
quantil

Minería de texto en redes sociales:

Termómetro de Convivencia Digital
CandiData

Junio 1 de 2018

quantil

Objetivo

Termómetro de Convivencia Digital (MinTIC)

Estructuración de un instrumento de Ciencia de Datos para ser aplicado a dos casos de estudio en las redes sociales y comentarios en medios digitales en el contexto colombiano

- Caracterizar del uso de lenguaje tóxico.
- Estudiar el impacto que el ciberacoso y la agresión en línea está teniendo en las interacciones virtuales, la participación y la diversidad en la red.

Este estudio es un Piloto para orientar futuras direcciones de estudio.

Alcance del piloto

- Dos estudios de caso: las conclusiones no son generalizables al ámbito extenso de interacciones digitales.
- Casos: Partido de las Farc y Violencia de Género en el Ámbito Digital:
  - 5 temas generales: política, religión, género, fútbol y temas ambientales.
  - Relevancia: política ocupa el puesto 1; fútbol y género cercanos en el puesto 2.

Alcance del piloto

- Dos meses del piloto: la limitación de tiempo llevó a desarrollar una metodología innovadora, pero exploratoria.
- El estudio ayuda a contestar preguntas con soporte cuantitativo; se genera un amplio conjunto de nuevas preguntas.

Preguntas de investigación

¿Qué factores de análisis están asociados a la toxicidad en la redes?

- Dos pasos:

Nivel de toxicidad

Descritores:
Contexto
Emisor
Receptor
Audiencia

Correlaciones

Fases del Piloto

El Piloto se dividió en tres fases independientes:

1. Recolección, almacenamiento y tratamiento de datos
2. Analítica
   - Análisis de sentimiento
   - Clústering
   - Redes conversacionales
3. Desarrollo visualización

Datos

- Fuentes: seleccionadas por volumen de conversación alrededor de casos seleccionados
  - Twitter
  - Portales de noticias: El Tiempo, RCN, Pulzo
  - Facebook (exploratorio)

Tema	Inicial	Final
Violencia de género	2017-03-01	2017-05-31
Partido FARC	2017-01-01	2017-10-12

Cuadro: Rango de fechas

Datos

- Twitter tiene a disposición del público tanto los textos como la información básica de perfil para la mayoría de sus usuarios.
- Para recuperar tweets históricos relacionados con los temas de interés se realizó un procedimiento iterativo de scraping sobre la aplicación web de Twitter.
- Para recuperar los perfiles de los autores se utilizó el API REST de la plataforma.

Datos

- Las aplicaciones web de los principales medios digitales del país tienen a disposición del público secciones de comentarios en las páginas de sus noticias donde los lectores pueden expresar su opinón.
- Para capturar esta información se seleccionaron los tres medios digitales con mayor cantidad de noticias relacionadas con los temas de interés.
- Para cada una de estas noticias se realizó un procedimiento de scraping sobre los comentarios.
- Sin embargo, no es posible extraer información de los usuarios.

Datos

Datos recolectados:

	Farc	Género	Total
Twitter	40,370	29,347	69,717
Comentarios	4,787	5,613	10,400

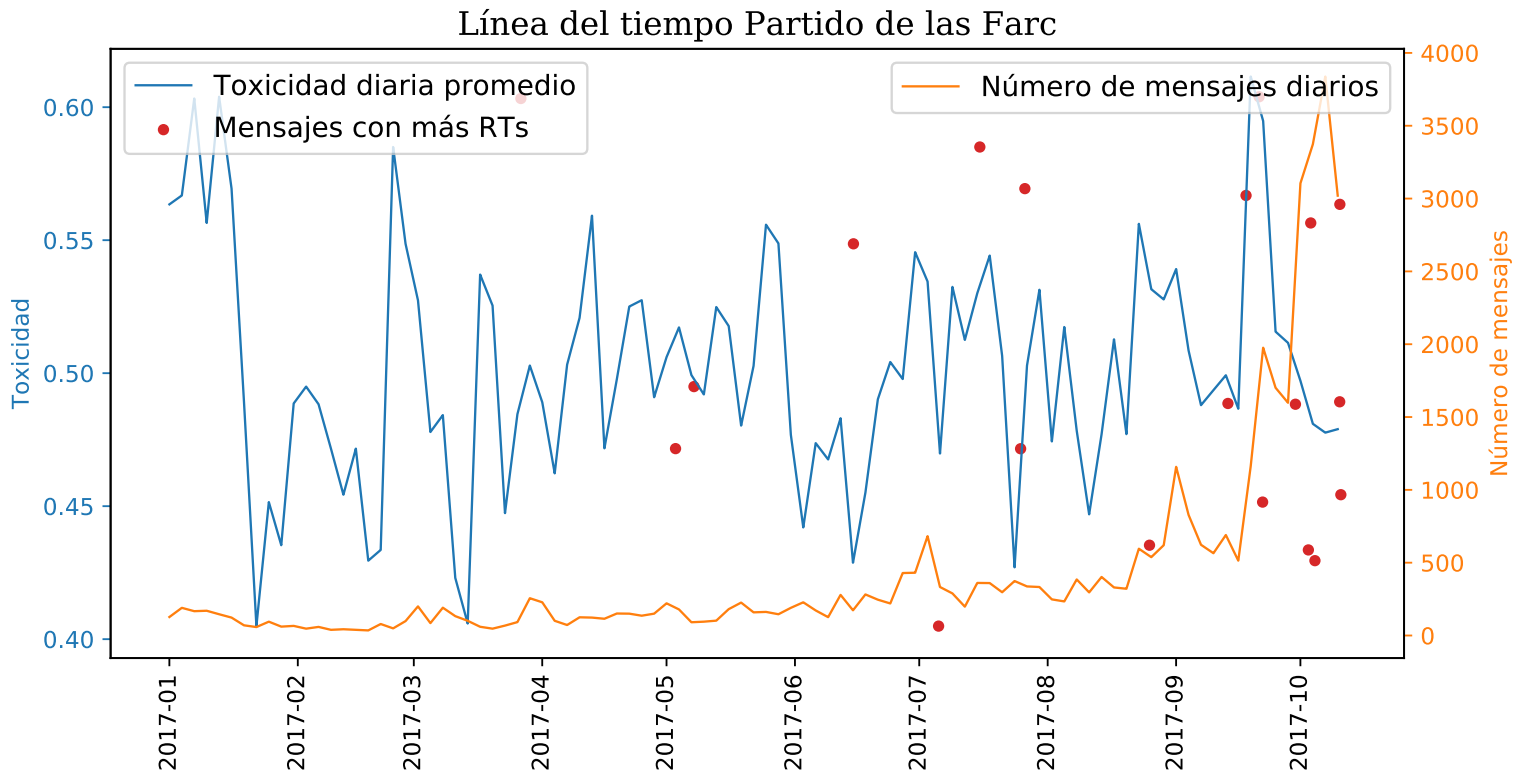
- Usuarios únicos Twitter: 35,919.
- Facebook exploratorio: se extrajeron 10 publicaciones en páginas publicas de 10 medios para el tema de Género.

Tratamiento de datos

- Por Ley 1581 de 2012 (Régimen General de Protección de Datos Personales) la información exraída constituye una colección de *datos sensibles*.
- Sin embargo, la Sentencia C-748/11 de la Corte Constitucional prohibe el tratamiendo de datos sensibles excepto cuando " (...) El Tratamiento tenga una finalidad histórica, estadística o científica."
- Por lo tanto, se implementó un procedimiento de supresión de identidad usando funciones hash criptográficas.

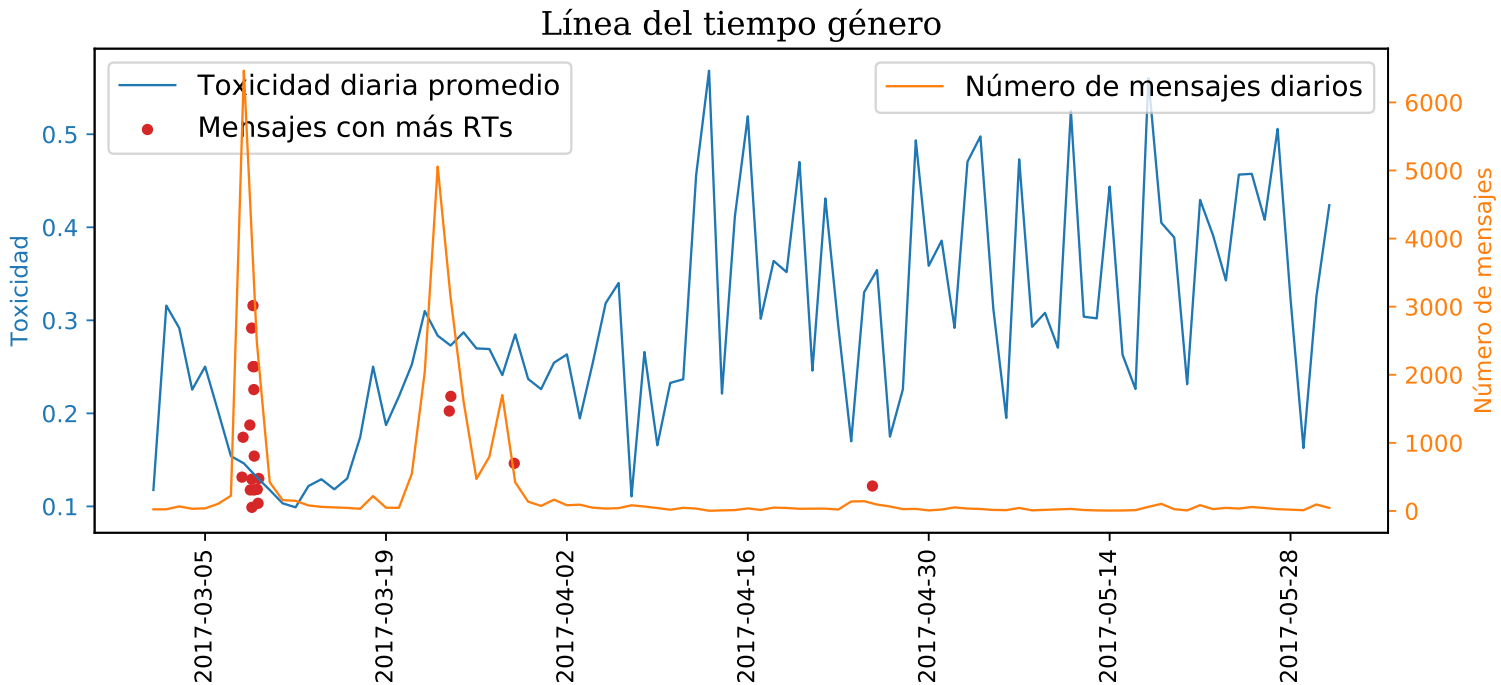
Descripción de caso Partido de las Farc

'@FARC_EPueblo Ivan marquez - no tienes nada que opinar perro hijueputa asesino de mierda malparido'



Descripción de caso: Violencia de Género

'#AndreaGuerreroPaLaCocina @AndreaGuerreroQ deje de joder que nadie le dice ni mierda por borracha. #NomasMujeresHablandoDeFutbol'



Análisis de Hashtags

- Usando expresiones regulares se extrajeron todos los *hashtags* utilizados en los tweets recolectados.
- Esto sirve como un análisis preliminar del tema y el tono de las conversaciones en las redes.

Hashtags Farc

- De los 40,370 tweets analizados, se identificaron 3,972 tweets con 1,573 hashtags diferentes.
- De los 78 hastags más repetidos, 36 tienen una connotación negativa, 36 neutra y 7 positiva.

Hashtag	Frecuencia	Hashtag	Frecuencia
#castrochavismo	34	#sialapaz	25
#terroristas	34	#fuerzatimo	25
#santos	32	#nuevopartido	18
#santosroboelplebiscito	24	#paz	13
#castrochavista	23	#nobeldepaz	11

Hashtags Género

- Se encontraron 4,654 de 29,347 tweets que usaban hashtags:

Hashtag	Frecuencia	Hashtag	Frecuencia
#diadelamujer	10546	#saqueenquilla	235
#diainternacionaldelamujer	840	#armero	184
#8m	397	#lapolemica	138
#mujeres	292	#andreaguerreroracista	82
#mujer	282	#conlatricolorpuesta	78
#8demarzo	227	#pabloarmeroesunguerrero	76
#niunamenos	222		

Análisis de sentimiento

- Dado el enorme volumen de comentarios en la red, no es posible analizar esta información mensaje por mensaje
- Se desea clasificar el nivel de toxicidad, provocación y calma de cada mensaje
- Metodologías de análisis de sentimiento (aprendizaje supervisado)

Análisis de sentimiento

- De acuerdo con la definición de Google a través del Api Perspective, se define **toxicidad** en discusiones como: “un comentario rudo, irrespetuoso o poco razonable que muy probablemente te haría irte de la discusión.”
- Se define el nivel de **provocación** (*troll*) de un mensaje respondiendo a la pregunta: ¿es el mensaje provocativo y dan ganas de responder para entablar un diálogo potencialmente tóxico?
- La variable **calma** se definió respondiendo la pregunta: ¿el mensaje llama a bajar el tono del debate, pide calma o tolerancia?

Análisis de sentimiento

- Se marcaron 1500 tweets y 500 comentarios para cada uno de los estudios de caso (partido de las Farc y violencia de género)
- La idea, a grandes rasgos, es que el algoritmo identifique características que hacen que un mensaje sea tóxico, a partir de ejemplos de mensajes marcados por humanos como tóxico o no tóxico
- Análogamente con el resto de variables de interés

Análisis de sentimiento

- Los textos constituyen datos no estructurados que deben ser traducidos es necesario traducirlos a un lenguaje matemático con el que el computador pueda trabajar.
- Se utilizó la metodología de **Bag of Words**:
  - Se crea un vocabulario juntando las palabras de todos los mensajes en un conjunto
  - Para reducir la dimensión del problema y disminuir el ruido de los datos, se eliminan palabras muy comunes (presentes en mas del 80 % de los textos), y palabras poco comunes (en menos de 5 textos)
  - Finalmente se eliminan palabras que no contienen mucha información como 'el', 'la', 'que', etc. (*stopwords*).

Análisis de sentimiento

- Luego, se construyó la **Matríz término-documento** *(DTM)* con pesos **tf-idf**:
  - Con el vocabulario construído, se calcula la frecuencia con que cada palabra aparece en un mensaje y se pondera por el porcentaje de los mensajes totales en los que aparece la palabra
  - Se construye una matriz donde cada fila representa un mensaje y cada columna un término del vocabulario
  - La entrada i, j de la matriz aumenta con el número de veces que aparece la palabra j en el mensaje i y disminuye con la porción de textos en el que aparece la palabra j

Análisis de sentimiento

Análisis supervisado:

- Para cada mensaje, se desea calcular la probabilidad de ser tóxico o provocativo (y), a partir de su vector (X) correspondiente de la DTM:

$$P(y = 1|X)$$

Modelo Logístico:

- Supone que la probabilidad de interés está dada por:

$$P(y_i = 1|X_i) = \frac{1}{1 + e^{-w^T \cdot X_i}},$$

  con $w^T = [\beta_0, \beta_1, \ldots, \beta_n]$ los parámetros del modelo.
- El parámetro β_j representa el peso de la palabra j en el modelo.

Modelo Logístico:

- Se entrena el modelo encontrando los parámetros que minimizan la siguiente función de costo, que captura el desvío entre el modelo y los mensajes marcados:

$$J(w, C) = \sum_{i=1}^{m} \left(y_i \log \left(\frac{1}{1 + e^{-w^T x_i}} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + e^{-w^T x_i}} \right) \right) + \frac{1}{C} \sum_{i=0}^{n} |\beta_i|.$$

con C un parámetro de regularización.

Análisis de sentimiento

- Para evaluar el poder predictivo del modelo y calibrar los parámetros, se realizó un procedimiento de validación cruzada usando la métrica de Área bajo la curva ROC
- 4 modelos: toxicidad y provocación para cada caso de estudio
- Otras metodologías de análisis supervisado: Naive Bayes, SVM, Boosted Trees

Análisis de sentimiento

Nombre del modelo	**Valor inverso de regularización**						
	0.1	**0.5**	**1.0**	**2.0**	**5.0**	**10.0**	**25.0**
Regresión Logística	0.50	0.73	0.76	0.76	0.72	0.69	0.67
Support Vector Machine	0.73	0.75	0.74	0.72	0.70	0.69	0.68
Boosted Trees	0.74	0.74	0.73	0.73	0.71	0.70	0.63
Naive Bayes	0.63	0.63	0.63	0.63	0.63	0.63	0.63

Cuadro: Área bajo la curva ROC para distintos modelos de **toxicidad** y valores del parámetro de regularización C. Caso de estudio: **Partido de las Farc**.

Análisis de sentimiento

Nombre del modelo	Valor inverso de regularización 0.1	0.5	1.0	2.0	5.0	10.0	25.0
Regresión Logística	0.58	0.65	0.65	0.66	0.64	0.62	0.59
Support Vector Machine	0.66	0.66	0.65	0.65	0.63	0.61	0.60
Boosted Trees	0.65	0.64	0.65	0.65	0.65	0.63	0.63
Naive Bayes	0.55	0.55	0.55	0.55	0.55	0.55	0.55

Cuadro: Área bajo la curva ROC para distintos modelos de **provocación** y valores del parámetro de regularización C. Caso de estudio: **Partido de las Farc**.

Análisis de sentimiento

Nombre del modelo	Valor inverso de regularización 0.1	0.5	1.0	2.0	5.0	10.0	25.0
Regresión Logística	0.72	0.80	0.80	0.79	0.76	0.75	0.74
Support Vector Machine	0.79	0.78	0.77	0.76	0.76	0.75	0.74
Boosted Trees	0.75	0.75	0.75	0.74	0.72	0.68	0.66
Naive Bayes	0.65	0.65	0.65	0.65	0.65	0.65	0.65

Cuadro: Área bajo la curva ROC para distintos modelos de **toxicidad** y valores del parámetro de regularización C. Caso de estudio: **violencia de género**.

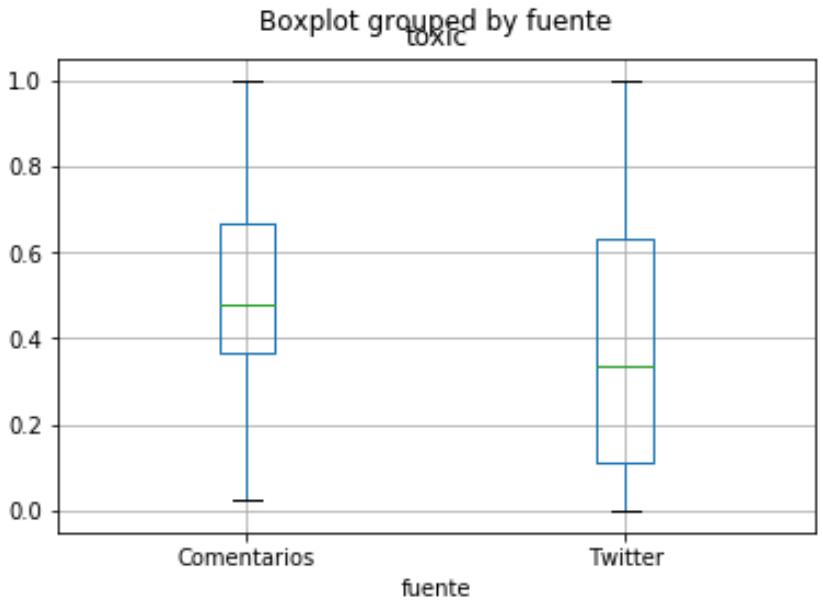
Análisis de sentimiento

Nombre del modelo	Valor inverso de regularización 0.1	0.5	1.0	2.0	5.0	10.0	25.0
Regresión Logística	0.50	0.57	0.61	0.63	0.63	0.63	0.62
Support Vector Machine	0.63	0.63	0.63	0.63	0.63	0.63	0.62
Boosted Trees	0.62	0.62	0.62	0.62	0.60	0.59	0.58
Naive Bayes	0.57	0.57	0.57	0.57	0.57	0.57	0.57

Cuadro: Área bajo la curva ROC para distintos modelos de **provocación** y valores del parámetro de regularización C. Caso de estudio: **violencia de género**.

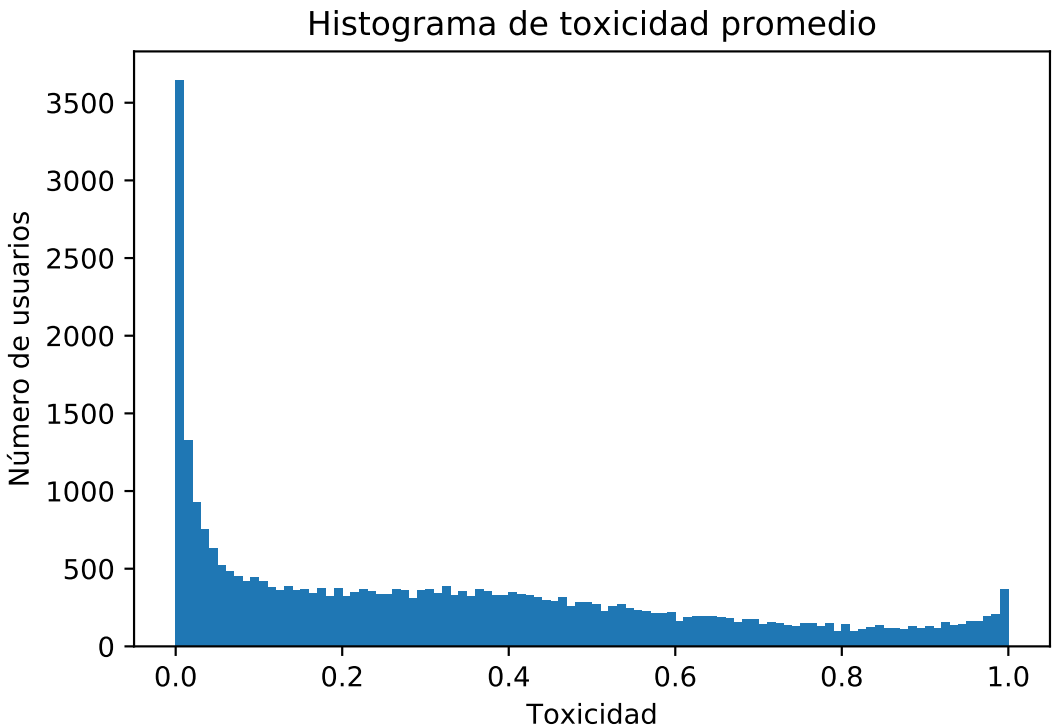
Análisis de sentimiento

La toxicidad promedio de los comentarios en los portales de los medios de comunicación utilizados fue de 0.42, superando a los tweets cuya toxicidad promedio fue de 0.38.



Análisis de sentimiento

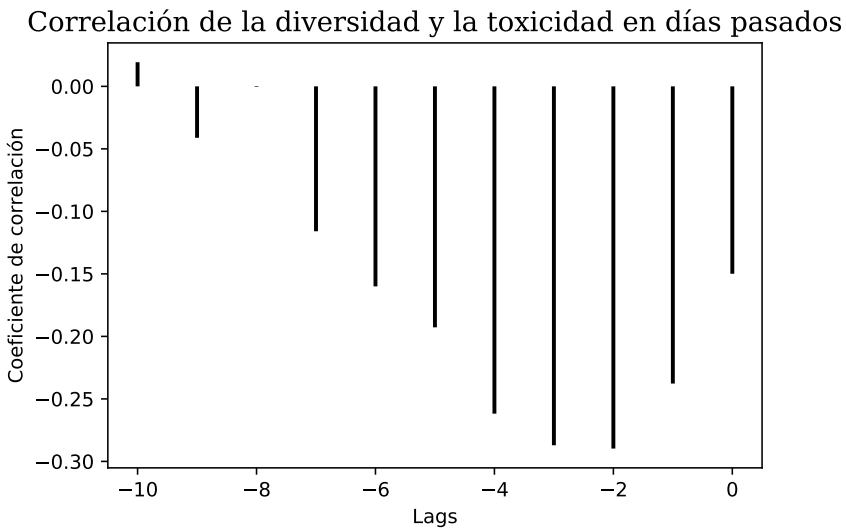
¿Se concentra la toxicidad en unos pocos?



	Toxicidad
count	31919.00
mean	0.33
std	0.29
25 %	0.06
50 %	0.28
75 %	0.52

Análisis de sentimiento

¿Los picos de toxicidad reducen la diversidad de la red? ¿Es decir, se pierde calidad del debate cuando hay un pico de insultos?



- Diversidad: número de participantes dividido por el número de mensajes (3 días).
- Entre más toxicidad en los días anteriores, menos diversidad.

Análisis de sentimiento

Modelo de calma:

Entre los 2000 tweets marcados para el tema Género, se encontraron 15 textos invitando a la calma en el debate (0.5 %).

- *el tuvo un error como todo ser humano y ya pidió disculpas ya la esposa lo perdona ya dejen de criticarlo*

Para el tema de las Farc se marcaron 17 textos como mensajes de calma (0.56 %)

- *932656 Aunque no estoy de acuerdo al 100 % en lo que opina, me gustó bastante su estilo para comentar.*

Clusters

- Adicionalmente, se buscó segmentar a los usuarios según las características básicas del perfil y otras variables construídas
- Grupos de interés: Anónimos, Visibles, Bots.
- Metdologías de agrupamiento (aprendizaje no supervisado): $K-$medias

Clusters

$K-$medias:

Si tenemos un conjunto de M observaciones $\{x_1, \ldots, x_M\}$, queremos encontrar conjuntos $C_1, \ldots, C_K$ que agrupen los datos, y que cumplan:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{x_1, \ldots, x_M\}$: cada observación pertenece al menos a uno de los K grupos.
2. $C_k \cap C_{k'} = \emptyset$ para todo $k \neq k'$: los grupos no se sobrelapan, i.e., ninguna observación pertenece a mas de un grupo.

Clusters

$K-$**medias:**

- Dado el número de clusters deseados K, queremos encontrar subgrupos de la base de datos $C_1, \ldots, C_k$, tales que la suma de sus respectivas varianzas sea lo más pequeña posible.
- Matemáticamente,

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} Var(C_k) = \min_{C_1,\ldots,C_k} \sum_{k=1}^{k} \left[\sum_{j=1}^{n} \left(\sum_{x_i \in C_k} (x_{i,j} - \bar{X}_j^k)^2 \right) \right]$$

- K se escogió utilizando la regla del codo.

Clusters

Bots:

cluster	count	followers_count mean	tweets_day mean	verified mean	msg_count mean	mean_repeats mean	mean_unique mean	retweets mean	score
0	33280	2,398.4	5.6	0.0	1.5	1.1	0.3	1.0	0.0
1	36	9,008.6	25.5	0.0	9.9	134.0	18.5	0.1	0.5
2	657	196,167.1	24.2	1.0	2.2	1.6	0.6	15.0	0.0
3	26	122,647.2	930.2	0.2	4.9	7.2	4.0	1.3	1.0
4	27	248,205.2	13.1	0.6	3.9	3.3	2.0	507.5	0.0
5	17	7,320.0	128.0	0.0	151.9	1.9	0.5	1.2	0.5
6	37	3,917,430.7	160.9	1.0	8.8	4.7	2.6	36.8	0.0
t 7	433	29,281.2	157.2	0.0	3.5	3.1	1.9	2.0	0.5
8	64	937.6	5.0	0.0	1.2	109.2	103.9	1.2	0.0
9	517	2,392.9	22.3	0.0	23.5	1.7	0.2	2.2	0.0
10	1	16,237,961.0	40.6	1.0	1.0	3.0	3.0	53.0	0.0
11	343	2,552.6	25.6	0.0	1.4	27.5	26.4	0.5	0.5
12	133	87,865.7	16.1	0.3	3.8	1.6	0.8	171.1	0.0

Clusters

Anónimos:

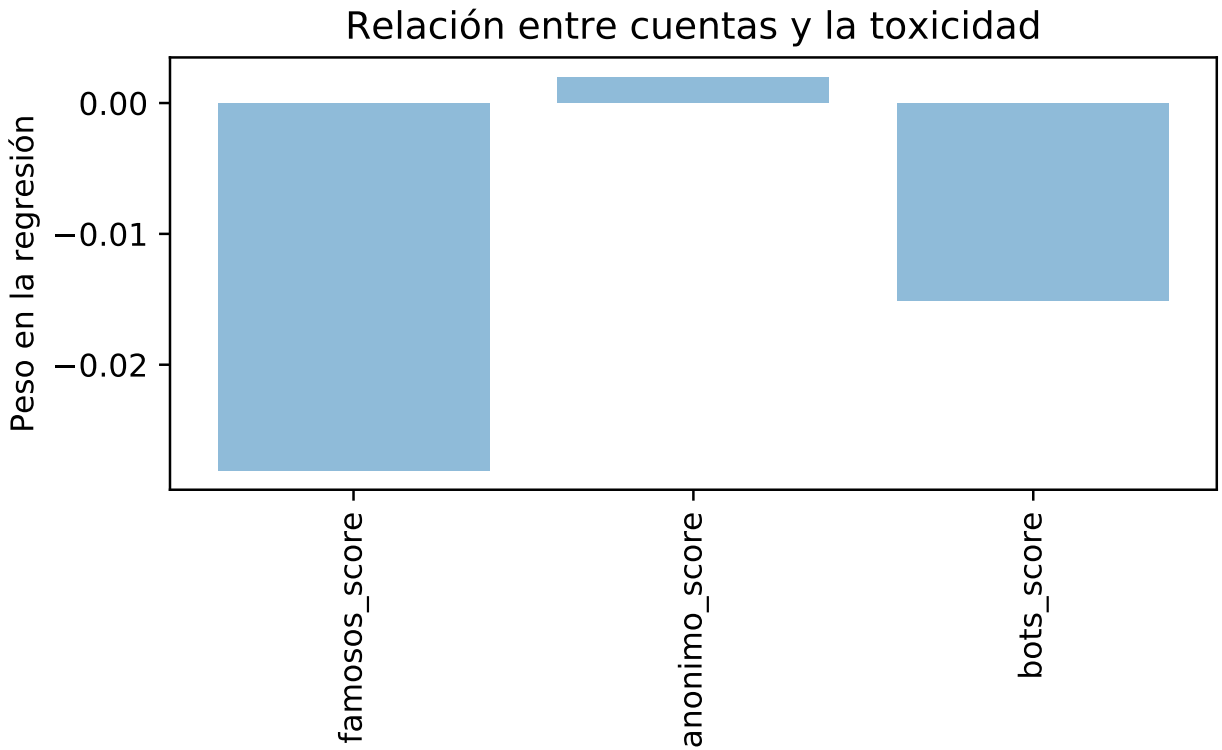
cluster	count	geo_enabled mean	verified mean	default_profile_image mean	gender_def mean	score
0	2679	1.0	0	0	0.0	0.5
1	13221	1.0	0	0	1.0	0.0
2	10883	0.0	0	0	1.0	0.5
3	1735	0.2	0	1	0.9	0.5
4	752	0.7	1	0	0.6	0.0
5	6301	0.0	0	0	0.0	1.0

Clusters

Visibles:

cluster	count	followers_count mean	verified mean	friends_count mean	following mean	listed_count mean	retweets mean	score
0	34641	2,421.1	0.0	759.2	0.0	18.8	1.0	0
1	50	1,638,087.4	0.9	5,776.8	0.0	5,367.6	25.2	1
2	82	276,594.6	0.3	2,439.2	1.0	603.2	29.7	1
3	608	124,496.3	1.0	2,523.2	0.0	515.9	13.9	1
4	19	4,846,204.2	1.0	19,734.2	0.4	14,794.1	45.9	1
5	3	1,696,799.0	0.7	673,827.7	0.0	4,005.7	3.6	0
6	116	85,662.6	0.3	3,333.3	0.0	305.6	191.6	1
7	2	11,346,163.0	1.0	798.0	0.0	52,453.5	36.9	1
8	20	300,493.6	0.6	2,144.7	0.1	890.5	560.3	1
9	30	226,817.3	0.1	109,075.9	0.0	1,243.8	24.0	0

Clusters

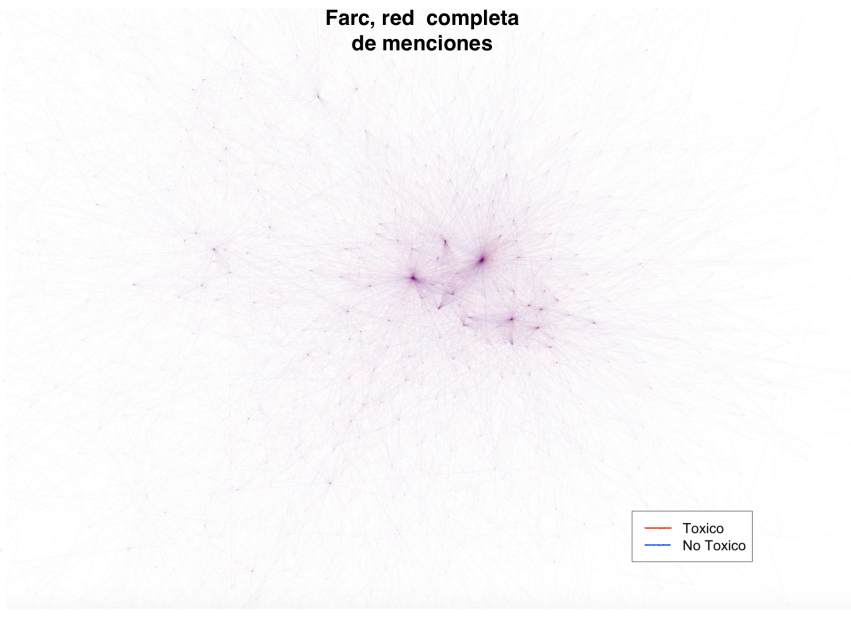

Relación entre cuentas y la toxicidad
Peso en la regresión
0.00
−0.01
−0.02
famosos_score
anonimo_score
bots_score

Redes

- Para visualizar las conversaciones al rededor de los dos temas en estudio, se construyeron redes de menciones entre los usuarios.
- Los nodos corresponden a los usuarios que participaron de la conversación, y existe una arista entre el nodo X y el nodo Y si el usuario X menciona en alguno de sus tweets al usuario Y

Redes

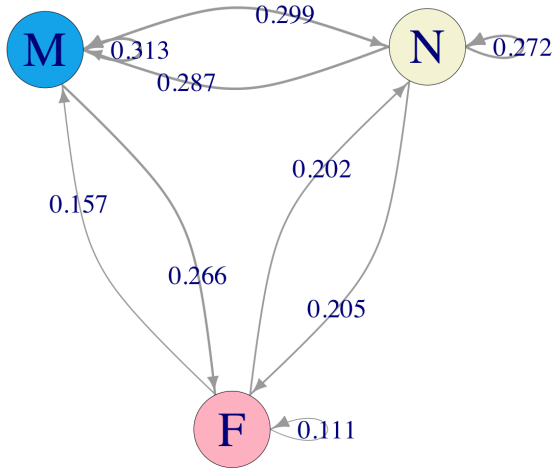
El caso de Andrea es un claro caso de acoso y el de las Farc es más un debate público como tal.



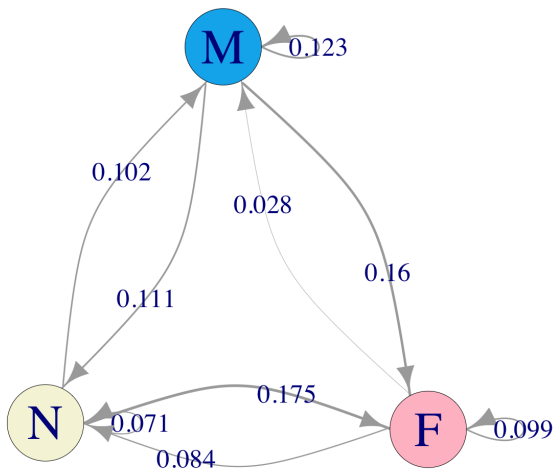
Volver

Redes

Farc: Proporción de Mensajes Toxicos entre Género



Género: Proporción de Mensajes Toxicos entre Género



Volver

Visualización

- Para facilitar la presentación e interpretación de los resultados obtenidos se desplegó un aplicativo local:

http://www.enticconfio.gov.co/termometro/

CandiData

- Análisis de redes sociales para los candidatos presidenciales en Colombia:

http://app.candidata.co

GRACIAS