quantil

BOLETÍN DE MATEMÁTICAS APLICADAS A LA INDUSTRIA

Termómetro de Violencia Digital

Diego Jara*, Simón Ramírez†, Mateo Dulce y Carlos Reyes ‡

No. 3

21 de marzo de 2018

Introducción

Para gran parte de la población, las redes sociales se han posicionado como el principal canal de comunicación y dispersión de información. La posibilidad de compartir un pensamiento, de aumentar la visibilidad de una marca, o de reenviar una noticia simultáneamente a miles de personas sigue planteando una nueva arquitectura para la dinámica social del país y del mundo.

Sin embargo, la posibilidad de mantener un perfil oculto y de evitar un enfrentamiento físico, en ocasiones ha aumentado el nivel de agresión en la forma en que los ciberciudadanos se comunican entre sí. El MinTIC, motivado por la evidencia de una baja cobertura de análisis matemático en la dinámica de estas conversaciones, ha impulsado el estudio riguroso de las interacciones digitales. Este es uno de los primeros esfuerzos en Colombia por cuantificar esta dinámica.

Manteniendo como enfoque el entendimiento del tipo de análisis y resultados factibles mediante técnicas de *Machine Learning* en el entorno de interacciones en redes sociales, se planteó como alcance el universo de conversaciones en Twitter y comentarios en medios noticiosos (de los que se recolectaron más de 80,000 elementos para el estudio) relacionados con dos temas puntuales: uno de ámbito político, y otro del ámbito de violencia de género en el entorno digital.

Por un lado, se desarrolló un esquema de aprendizaje supervisado (con marcaciones humanas realizadas a 4,000 mensajes) para clasificar todos los mensajes y usuarios según su toxicidad y su nivel de provocación. Por otro, se utilizaron técnicas de segmentación para guiar la identificación de usuarios anónimos, usuarios con alta visibilidad, y "bots". Finalmente, se buscó identificar el género de cada usuario a partir de los nombres reportados en las redes.

Con esta información, se estudió la correlación y nivel de asociación de distintas variables, y se construyeron redes dirigidas para analizar la dinámica de la interacción entre los usuarios. Esto permitió descubrir realidades de la dinámica digital que, si bien no son generalizables a todas las temáticas e

*Co-Director General, Quantil.
†Director de IT, Quantil.
‡El equipo del trabajo incluyó en adición a Alejandro Feged, José Luis Gutiérrez, Germán González, Olga Barrios y Felipe González.

Resumen

Este estudio aplica técnicas modernas de estadística para analizar la violencia digital en redes sociales. Más precisamente, se definen dos temas que hayan generado polémica, para analizar el tono del lenguaje usado por participantes en redes sociales. En adición, se plantean técnicas para identificar actitudes provocadoras (típicas de *trolls*) y para clasificar a los usuarios según su anonimidad, visibilidad, y factibilidad de ser un "bot". Finalmente, se construyen redes dirigidas a partir de conversaciones en Twitter relacionadas con estos dos temas para analizar las interacciones de los participantes.

Principalmente, el estudio ayuda a corroborar la premisa de que el análisis matemático puede ser una herramienta valiosa para el análisis de interacciones entre usuarios de redes sociales. Aunque el alcance del estudio no permite generalizar, en el ámbito de los casos estudiados se observa: (i) entre más diversa la discusión, menos tóxica la interacción; (ii) comentarios a noticias son más tóxicos que Tweets; (iii) hay más usuarios no tóxicos que tóxicos, pero la toxicidad no se concentra en pocos usuarios; (iv) el uso de hashtags, ser un usuario visible y tener foto de perfil se asocia con mensajes menos tóxicos; (v) se presentan más mensajes tóxicos de hombres a mujeres que al revés; (vi) la audiencia es pasiva a la hora de tratar de calmar discusiones agresivas.

Boletín de Matemáticas Aplicadas a la Industria es una publicación de Quantil S.A.S. Las opiniones expresadas en los artículos son las de sus autores y no necesariamente reflejan el parecer y la política de la compañía o de su junta directiva.

MINTIC

GOBIERNO DE COLOMBIA

interacciones en redes sociales, permiten aterrizar conclusiones en el contexto de los dos temas analizados. Algunos de estos resultados corresponden con hallazgos generales de otros estudios: por ejemplo, los comentarios más tóxicos tienden a utilizar más signos de exclamación y mayúsculas, y menos hashtags. Sin embargo, otros resultados pueden considerarse más novedosos: por ejemplo, se encontraron más usuarios no tóxicos que tóxicos, pero esta toxicidad no se concentra en unos pocos usuarios.

El estudio encontró que los dos casos estudiados presentaban dinámicas muy diferentes: uno era más centralizado en un usuario, casi revelando un patrón de acoso, y el otro parecía más un debate público, sin un nodo central. Como conclusión general, este estudio permitió demostrar la posibilidad de combinar la ciencia de datos y teoría de redes con información no estructurada - mensajes y comentarios - para identificar en tiempo real aspectos estructurales de la agresión digital en las redes.

Datos y Descriptivas

Para los dos casos de estudio seleccionados ("Grupo Político" y "Violencia de Género en Ámbitos Digitales") se recolectaron 69,717 tweets, y 10,400 comentarios históricos de los portales de los principales medios digitales del país, como se resume a continuación:

Plataforma \Tema	Grupo Político	Género
Twitter	40,370	29,347
Comentarios	4,787	5,613

Cuadro 1. Datos por tema y plataforma de información.

Adicionalmente, se extrajo la información (id, número de seguidores, seguidos, favoritos, entre otros) de 35,919 usuarios de Twitter que participaron en las conversaciones de los temas de interés.[4]

Dado el alto volumen de comentarios, no se analiza cada mensaje individualmente. En contraste, se utiliza aprendizaje de máquinas a partir de una muestra de entrenamiento marcada manualmente por humanos. Particularmente, en este estudio se buscó clasificar los niveles de toxicidad, provocación y calma de cada mensaje de acuerdo con las siguientes definiciones:

1. **Toxicidad**: 1, si es un comentario rudo, irrespetuoso o poco razonable que muy probablemente te haría irte de la discusión Wulczyn, Thain, y Dixon (2016); 0, de lo contrario.
2. **Provocación**: 1, si el mensaje es provocativo y dan ganas de responder para entablar un diálogo potencialmente tóxico Buckels, Trapnell, y Paulhus (2014); 0, de lo contrario.
3. **Calma**: 1, si el mensaje llama a bajar el tono del debate, pide calma o tolerancia; 0, de lo contrario.

Para este ejercicio, se marcaron 1500 tweets y 500 comentarios para cada caso. Posteriormente, se entrenaron distintos modelos de clasificación (ver sección: Modelamiento Matemático) y se predijo el nivel de toxicidad, provocación y calma para cada mensaje en la base de datos, usando el mejor modelo de clasificación entrenado.

Finalmente, para los usuarios que contaban con múltiples tweets sobre los temas de interés, se utilizó un promedio simple de la toxicidad, provocación y calma de sus mensajes, con el objetivo de otorgar una única calificación para cada usuario.

Modelamiento Matemático

Análisis de sentimiento

Existe una gran variedad de algoritmos para clasificar texto de acuerdo con su positividad o negatividad; esto se conoce en la literatura como análisis de sentimiento. Los algoritmos utilizados en este ámbito van desde regresiones logísticas hasta modelos complejos de análisis semi-supervisado y redes neuronales recurrentes.

Buscando una alta eficacia en la clasificación de mensajes tóxicos o provocativos, se probaron modelos relativamente sencillos que se pueden entrenar con un número reducido de datos: regresión logística, Naive Bayes, Boosted Trees y Support Vector Machines con kernel lineal. Estos modelos son muy versátiles y aplicables a una gran variedad de problemas; en particular, han sido aplicados con éxito en muchos problemas de análisis de sentimiento. En el Cuadro 2 se muestra el desempeño de los distintos modelos para la clasificación de toxicidad sobre un conjunto de validación. Debido a la ambigüedad en la definición y marcación del nivel de provocación, los modelos alcanzan un máximo de 0.76 de área bajo la curva ROC. La característica de "provocación" fue más retante, alcanzando un área máxima de 0.66.

	Valor inverso de regularización						
Nombre del modelo	**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 2. Área bajo la curva ROC para distintos modelos de toxicidad y valores del parámetro de regularización C. Caso de estudio: Grupo Político.

Al agregar los mensajes por usuario se obtiene el nivel de toxicidad del usuario; el histograma de nivel de toxicidad de

[4] Para los comentarios no fue posible extraer la información de los autores.

los usuario se muestra en la Figura 1.

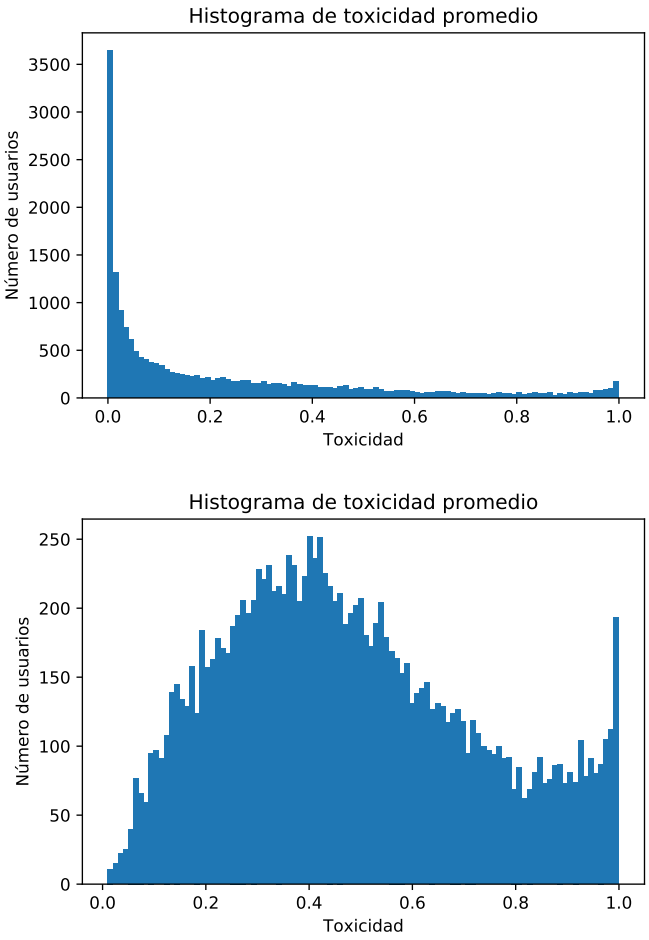

Figura 1. Nivel de toxicidad de los usuarios; arriba: caso de género; abajo: caso Grupo Político. En promedio las conversaciones alrededor del grupo político eran más tóxicas.

Segmentación de usuarios

Para enriquecer el análisis, se identificaron usuarios anónimos, usuarios visibles en la conversación y cuentas automatizadas ("bots"). Se tomaron distintas características de los perfiles de los usuarios de Twitter para segmentarlos en distintos grupos utilizando un algoritmo de k-medias. Por ejemplo, para identificar a los usuarios anónimos se utilizaron las variables que indican si la cuenta es verificada, si el nombre utilizado tienen un género definido, si tiene habilitado el sistema de georreferenciación y si ha cambiado la imagen original de su perfil. Una vez se identifican los grupos, se les asigna manualmente a cada uno una calificación de anonimidad según las variables de cada grupo tal como se muestra en el Cuadro 3. Para la caracterización de bots y de usuarios visibles se procede de la misma manera (usando otras variables).

cluster	count	geo_enabled mean	verified mean	default_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

Cuadro 3. Segmentación de perfiles en grupos de distinto grado de anonimidad.

Resultados

El análisis permite identificar dinámicas de las conversaciones de forma cuantitativa, en cuatro dimensiones definidas por el MinTIC: Contexto, Emisor, Receptor y Audiencia. Algunos resultados a resaltar son los siguientes:

- Contexto
  - Se observa (levemente) más toxicidad en comentarios a noticias que en tweets.
  - Entre más diversa la discusión, menos tóxica la interacción; picos de toxicidad tienden a ser seguidos por reducciones en diversidad.
  - El comportamiento de las redes asociadas a cada caso de estudio es distinto: parece poder asociarse a una dinámica de debate público en el caso del Grupo Político, y de acoso conversacional en el caso de violencia de género en ámbitos digitales.
- Emisor
  - Hay más usuarios no tóxicos que tóxicos, pero esta toxicidad no se concentra en pocos (ver Figura 1).
  - Comentarios tóxicos tienden a utilizar más signos de exclamación y mayúsculas, y menos hashtags.
  - Usuarios visibles (con cuentas verificadas o con muchos seguidores) tienden a ser menos tóxicos; igualmente los usuarios que tienen foto del perfil tienden a ser menos tóxicos.
- Receptor
  - Se observan más mensajes tóxicos de hombres a mujeres que de mujeres a hombres, y de hombres a hombres que de mujeres a mujeres.
  - Los usuarios más atacados son más centrales en la red.
- Audiencia
  - Se evidencian muy pocos mensajes de calma.
  - Los mensajes más tóxicos en promedio exhiben menos retweets, pero la diferencia con mensajes no tóxicos es pequeña.
  - Momentos más tóxicos de conversaciones fueron seguidos por momentos con poca interacción entre usuarios.

Herramienta

Con el fin de facilitar la apropiación de la investigación, se desarrolló una aplicación web que permite a los usuarios interactuar con los modelos estimados. La aplicación cuenta con una interfaz gráfica sencilla - desarrollada en Javascript utilizando el framework Angular y las librerías Plotly y D3 para las visualizaciones - que ofrece cuatro vistas principales:

- Estadísticas descriptivas y caracterización de los textos capturados para cada uno de los temas.
- Estadísticas descriptivas de los modelos predictivos y termómetro de convivencia digital.
- Librerías de términos asociados a dimensiones de la violencia en entornos digitales.
- Visualización de toxicidad en redes dinámicas de menciones.

(a) Estadísticas descriptivas

(b) Modelos predictivos

(c) Termómetro de características

(d) Análisis de redes

Figura 2. Funcionalidades selectas de la aplicación web.

Consideraciones Finales

El trabajo desarrollado evidencia el valor de analizar cuantitativamente las interacciones digitales, y sugiere bastantes direcciones de posible trabajo futuro. Por ejemplo, hacer seguimiento de temas generales y más casos específicos, profundizar en las definiciones de *trolls*, y hacer monitoreo en tiempo real, son algunas de las direcciones en las que se puede extender lo que se presenta en este trabajo. Confiamos que el futuro cercano convocará más a la investigación cuantitativa de interacciones sociales en ámbitos digitales en el país.

Referencias

Buckels, E. E., Trapnell, P. D., y Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and individual Differences*.

Wulczyn, E., Thain, N., y Dixon, L. (2016). Ex machina: Personal attacks seen at scale. *CoRR, abs/1610.08914*. Descargado de http://arxiv.org/abs/1610.08914