Recuperación, procesamiento y clasificación de tuits para visualizar estructuras de interacción

Carlos Pérez[1], Jorge Cortés[1], Aarón Ramírez[1], Rocío Abascal-Mena[1], Alejandro Molina-Villegas[2]

[1] Universidad Autónoma Metropolitana-Cuajimalpa, México
{2143805174, 2143805218, 2113066212}@alumnos.cua.uam.mx
mabascal@correo.cua.uam.mx
[2] Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, México
amolina@conabio.gob.mx

Resumen. En un contexto de medios sociales digitales, donde existen múltiples formas de vinculación entre usuarios, resulta importante contar con herramientas que permitan analizar los procesos de interacción presentes en estas plataformas. El análisis de redes sociales utiliza frecuentemente diagramas nodo-enlace para representar las relaciones entre un conjunto de actores. Sin embargo, la representación visual de grafos con información adicional en vértices y aristas es una tarea compleja y pocos programas cuentan con esta característica. Presentamos una propuesta para la recuperación y procesamiento de tuits con el fin de visualizar redes de Comunicación Política en Twitter. El sistema incluye clasificación automática de usuarios, diferenciación del tipo de aristas dependiendo de si es una mención, una respuesta o un retuit, así como visualización interactiva de los grafos.

Palabras clave. Visualización de redes complejas, Interacción en medios sociales, Clasificación automática, Ciencias de datos en comunicación política

Recovery, Processing and Classification of Tweets to Visualize Structures of Interaction

Abstract. In the context of digital social media, where users have multiple ways to interact with others, it is important to have tools to analyze the interaction processes within these platforms. Social network analysis frequently uses node-link diagrams to represent relationships among social actors. However, the visual representation of network graphs with additional information in vertices and edges is a complex task and few programs provide this feature. We propose a system for the recovery and processing of tweets to visualize Political Communication networks. The

system includes the automatic classification of Twitter users, differentiation between retweets, mentions, and replies, as well as an interactive visualization of network graphs.

Keywords. Visualization of Complex Networks, Interactions on Social Media, Automated Classification, Data Science in Political Communication

1. Introducción

Los medios sociales se han instalado progresivamente en nuestra vida diaria, alterando los métodos de comunicación y los intercambios de información. Estas plataformas sociales continúan evolucionando y permitiendo nuevas formas de acción colectiva. A partir de ello, es posible identificar una nueva Comunicación Política apoyada en el uso de medios sociales. Los actores sociales trascienden de ser consumidores hacia productores creativos de información sustantiva, llamados prosumidores.

Twitter, con sus no más de 140 caracteres, se ha convertido en una herramienta de manifestación social y política en la que los prosumidores no sólo crean mensajes sino que reproducen y responden creando un medio de colaboración. Es así como el análisis de datos, recuperados de Twitter, puede proporcionar un medio para observar la sociedad contemporánea. A partir de los intereses, motivaciones y actitudes de los usuarios, es posible descubrir patrones de comportamiento [6]. Los tuiteros tienen a su disposición diversas formas de interacción, como las menciones, las respuestas, los retuits y los *likes*. Esta variedad presenta retos para el análisis de las interacciones en Twitter, especialmente cuando se asigna algún significado a cada uno de los tipos de interacción en el contexto de investigaciones específicas. Una forma frecuente de examinar las interacciones presentes en los medios sociales es generar grafos que representan relaciones entre actores con aristas (o enlaces) y vértices (o nodos), respectivamente. Sin embargo, no todos los programas especializados en análisis de redes permiten dibujar múltiples aristas entre un mismo conjunto de nodos. La última versión de Gephi (http://gephi.org) los soporta, pero únicamente en su laboratorio de datos y no en la representación gráfica; por su parte Social Network Visualizer (http://socnetv.sourceforge.net/) sí los muestra, pero de manera separada, es decir, dibuja un grafo por cada tipo de interacción.

Asimismo, hay complicaciones relacionadas con la recuperación y el filtrado de los datos, pues el software especializado suele restringir la personalización de la salida de datos recuperados. Los formatos empleados para su almacenamiento, al no estar destinados para su manipulación a través de distintos programas, pueden ocasionar pérdidas de información y, además, limitar su compatibilidad con herramientas de visualización. Una ausencia de capacidades para el tratamiento del corpus reduce las opciones para la depuración y extracción de características representativas de los conjuntos de datos.

Por todo lo anterior, en este artículo se proponen técnicas de recuperación, depuración y procesamiento de datos de Twitter para la visualización y análisis de datos relacionales complejos. La visualización se utiliza en el estudio de las interacciones entre ciudadanos, políticos y medios de noticias en Twitter, con el objetivo de conocer de qué manera interactúan dentro de este medio social.

El artículo está organizado de la siguiente manera: en la Sección 2 se presenta el estado del arte de los principales trabajos en el análisis de Twitter. La propuesta de recuperación, procesamiento y visualización del conjunto de datos es detallada en la Sección 3. Un análisis de los resultados es presentado en la Sección 4. Finalmente, se da un panorama general sobre el estado actual del trabajo y lo que se espera en un futuro.

2. Estado del arte

Desde el surgimiento de la Web 2.0, el modelo de Comunicación Política tradicional se ha redefinido al permitir un acercamiento de la ciudadanía con los políticos y los medios. Es así como el llamado régimen mediático se ha visto opacado con la participación de los ciudadanos en las redes sociales, quienes interactúan entre ellos creando, retransmitiendo e interactuado con otros actores. Esto amplía el espacio de opinión pública y, potencialmente, podría mejorar las condiciones democráticas a partir de una mayor participación y representación de la ciudadanía [12].

Son numerosos los estudios que se han centrado en el estudio de Twitter. En específico, se encuentran trabajos sobre la detección de actores influyentes ([4], [10], [14], [18]), el desarrollo de campañas políticas en Internet, tanto en México ([3]) como en otros paises ([5], [7], [9], [10]) y la predicción y análisis de los usuarios de acuerdo con lo que comparten y sus principales contactos como es el caso de ([1], [11], [17]).

Sin embargo, en los estudios antes mencionados, el análisis está centrado en el contenido y sentimiento expresado en los tuits o en la actividad de personajes específicos, no en las interacciones entre ellos.

Consideramos que las interacciones es un aspecto importante a estudiar ya que como se menciona en [15], las interacciones de los actores de no élite, nombrados por Chadwick [2], tienen mucho éxito al utilizar redes sociales como Twitter. Sabiendo que en cuanto más los medios tradicionales realicen difusión en los medios digitales será más probable que los ciudadanos activos, que utilizan las mismas herramientas, puedan influir en la cobertura de los medios de comunicación. De igual forma, los actores de no élite tienen cuidado en su interacción con las élites en línea, incluyendo políticos y periodistas, haciendo pues que la interacción tenga un papel importante en la cobertura de las noticias.

La propuesta está compuesta de 3 pasos que incluyen: 1) recuperación de tuits, 2) procesamiento y 3) visualización. La recuperación o minado de tuits es, generalmente, realizado a partir del uso de *hashtags* asociados a los tuits. Sin embargo, la detección del ruido causado por *hashtags* que no tienen relación alguna con el contexto es una tarea ardua. En general, las investigaciones lo

afrontan como un problema de clasificación proponiendo, en algunos trabajos, un enfoque supervisado basado en grafos con el fin de inferir las categorías de intención de los tuits ([6], [13], [19]).

Por su parte, la recuperación de tuits tiene como objetivo el filtrado mediante un análisis del tuit. Este procesamiento incluye la limpieza de los tuits a partir de una comparación con términos o *stopwords*. El filtrado, y en algunos casos jerarquización, ha sido muy estudiado para los casos de documentos estáticos. Sin embargo, en las redes sociales existen nuevos factores que lo vuelven difícil como lo es el uso de idiomas distintos ([3]), estilos fragmentados de redacción, la ambigüedad, y la restricción propia de los 140 caracteres como en el caso de Twitter. Algunos métodos están basados en modelos probabilísticos o clasificadores como Naïve Bayes. Sin embargo, la gran parte de los trabajos encontrados son aplicados en el análisis de sentimientos en cuyo caso es muy importante el análisis sintático y semántico de todo el tuit. No hay ejemplos de aplicaciones en el que se pueda estudiar la interacción de los actores sin tomar en cuenta el contenido del tuit.

A la fecha, el impacto de Twitter en la Comunicación Política, en particular en el contexto mexicano, no ha sido abordado suficientemente aún para ofrecer un panorama claro sobre las dinámicas entre actores dentro del medio. Es de gran importancia estudiar el tipo de interacciones que se dan entre los diferentes actores ya que, como lo menciona [15], los modelos de toma de decisión periodística y de Comunicación Política necesitan incorporar el papel de las plataformas en los medios sociales, como Twitter, debido a su gran importancia.

En la siguiente sección se presenta la propuesta para la recuperación, procesamiento y visualización de estructuras de interacción.

3. Propuesta de recuperación, procesamiento y visualización

A grandes rasgos, nuestra propuesta para obtener una visualización de conjuntos de datos para análisis tiene tres etapas principales: la Recuperación, el Procesamiento y la Visualización. Entre las características principales de este flujo, destacan que las primeras dos etapas consideran una salida de datos para su posterior visualización en un sistema desarrollado a medida. De igual manera, destaca la inclusión de procesamiento en paralelo para la generación de un modelo de clasificación automática de perfiles de usuario. La Figura 1 detalla el proceso.

3.1. Recuperación

Se automatizó el proceso de recuperación de información mediante un *script* para la captura de publicaciones mediante la API de Twitter empleando la librería Tweepy (http://www.tweepy.org/). Entre otros parámetros, se limitó la recuperación al idioma español y conteniendo los términos especificados incluidos en los metadatos determinados por la plataforma.

Como detalla la Figura 2, tras la obtención de cada tuit, se realizó una búsqueda de dos niveles de respuestas para obtener entradas relacionadas con la temática recuperada que no incluyeran necesariamente los términos inicialmente establecidos.

Uno de los principales problemas de la recuperación consistió en que al realizar peticiones repetidas se obtenían datos duplicados. Por lo tanto, se empleó el metadato *id* de cada publicación y se consultó su existencia en el conjunto de entradas recuperadas. Para evitar una disminución en el rendimiento del programa, debido al gran número de tuits almacenados, los archivos se guardaron en texto plano. Los documentos fueron separados por día de publicación –obtenido del metadato *created_at*– para reducir el tiempo de cómputo.

3.2. Depuración

Los datos recuperados fueron procesados con el fin de conservar publicaciones relevantes para su análisis. En la Figura 3 se muestra la serie de filtros ordenados, cuyos resultados fungieron como entrada del paso subsecuente para reducir el número de entradas.

El primer paso del proceso de depuración implicó descartar los tuits sin interacción y, por lo tanto, ninguna conexión con otros actores.

Definimos que un tuit tiene interacción si cumple con alguno de los siguientes criterios:

- es un retuit;
- menciona a otro usuario o;
- es una respuesta a un tuit previo.

Mediante estos criterios se conservaron aquellos tuits que no estaban aislados y eran susceptibles a representarse mediante un grafo.

Es posible encontrar tuits publicados por métodos automáticos conocidos como *bots* y cuyo propósito es popularizar o desprestigiar a una persona o un determinado tema. Una característica común en los *bots* reside en su plataforma

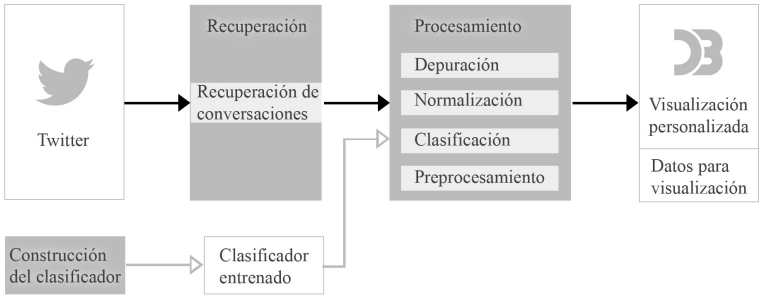

Fig. 1. Flujo de recuperación, procesamiento y visualización

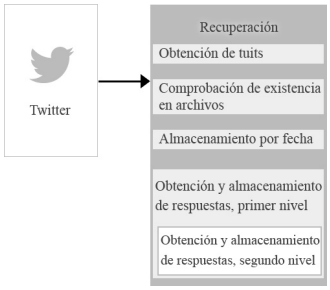

Fig. 2. Flujo de recuperación de publicaciones

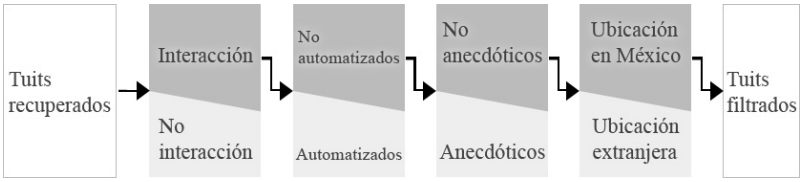

Fig. 3. Flujo de depuración de entradas

de publicación, diferente a las oficiales creadas por Twitter. Algunas de estas plataformas son IFTTT (https://ifttt.com/) y Tapbots (http://tapbots.com/). En nuestra propuesta, el módulo de depuración excluye este subconjunto, conservando así únicamente, los tuits presumiblemente auténticos. Cabe mencionar que aún cuando esta estrategia resultó suficiente para nuestros experimentos, no existe actualmente un método infalible para filtrar *bots*.

Se incluyó un filtro de tuits que tratasen de una experiencia anecdótica o contenido promocional. Con este fin, se creó una lista de 43 términos para detectar y excluir estas ocurrencias. Sólo se mantuvieron los tuits que no tenían ocurrencia alguna de los términos establecidos en la lista.

Finalmente, se descartaron los tuits con una ubicación geográfica fuera de México. Para este proceso se usó una lista de localidades de la República Mexicana y se conservaron los tuits emitidos desde alguna de las ubicaciones listadas y aquellos sin ubicación.

Al concluir el proceso de depuración, prevalecieron los tuits que tenían interacción, que no fueron publicados por *bots*, que no eran de carácter anecdótico o promocional y que fueron emitidos desde México.

3.3. Normalización y clasificación

Las cuentas presentes en el corpus depurado fueron clasificadas en tres categorías: medio, político y ciudadano. Este proceso de clasificación fue realizado manualmente en primera instancia y luego automáticamente mediante algorit-

mos de aprendizaje supervisado. En la Figura 4 se detallan los componentes del proceso para el entrenamiento del clasificador.

Se usó la descripción del perfil de cada usuario como criterio para determinar la clase de cada cuenta, así como su rol en Twitter a partir de esta información.

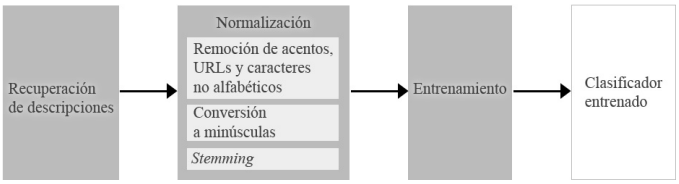

Fig. 4. Flujo de generación del clasificador bayesiano

Primeramente se clasificó de manera manual un subconjunto de perfiles pertenecientes a las categorías medio y político. Para este proceso, se automatizó la descarga de sus descripciones de perfil (biografías). Seguidamente, se empleó un *script* para remover hipervínculos, signos de puntuación y caracteres especiales, con la finalidad de conservar sólo caracteres alfanuméricos. El texto restante de cada descripción fue transformado a minúsculas y cada palabra reducida a su raíz léxica (*stem*) usando la implementación del algoritmo de Porter contenido en el módulo *Snowball Stemmer* de la biblioteca *NLTK*.

Como resultado del procesamiento descrito anteriormente, se obtuvo la representación de *Bolsa de Palabras* de un subconjunto de descripciones de perfiles de usuarios de medios y políticos. A este proceso se le conoce como codificación y sirve para transformar texto en vectores numéricos que la máquina pueda procesar. La manera precisa de codificar la información depende de cada problema particular pero lo que es indispensable para este tipo de métodos es transformar la información textual en datos vectoriales que son la materia prima del aprendizaje supervisado.

La clasificación automática de perfiles de usuario se realizó mediante técnicas de reconocimiento de patrones basado en el teorema de Bayes.

Básicamente, el teorema de Bayes (ver Ecuación 1), define una manera de calcular probabilidades condicionales. Por ejemplo, si $P(mediatico)$ es la probabilidad a priori de que un perfil de usuario sea un perfil de medios. $P(mediatico|x)$ sería la probabilidad a posteriori, de ser un perfil de medios, basada en el contenido del perfil, siendo x nuestra nueva observación, es decir, la codificación de un perfil de usuario de Twitter en forma de vector.

$$P(mediatico|x) = \frac{P(x|mediatico)P(mediatico)}{P(x)} \tag{1}$$

Así, para aplicar el teorema de Bayes al reconocimiento de patrones, se define $p(x)$ como la probabilidad de que exista una codificación como la de la entrada x.

Usando este marco metodológico, solamente hace falta conocer, a través de muchos ejemplos, cómo se comportan los elementos de una clase particular (descripciones de perfiles mediáticos, o sea la variable *mediatico*), y de ésta forma hallar la función de distribución asociada a esa clase, que será asímismo $p(x|mediatico)$. Lo que nos permite saber, a partir de la probabilidad de una clase, la probabilidad de dicha clase una vez obtenido el patrón x. En concreto, de entre todas las clases posibles, debemos escoger la de mayor probabilidad a posteriori.

Los datos de entrenamiento ingresados al clasificador de nuestro sistema fueron las cuentas de usuarios de medios y políticos representados por sus descripciones previamente clasificadas por inspección.

De este modo, el clasificador entrenado es empleado como parte del sistema para automatizar la categorización de nuevas cuentas. Aunque la precisión de dicho modelo no es perfecta, es lo suficientemente buena para ser usada en el sistema en producción. La Figura 5 muestra en detalle el número de instancias de perfiles clasificados como políticos que en efecto sí son políticos (verdaderos positivos para p). Análogamente, se muestran los verdaderos positivos de perfiles mediáticos (verdaderos positivos para m); así como los correspondientes falsos positivos para p y falsos positivos para m. Se puede deducir, a partir de esta matriz, que el 95.6% de perfiles es clasificado correctamente. En la sección 4. se muestran más detalladamente los resultados de la clasificación automática de perfiles de usuario.

		Clasificación automática	
		p	m
Real	p	1183	59
	m	40	1001

Fig. 5. Matriz de confusión resultante de la clasificación automática de perfiles de usuarios de twitter p= perfiles políticos; m=perfiles mediáticos

3.4. Visualización

Los diagramas nodo-enlace son empleados frecuentemente para representar visualmente datos relacionales, pues a partir de ellos es posible obtener una idea general de los patrones de actividad dentro de una red. Debido a ello, se decidió emplearlos para visualizar las interacciones entre los usuarios de Twitter.

El primer paso fue establecer parámetros visuales para los nodos y los enlaces. Cada nodo representa una cuenta de Twitter y su color indica la pertenencia a

una de las tres categorías. Siguiendo las recomendaciones de [16], se utilizaron tonos notoriamente distintos entre sí: magenta, cian y gris. De igual forma, se asignaron tres colores para diferenciar los tres tipos de interacción. El fondo sobre el que se dibujarían las redes debía procurar un buen contraste para diferenciar los tonos, así que se optó por un color oscuro.

El tamaño de los elementos también representa una dimensión de los datos. En el caso de los nodos, el área se calcula con base en el número de enlaces recibidos (*in-degree*) o bien de los emitidos (*out-degree*). De esta manera, se puede identificar a las cuentas más solicitadas o las más activas. En cuanto a los enlaces, el grosor de las líneas aumenta según el número de interacciones entre dos cuentas.

La principal característica de nuestras redes es la multiplicidad de formas de interacción entre las cuentas. Para observar los tres tipos de comunicación, cada uno se dibuja con un enlace y, de existir más de un tipo de vinculación entre dos cuentas, se añade otra línea y se modifica su curvatura para diferenciarla de la primera. Así, pueden haber hasta seis enlaces entre dos nodos A y B: tres de A hacia B y viceversa.

Se decidió trabajar con D3 –una librería de JavaScript para generar y manipular documentos web con datos–, debido a que soporta un amplio número de representaciones gráficas y brinda gran control sobre los atributos visuales y la interactividad. Para obtener grafos más legibles se modificaron los atributos predeterminados del algoritmo de fuerza de D3. Se redujo la gravedad y se aumentó la longitud de las aristas con el fin de dispersar el grafo y observar mejor las interacciones. En la Figura 6 se muestra una de las redes obtenidas.

3.5. Interactividad

Sintetizar en una imagen el conjunto de datos presentes en redes multivariantes es una tarea desafiante y, en ocasiones, resulta imposible mostrar todos los datos de manera útil [8]. En los diagramas nodo-enlace, surgen problemas de legibilidad a medida que el volumen de datos aumenta. La posibilidad de interactuar con la representación es un aspecto esencial para obtener información de la visualización. La inclusión de funciones como *zoom*, *panning*, resaltado, filtrado o búsqueda permite a los usuarios ubicar zonas y actores de interés [20].

En la presente propuesta de visualización, se incluyen las funciones de zoom y arrastre para acercar, alejar y desplazar el grafo. El usuario puede ocultar los enlaces para observar solamente la distribución de los nodos en el grafo. En caso de que necesite concentrarse sólo en los enlaces, puede quitar el color a los nodos.

Para ubicar rápidamente las cuentas principales, se colocó un botón que muestra los nombres de los diez nodos con mayor grado de entrada. Se incluyó un campo de búsqueda para localizar cuentas específicas. Al hacer clic en *Buscar*, se resaltan el nodo y sus vecinos. Asimismo, cuando se pasa el cursor sobre un nodo, se muestran su nombre, grado de entrada y grado de salida. Dar clic en el nodo selecciona su red inmediata y muestra la frecuencia de interacción de los enlaces cuyo peso es mayor a uno, porque a simple vista no es posible comparar

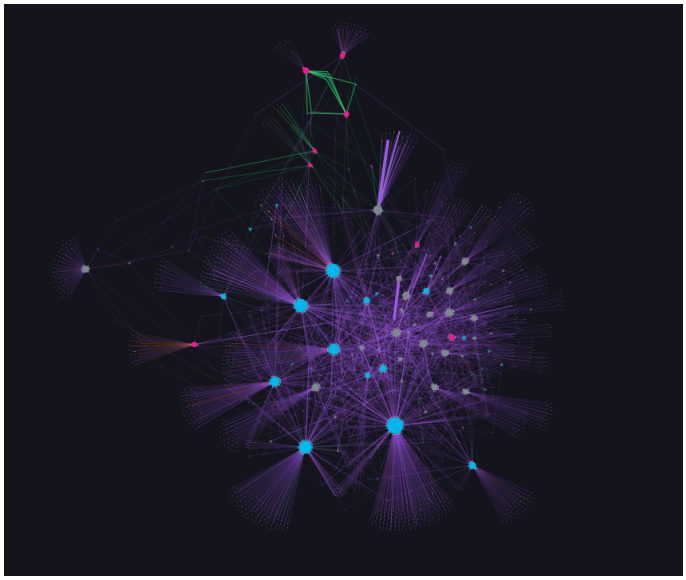

Fig. 6. Red de interacciones

con precisión su grosor. Un par de radio *buttons* permite alternar entre los valores de grado de entrada o salida para determinar el tamaño de los nodos.

En cuanto a los filtros, es posible observar cada uno de los tipos de interacción por separado o cualquier combinación de ellos. Otro conjunto de botones filtra los nodos según su clasificación. Además, oculta aquellos nodos relacionados únicamente con cuentas de la clasificación filtrada. Por último, los usuarios pueden observar sólo aquellas relaciones recíprocas entre nodos. En la Figura 7 se muestra una red y en la parte izquierda de la pantalla el menú que contiene las herramientas de filtrado.

4. Resultados

El método propuesto para la obtención y tratamiento de datos, permite una simplificación en el flujo de análisis de grandes cantidades de entradas. La recuperación de cadenas de conversaciones implicó la incorporación de datos relacionados con la temática recuperada que, de otra manera, sería complejo vincular. La depuración automatizada de publicaciones posibilitó la obtención de corpus condensados con una menor variación temática entre sus publicaciones. Al no contar con un corpus anotado ni con un *gold standard* para evaluar el desempeño con los datos actuales, se utilizó una validación cruzada con 10 pliegues

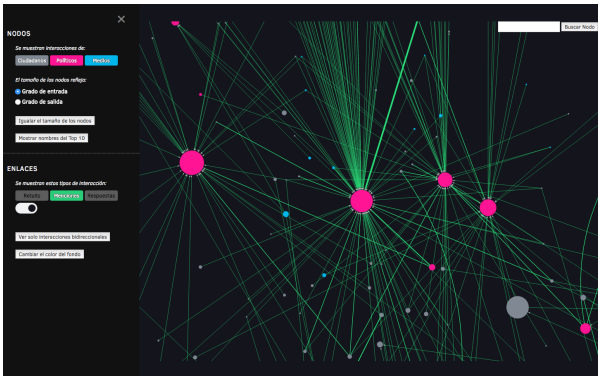

Fig. 7. Controles de la visualización que permiten navegar y filtrar el grafo

en Weka. Los resultados detallados de la evaluación y del clasificador de perfiles son presentados en las Tablas 1 y 2 respectivamente.

Tabla 1. Resultados de la validación cruzada con 10 pliegues para el modelo de clasificación

Indicador	**valor**	
Instancias clasificadas correctamente	2184	95.6%
Instancias clasificadas incorrectamente	99	4.3%
Coeficiente kappa	0.9127	
Error promedio absoluto	0.0537	
Error cuadrático Medio	0.1822	
Error absoluto relativo	10.82 %	
Número de instancias en la evaluación	2283	

Así, la clasificación automatizada permite obviar parte de la manipulación de la información recabada.

Pese a los resultados, existe un gran número de instancias incorrectamente categorizadas, debido a las formas de uso de Twitter, en específico la capacidad de los usuarios para ingresar información arbitraria a su perfil, dando lugar a inconsistencias y, por ende, disminuyendo la precisión del sistema de clasificación.

A partir del análisis de las redes generadas, se observa un uso principalmente informativo de Twitter, pues la forma de interacción más empleada es el retuit y se concentra alrededor de cuentas de medios de noticias. La actividad de los usuarios se enfoca en la difusión de notas periodísticas sobre los temas analizados. Un rasgo característico es que la mayoría de los usuarios están vinculados solamente a un medio, siendo muy pocos quienes retuitean contenido de varias fuentes. Las menciones son utilizadas generalmente para dirigir mensajes

Tabla 2. Resultados de la clasificación automática de usuarios

Clase	**Precisión**	**Cobertura**	**F-measure**
Político	0.967	0.952	0.960
Medio	0.944	0.962	0.953
Promedio ponderado	0.957	0.957	0.957

a ciertos actores o para hablar sobre ellos. Es por ello que existen regiones de color verde en los grafos donde, por lo regular, hay cuentas de políticos. La respuesta es el mecanismo de interacción menos utilizado, lo cual parece indicar poca propensión al diálogo en las redes que se estudiaron.

La presencia de cuentas de políticos y medios es muy escasa. En promedio, el 90% de los nodos fue clasificado como ciudadano. Se observó en los actores políticos y de medios un uso estratégico de Twitter. Por ejemplo, cuentas pertenecientes al mismo partido político retuitean el contenido publicado por algún líder de su organización. Los medios muestran un comportamiento similar, con periodistas difundiendo las notas del medio en el que laboran. Hay poca vinculación entre estas dos esferas de actores. Los portales digitales de noticias parecen estar más dispuestos a interactuar con sus seguidores o a mencionar figuras políticas en sus tuits.

En cuanto a las cuentas identificadas como ciudadanos, algunas pueden recibir el mismo número de interacciones que los medios de noticias. Este es un rasgo distintivo de la comunicación en medios sociales. Anteriormente, el acceso a los medios de comunicación estaba más restringido. Es notable que las cuentas de este tipo de actor hacen uso de todas las formas de interacción disponibles. En este sentido, la reconfiguración de la Comunicación Política está ocurriendo desde la ciudadanía. No obstante, es necesario un mayor involucramiento de los otros dos actores, pues, según lo observado en estas redes, los políticos y los medios repiten las estrategias que emplean en el mundo *offline*. Los actores políticos están en la plataforma para mostrar su presencia y ganar adeptos, mientras que los medios de noticias lo usan para difundir sus notas.

Los nodos con mayor grado de entrada –que suelen pertenecer a figuras políticas relevantes o a medios consolidados– rara vez establecen interacción con aquellos que los mencionan o que responden alguno de sus tuits. Por el contrario, políticos de menor jerarquía y periodistas muestran mayor disposición a intercambiar mensajes.

En este artículo se presenta una identificación de los actores que permite dar seguimiento al tipo de interacciones que tienen entre ellos. Sin embargo, la clasificación de actores puede ser un trabajo complejo, inicialmente manual, en el que el conocimiento previo del experto sobre los usuarios juega un papel fundamental. La automatización del proceso del clasificación no es siempre exacto debido a la limitada información en los perfiles de usuario. Aún así, el trabajo presentado es un acercamiento que puede ser complementado para la clasificación de los usuarios previo a un análisis manual.

5. Conclusiones y trabajo futuro

En este artículo se presentó una propuesta para la recuperación y procesamiento de tuits con el objetivo de visualizar estructuras de interacción entre medios de comunicación, ciudadanos y políticos. Si bien, el trabajo es muy importante en el área hay cuestiones que, por la gran cantidad de datos obtenidos, son difíciles de automatizar. Por ejemplo, la caracterización o perfilado de actores en contextos diversos. Para ello, es importante contar con un proceso manual previo que permita catalogar a los actores de acuerdo con sus funciones.

La clasificación también puede ser complementada con la consideración de otros parámetros, como el comportamiento de publicación, relación entre usuarios seguidos y suscriptores, detección afinada de *pseudo-bots*, así como la implementación de una función de categorización manual por parte del usuario, que, a su vez, sea incorporada en el entrenamiento del clasificador.

Estas afinaciones permtirían, a su vez, una mejor detección de *bots*, algo que se traduciría en un reducción mayor de la cantidad de entradas y un corpus refinado con respecto a la temática de búsqueda.

También se identificó la necesidad de diferenciar aún más la clasificación de las cuentas. Por ejemplo, los medios de noticias pueden ser subdividos en medios tradicionales, medios digitales y periodistas, ya que su comportamiento en Twitter suele ser distinto.

Por otra parte, es necesario complementar la red con gráficos estadísticos y listas con los actores principales. Otra característica necesaria es el funcionamiento dinámico de los filtros. En otras palabras, la red, y atributos como el grado de entrada y de salida, deben actualizarse con cada filtro aplicado.

Otra área que puede explorarse es el dibujo de redes dinámicas. La observación del despliegue de la red aportaría más datos para la comprensión del fenómeno.

El estudio de las interacciones entre usuarios de Twitter en términos de Comunicación Política debe continuar para incrementar el conocimiento sobre el impacto de la tecnología en los procesos comunicativos de los integrantes de una sociedad. Según lo observado en esta investigación, el análisis de redes es un buen punto de partida que, sin embargo, debe enriquecerse con otras aproximaciones como el análisis de sentimiento o de contenido.

Agradecimientos

Agradecemos el apoyo de la coordinación de la Maestría en Diseño, Información y Comunicación (UAM-C) y de la Red Temática en Tecnologías del Lenguaje del CONACyT, México.

Referencias

1. Bruns, A., Moe, H.: Structural layers of communication on twitter. In: Weller, K., Bruns, A., Burgess, J., Mahrt, M., Puschmann, C. (eds.) Twitter and Society, chap. 2, pp. 15–28. Peter Lang (2014)

2. Chadwick, A.: The Hybrid Media System: Politics and Power. Oxford University Press (2013)
3. Cossu, J.V., Abascal-Mena, R., Molina-Villegas, A., Torres-Moreno, J.M., San-Juan, E.: Bilingual and cross domain politics analysis. Avances en la Ingeniería del Lenguaje y del Conocimiento 85, 9–19 (2014)
4. Dubois, E., Gaffney, D.: The Multiple Facets of Influence: Identifying Political Influentials and Opinion Leaders on Twitter. American Behavioral Scientist 58(10), 1260–1277 (2014)
5. Graham, T., Jackson, D., Broersma, M.: New platform, old habits? candidates' use of twitter during the 2010 british and dutch general election campaigns. New Media & Society 18(5), 765–783 (2016)
6. Huang, H., Cao, Y., Huang, X., Ji, H., Lin, C.Y.: Collective tweet wikification based on semi-supervised graph regularization. ACL 1, 380–390 (2014)
7. Jungherr, A.: The logic of political coverage on twitter: Temporal dynamics and content. Journal of Communication 64(2), 239–259 (2014)
8. Kerren, A., Purchase, H.C., Ward, M.O.: Multivariate Network Visualization. Springer International Publishing, Cham (2014), http://link.springer.com/10.1007/978-3-319-06793-3
9. Kruikemeier, S.: How political candidates use twitter and the impact on votes. Computers in Human Behavior (34), 131–139 (2014)
10. Lahuerta-Otero, E., Cordero-Gutiérrez, R.: Looking for the perfect tweet. the use of data mining techniques to find influencers on twitter. Computers in Human Behavior 64, 575–583 (2016)
11. Makazhanov, A., Rafiei, D., Waqar, M.: Predicting political preference of twitter users. Social Network Analysis and Mining 4(1), 1–15 (2014)
12. McNair, B.: An introduction to political communication. Taylor & Francis (2011)
13. Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A.: Ranked wordnet graph for sentiment polarity classification in twitter. Social Network Analysis and Mining 28(1), 93–107 (2014)
14. Neves, A., Vieira, R., Mourão, F., Rocha, L.: Quantifying complementarity among strategies for influencers' detection on twitter. Procedia Computer Science 51, 2435–2444 (2015)
15. Newman, T.P.: Tracking the release of ipcc ar5 on twitter: Users, comments, and sources following the release of the working group i summary for policymakers. Public Understanding of Science (2016)
16. Pfeffer, J.: Fundamentals of visualizing communication networks. China Communications 10(3), 82–90 (2013), http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6488833
17. Smith, M., Rainie, L., Himelboim, I., Shneiderman, B.: Mapping Twitter Topic Networks: From Polarized Crowds to Community Clusters. The Pew Research Center (February 20), 1–57 (2014), http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters
18. Subbian, K., Aggarwal, C.C., Srivastava, J.: Querying and tracking influencers in social streams. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. pp. 493–502. ACM (2016)
19. Wang, J., Cong, G., Zhao, W.X., Li, X.: Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets. AAAI pp. 318–324 (2015)
20. Wybrow, M., Elmqvist, N., Fekete, J.D., von Landesberger, T., van Wijk, J.J., Zimmer, B.: Interaction in the visualization of multivariate networks. In: Kerren,

A., Purchase, H.C., Ward, M.O. (eds.) Multivariate Network Visualization, chap. 6, pp. 97–126. Springer International Publishing (2014)