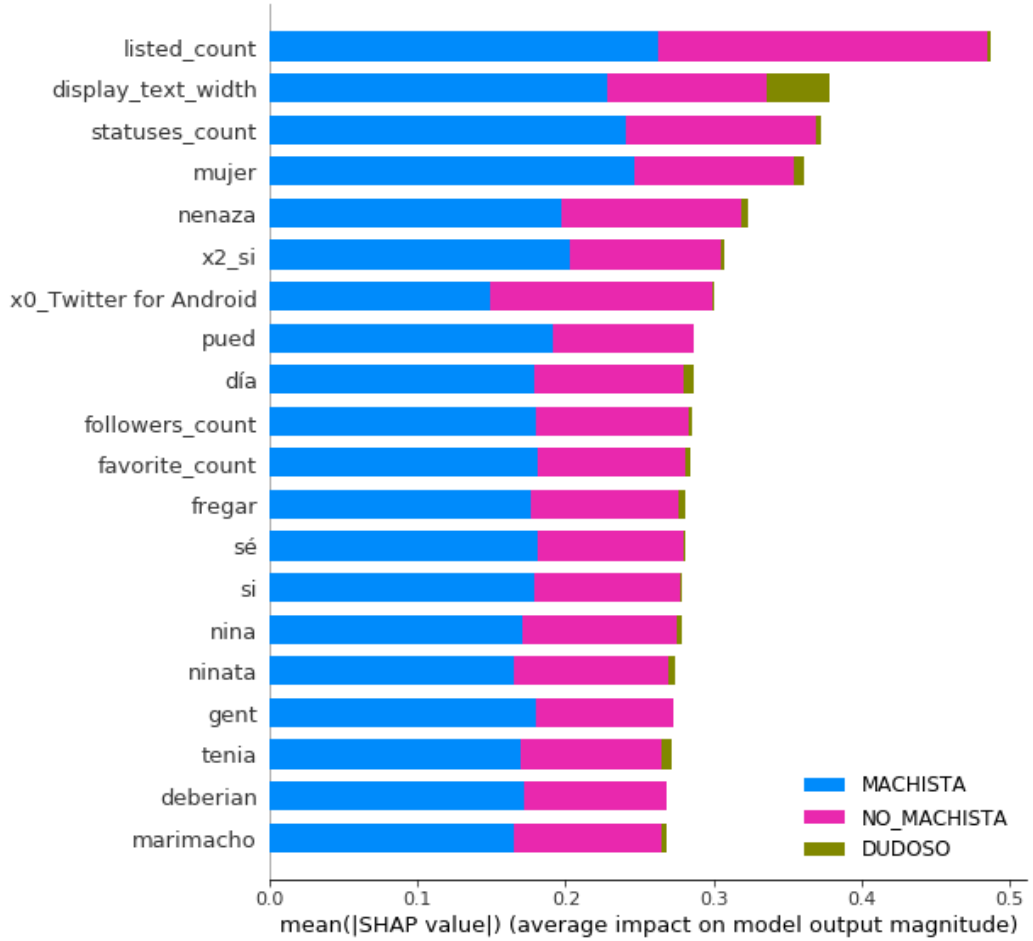
- **LR (EXPERIMIENTO 1):**

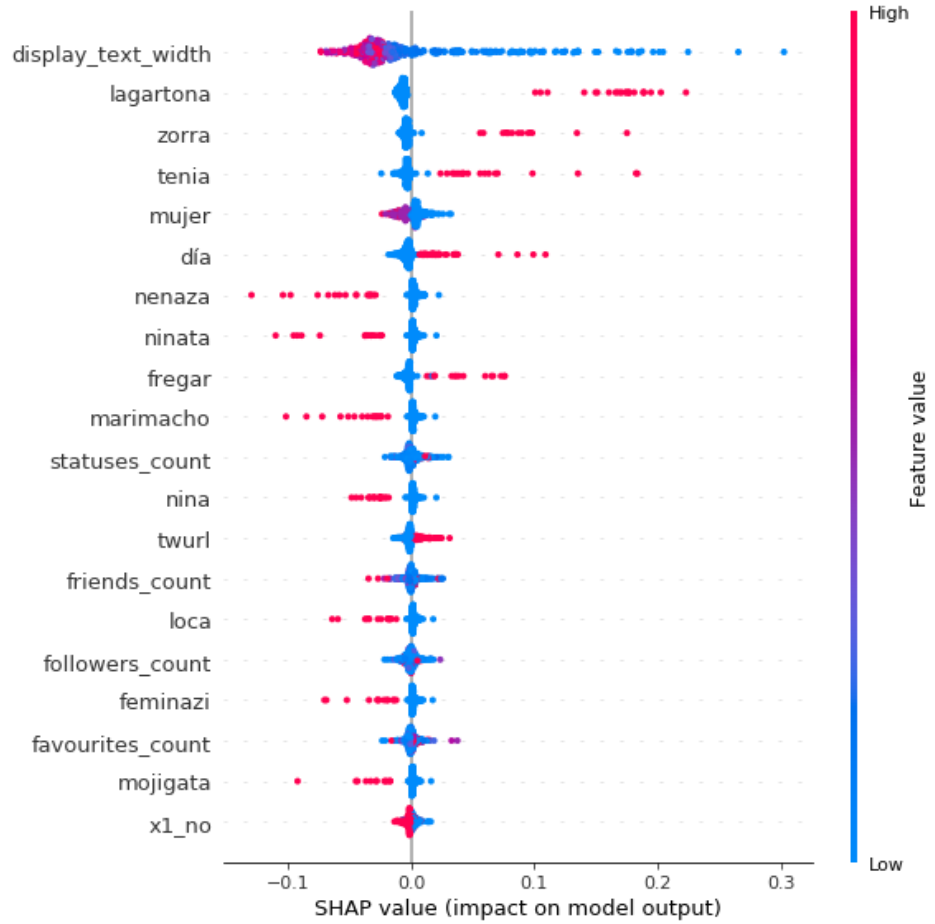
RF	DUDOSO	MACHISTA	NO_MACHISTA
DUDOSO	**116**	36	36
MACHISTA	96	**479**	247
NO_MACHISTA	91	250	**1169**

5. RESULTADOS

- Empeoramiento SVM: se trabajan con 222 atributos en total antes de aplicar el algoritmo de clasificación.
- RF: precisión del método para detectar los tweets no machistas
- SVM y LR similares: frontera de decisión lineal.

5. RESULTADOS





5. ANÁLISIS DE ERRORES

"@damita2808 @berege7 @Mariagtriana Y los ojos? Uff demasiado dureza en la mirada para ser chica...no?"