

TWITTER, DEL SONDEO A LA SONDA: NUEVOS CANALES DE OPINIÓN, NUEVOS MÉTODOS DE ANÁLISIS



M. Luz Congosto

Universidad Carlos III, Departamento de Telemática, grupo Webtlab.
mariluz.congosto@gmail.com



Pablo Aragón

Fundació Barcelona Media.
pablo.aragon@barcelonamedia.org

Resumen

Twitter proporciona un caudal de opinión totalmente espontáneo y al alcance del que quiera recogerlo. El volumen de información supera con creces el volumen de cualquier sondeo y el muestreo es continuo. Sin embargo, la información es desestructurada, limitada a 140 caracteres y presenta numerosas dificultades a la hora de segmentar por edad, sexo, nivel de estudios o económico. Para obtener resultados concluyentes de esta nueva fuente de opinión es preciso aplicar nuevos métodos de análisis que permitan comprender la reacción de los ciudadanos tanto ante eventos señalados como en el día a día.

En este artículo se analiza el caso particular de las elecciones del 20N en España a fin de determinar nuevas vías de análisis de campañas electorales.

Palabras clave: Twitter, minería de datos, predicciones.

Abstract

Twitter provides a completely spontaneous flow of opinion ready to be picked up. The amount of information far exceeds the volume of any survey and the sampling is continuous. However, the information is unstructured, limited to 140 characters and presents numerous difficulties to be segmented by age, sex, educational or economic level. In order to take advantage of this new source of opinion, we need to apply new analytical methods to understand the reaction of citizens to specific events and day to day.

In this article we study the case of the Spanish General Elections on November 20 to determine new methods for analyzing political campaigns.

Keywords: Twitter, data mining, forecast.

I. INTRODUCCIÓN

El desarrollo de las redes sociales ha creado nuevas vías de comunicación y expresión de opinión de los ciudadanos. En cada proceso electoral va aumentando el interés de los gestores de las campañas en tener una presencia activa en las redes sociales, entre ellas Twitter. Actualmente casi todos los partidos y candidatos principales tienen un perfil creado en Twitter que gestionan en colaboración con su equipo, con mucha más intensidad en períodos electorales. Por otra parte, los ciudadanos pueden interactuar con ellos, hacerse eco de sus mensajes, seguir sus lemas, mostrar su desacuerdo o simplemente citarlos por su nombre real. Todo esto se realiza por medio de breves mensajes de 140 caracteres, llamados *tweets*, que imponen una concisión en el contenido.

Estas pequeñas piezas de información pueden contener además del texto una serie de componentes que enriquecen el contenido como etiquetas (*#hashtags*) para enfatizar el contenido del mensaje y hacer que se propague mejor por la Red, marcas que determinan que el mensaje no es original sino una retransmisión (RT, Via), menciones específicas a otros usuarios de Twitter (@usuario) y enlaces a sitios Web que amplían el contenido del mensaje. Todos estos componentes debidamente analizados en su conjunto pueden proporcionar información sobre qué tipo de actividad se está produciendo en las campañas electorales, qué lemas son los que más se están usando, qué mensajes se están difundiendo, qué candidatos o partidos están captando la atención de los ciudadanos, qué medios de comunicación o blogs están siendo más citados y qué tipo de relaciones existen entre los usuarios que están interactuando.

La aplicación de las tecnologías de minería para extraer valor del caudal continuo de millones de tweets generados por los usuarios en Twitter está abriendo nuevas líneas de investigación, entre ellas la predicción de resultados electorales. Actualmente existen dos corrientes de opinión científica sobre la capacidad de predecir resultados electorales analizando la participación ciudadana en Twitter. Por una parte, hay una opinión favorable, cuyo caso más destacado es el estudio realizado por Tumasjan (Tumasjan, 2010) y en el otro extremo una corriente crítica desde los trabajos de Panagiotis (Panagiotis, 2011), (Gayo-Avello (Gayo-Avello, 2011) y Jungherr (Jungherr, 2011). En el momento actual no existe un método validado y reproducido en distintos escenarios que permita predecir elecciones. No obstante, el análisis de elecciones

puede contribuir conocer mejor el entorno social en las campañas electorales y a aportar nueva luz hacia nuevos procesos de predicción de resultados.

A continuación se analizan algunos aspectos de la campaña de las elecciones generales del 20N en Twitter, como la participación de los ciudadanos, el estado emocional de los políticos, la popularidad de los candidatos, las relaciones entre ellos y la predicción de resultados electorales.

II. PARTICIPACIÓN DE LOS CIUDADANOS DURANTE LA CAMPAÑA

La participación ciudadana en Twitter ha estado marcada por los grandes eventos como se puede apreciar en la Figura 1 que muestra la evolución diaria del número de personas que participaron y de la cantidad de tweets que se publicaron. Destacan las siguientes fechas: el 15 de octubre (manifestación global de indignados), el 7 de noviembre (debate televisivo de Rubalcaba-Rajoy) y el 20 de noviembre (día de las elecciones).

En la fecha de mayor expectación, el 7 de noviembre, el número de ciudadanos que opinaron fue de casi 170.000, pudiéndose considerar como el tope de participación en temas políticos, lejos aún del techo alcanzado en eventos deportivos de 500.000 usuarios (Figura 1).

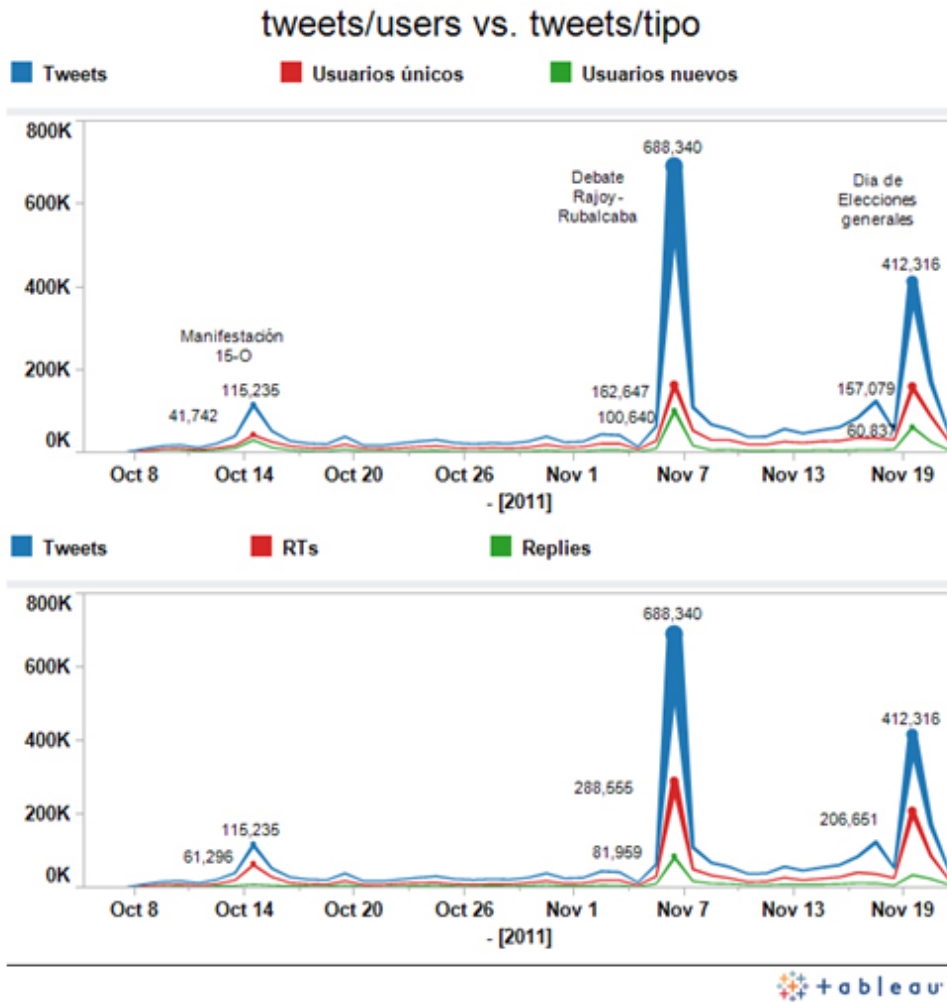
También se observa que durante la campaña tanto los usuarios como los tweets fueron el doble que en la precampaña, lo que denota un aumento de la atención de los ciudadanos hacia la política conforme se aproximaban las elecciones.

III. ANÁLISIS DE EMOTIVIDAD DE LOS POLÍTICOS

Para el análisis de emotividad se ha utilizado el corpus SPANNEW (Redondo, 2007) que mide los siguientes factores emocionales: **valencia** (valores altos corresponden a sentimientos de felicidad, satisfacción y esperanza y valores bajos a tristeza, descontento y desesperación), **excitación** (capta la asociación de palabras con sentimientos de entusiasmo, la ira, o el frenesí, y sus opuestos) y **dominancia** (se centra en sentimientos de dominación frente a los sentimientos de sumisión o temor).

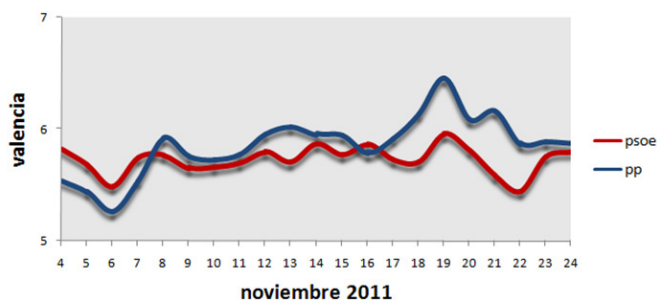
Para los políticos de los dos partidos principales, PP y PSOE, se observa un incremento en la **valencia** de

Figura 1. Participación general.



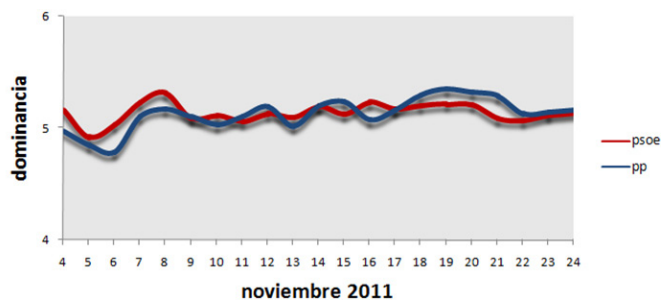
los tweets del PP a partir del día del debate electoral superando a los del PSOE durante toda la campaña, a excepción del 16 de noviembre. Los últimos días de campaña la valencia del PP crece notablemente respecto al PSOE, y tras los resultados, los valores de la valencia decrecen de manera significativa en el PSOE.

Figura 2. Análisis de emotividad (valencia).



Para los valores de **excitación** no se encontraron resultados significativos. Respecto a la **dominancia**, PP y PSOE se alternan como líderes en la campaña electoral. Al igual que en la valencia, tras el debate Rajoy-Rubalcaba los valores crecen notablemente en ambos partidos. Sin embargo en el final de la campaña y tras las elecciones, el PP supera de manera clara al PSOE en dominancia.

Figura 3. Análisis de emotividad (dominancia).



IV. POPULARIDAD DE LOS CANDIDATOS

La popularidad de los candidatos de los principales partidos se ha medido tanto en las menciones recibidas como en el número seguidores. Las menciones corresponden a una medida dinámica ya que pueden realizarse tantas veces como se quiera. Es por tanto una medida que fluctúa en el tiempo y que indica la atención captada por cada candidato en cada momento. En el caso de las elecciones del 20N @Con Rubalcaba fue el más mencionado aunque el día de debate y el de las elecciones fue sobrepasado por @MarianoRajoy como se puede apreciar en la Figura 4.

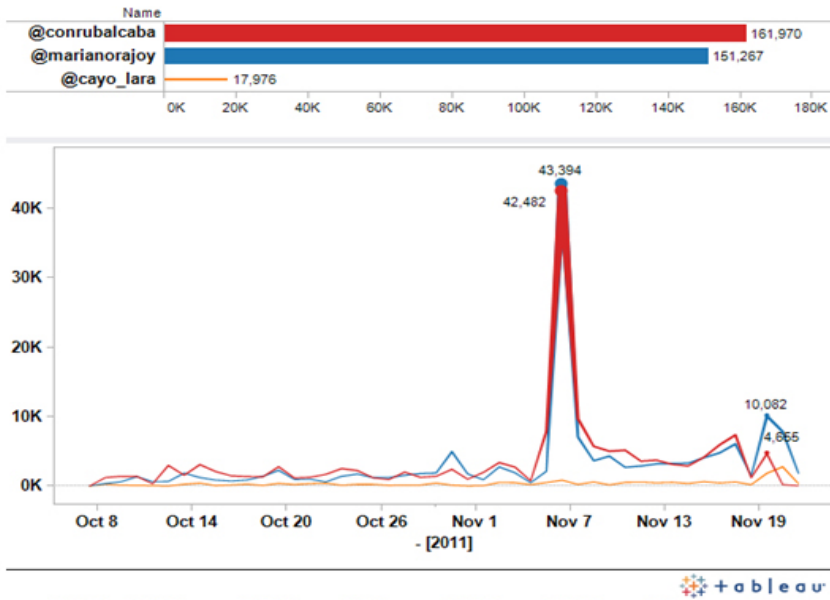


Figura 4. Menciones a los @candidatos.

El número de seguidores es una medida estática ya que la acción de seguir a un candidato se realiza una vez. Aunque este indicador por sí mismo no refleja la aceptación de un usuario en Twitter (Cha, 2010) no deja de ser interesante observar su evolución en el tiempo para descubrir qué factores han podido influir en su crecimiento. En la Figura 5 se muestra el acumulado de nuevos seguidores desde el inicio de la campaña. Para @marianorajoy el incremento de seguidores fue notablemente mayor en los días del debate y las elecciones. Para @ConRubalcaba este incremento solo ocurrió en el día del debate y para @cayo_lara el incremento fue uniforme.

De esta gráfica podría deducirse que los **acontecimientos especiales** relacionados con las elecciones estimulan a los usuarios a seguir a los candidatos protagonistas y que esa adhesión tiene alguna relación con el **éxito** del candidato (Figura 5).

V. RELACIONES ENTRE LOS POLÍTICOS

El RT es el mecanismo de difusión de mensajes en Twitter. Generalmente, se retransmiten los mensajes con los que se está de acuerdo, por este motivo en la red formada por los RTs emergen las comunidades de usuarios de ideología afín. Como indica Conover

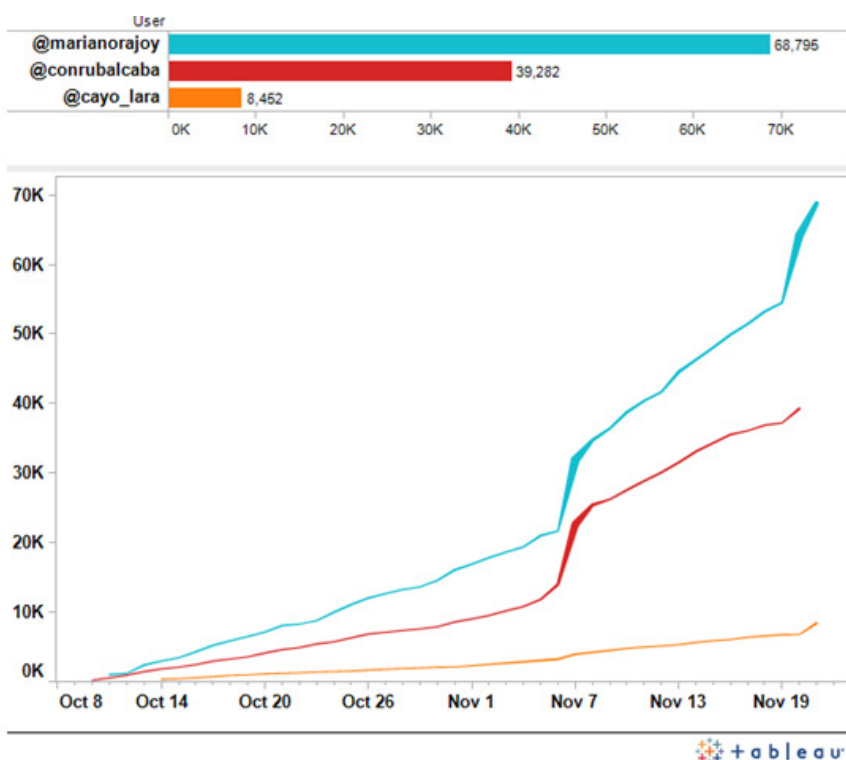


Figura 5. Nuevos seguidores de los candidatos.

(Conover, 2011) analizando la red de RTs se puede determinar en el 94.9% de los usuarios su polaridad política.

Analizando la forma en que los miembros de los partidos hacen RT queda de manifiesto la endogamia en estas comunicaciones ya que prácticamente la totalidad de mensajes retransmitidos se realizan dentro del propio partido.

	pp	psoe	iu	upyd	equo
pp	4186	0	0	3	0
psoe	8	4729	3	13	3
iu	1	3	964	2	2
upyd	3	13	40	7013	2
equo	0	4	6	3	960

Al generar el grafo con los RTs (Figura 6) entre miembros de un mismo partido se evidencia que EQUO, UPYD e IU (partidos nacionales con menor cobertura mediática) generan redes de difusión con un mayor coeficiente de clusterización respecto a los partidos con mayor atención de prensa tradicional nacional y local (PP y en mayor medida, PSOE).

Sin embargo, mientras en EQUO (82%) y UPYD (73%) la mayoría de los usuarios pertenecen a la componente gigante, en IU (44%) figuran menos de la mitad.

guración como plataforma de coalición de diferentes partidos.

V. MÉTODOS DE PREDICCIÓN

5.1. Medidas basadas en el cómputo de menciones

Siguiendo el método aplicado por (Tumasjan, 2010) se contabilizaron las menciones y se compararon con los resultados. Al no estar representados todos los partidos, los resultados electorales se normalizarán para que la suma del porcentaje de los votos de los cinco partidos analizados fuera el 100%.

El resultado de este método aplicado a las elecciones 20N se ha obtenido un MAE (Mean Absolute Error) de 1.66%. (Tabla 2).

5.2. Medidas basadas en la polaridad política

Como alternativa al cómputo por mención se evalúa la opción de medir la polaridad política en los usuarios. Esta medida está basada en las conclusiones de varios trabajos de investigación. Por un lado se apoya en el trabajo de Conover (Conover, 2011) que determina que tanto el RT como el uso de *#hashtags* permiten determinar la polaridad política de los usuarios. También en los estudios de Golbeck (Golbeck Jen, 2010) que midió la polaridad de los usuarios en función de la polaridad de los políticos a los que seguían. Finalmente, para la realización de filtros ha sido de mucha

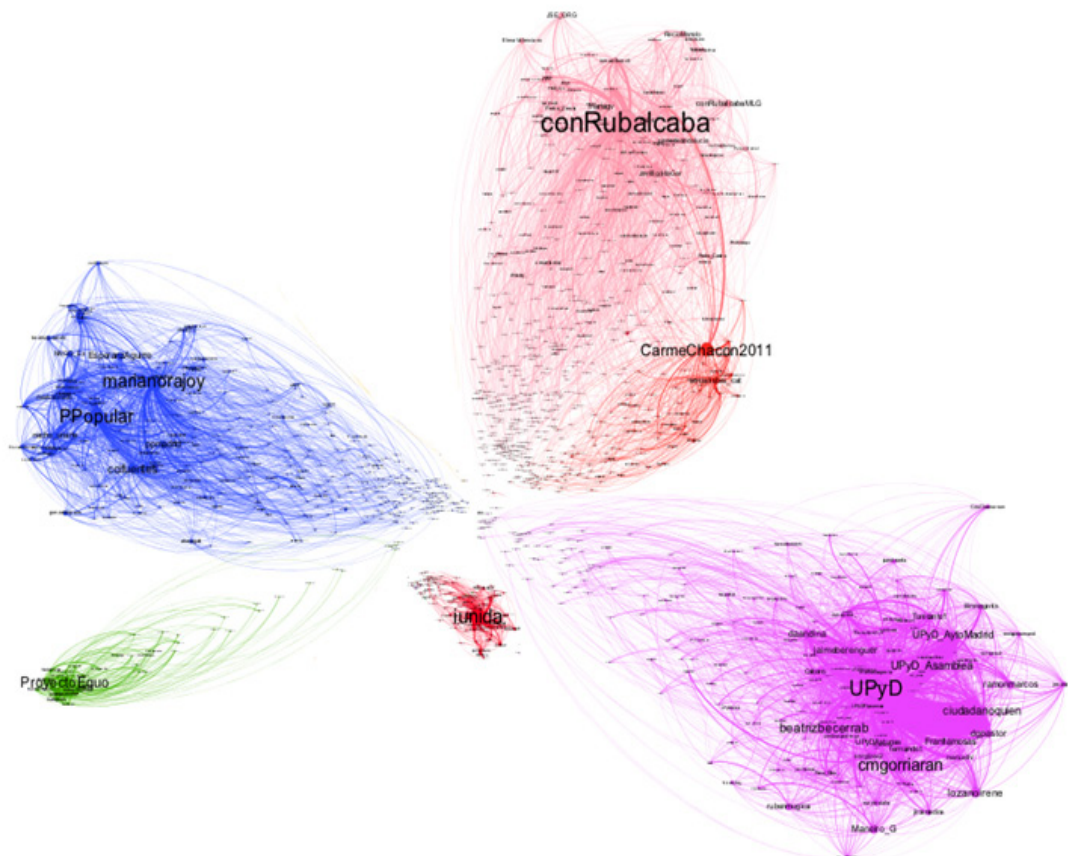


Figura 6. Grafo de RTs

Tabla 2. Menciones.

Partido	(v) %Votos	(vn) %Votos normalizados	Menciones	(vt) %Votos Twitter	vn-vt
PP	44,62%	51,8%	1.142.196	53,72%	+1,73%
PSOE	28,73%	33,47%	746.317	35,10%	+1,63%
IU	6,92%	8,06%	83.064	3,91%	-4,16%
UPyD	4,69%	5,46%	126.953	5,97%	+0,51%
Equo	0,88%	1,3%	27.866	1,31%	+0,29%

Tabla 3. Usuarios en general.

Partido	(v) %Votos	(vn) %Votos normalizados	Usuarios	(vt) %Votos Twitter	vn-vt
PP	44,62%	51,98%	11.610	50,25%	-1,73%
PSOE	28,73%	33,47%	5.539	23,97%	-9,50%
IU	6,92%	8,06%	1.568	6,79%	-1,28%
UPyD	4,69%	5,46%	2.245	9,72%	+4,25%
Equo	0,88%	1,03%	2.143	9,28%	+8,25%

utilidad aplicar las propuesta de Mislove (Mislove, 2011) para identificar personas y su género.

La polaridad de los usuarios se calcula en función de los RTs realizados a usuarios Twitter que tienen afiliación política conocida y por el uso de los *#hashtags* de la campaña promovidos por los partidos.

Se detectó polaridad política en el **26.81%** de los tweets y al **47.07%** de los usuarios. Se realizó una prueba para validar el algoritmo con la lista de políticos/partidos para determinar si se les clasificaba correctamente y se detectó polaridad política en el **86.42%** de ellos y en los detectados se obtuvo **99%** de aciertos. Del conjunto de usuarios se seleccionó a los que tenían una ligera polaridad política (al menos 3RTs/*#hashtags*) y se calculó para los siguientes escenarios:

- » Usuarios en general MAE: 5.00% (Tabla 3)
- » Usuarios identificados como hombres MAE: 6.49% (Tabla 4)
- » Usuarios identificados como mujeres MAE: 3.88% (Tabla 5)

El mejor resultado (3,88%) corresponde al grupo identificado como **mujeres** y el peor (6,49%) al grupo reconocido como **hombres**.

VI. CONCLUSIONES

Con el método “**Medidas basadas en el cómputo de menciones**” se han conseguido mejores resultados que con las “**Medidas basadas en la polaridad política**”. En ambos procesos de evaluación se ha observado que

tanto los partidos como sus seguidores hacen un uso diferente de Twitter. Algunos partidos utilizan preferentemente el RT para difundir mensajes mientras que otros tienden a emplear los *#hashtags*. Al medir todos los tipos de menciones los distintos usos se equilibran y se obtienen mejores resultados.

Con el método “**Medidas basadas en la polaridad política**”, se han obtenido unos resultados menos precisos debido a las siguientes barreras que habrá que tener en cuenta para mejorar los algoritmos de predicción:

- » **Demografía:** La muestra de usuarios no es completa, no están representados todos los segmentos de la sociedad. Los usuarios de Twitter son jóvenes y en un alto porcentaje con estudios superiores.
- » **Opinión oculta:** No todos los usuarios manifiestan sus preferencias políticas. Por un lado los usuarios Twitter durante la campaña han difundido mayoritariamente informaciones de prensa progresista por lo que se supone que hay un sesgo de opinión favorable hacia la izquierda. Sin embargo, el porcentaje de polaridad de los usuarios hacia el PSOE ha estado muy por debajo de los votos obtenidos.
- » **Sobreopinión:** En contraposición al punto anterior, los partidos más pequeños (UPyD y Equo) al tener el acceso más difícil a la TV, la radio y la prensa que los partidos grandes, basan su campaña en las redes sociales, entre ellas Twitter. Sus simpatizantes participan activamente haciendo que el porcentaje de polaridad de estos partidos esté muy por encima de los votos obtenidos.
- » **Entidades vs. Personas:** Es difícil distinguir en

Tabla 4. Usuarios identificados como hombres.

Partido	(v) %Votos	(vn) %Votos normalizados	Usuarios	(vt) %Votos Twitter	vn-vt
PP	44,62%	51,98%	4.431	46,21%	-5,77%
PSOE	28,73%	33,47%	2.270	23,68%	-9,79%
IU	6,92%	8,06%	709	7,39%	-0,67%
UPyD	4,69%	5,46%	1.220	12,72%	+7,26%
Equo	0,88%	1,03%	958	9,99%	+8,97%

Tabla 5. Usuarios identificados como mujeres.

Partido	(v) %Votos	(vn) %Votos normalizados	Usuarios	(vt) %Votos Twitter	vn-vt
PP	44,62%	51,98%	2.937	49,94%	-2,04%
PSOE	28,73%	33,47%	1.652	28,09%	-5,38%
IU	6,92%	8,06%	340	5,78%	-2,28%
UPyD	4,69%	5,46%	457	7,77%	+2,31%
Equo	0,88%	1,03%	495	8,42%	+7,39%

Twitter una entidad de una persona debido a que es opcional cumplimentar el nombre en el perfil de Twitter. En la muestra analizada solo se ha identificado el **63,46%** como personas. Esto dificulta conocer si un usuario Twitter es un posible votante.

- » **Anonimato vs. "Identidad real"**: Los resultados del grupo de usuarios genéricos han sido mejores que los del grupo identificado como personas. Los usuarios con identidad real son más propensos a la opinión oculta.
- » **Hombres vs. Mujeres**: Se ha detectado diferencia de comportamiento según **género**. Los hombres son más propensos a ocultar su opinión o a sobreopinar que las mujeres.

A pesar de los buenos resultados de las "**Medidas basadas en el cómputo de menciones**" no se puede generalizar este proceso porque depende de muchos factores como el entorno socio-cultural de las elecciones, el periodo de la muestra, los eventos de campaña, la recogida de datos en Twitter, los partidos analizados y la forma de cómputo.

La validación de este método para predicción requeriría sistematizar todos los pasos y comprobar si los resultados son similares para otras elecciones en diferentes entornos.

REFERENCIAS BIBLIOGRÁFICAS

- Cha, M. (2010). *Measuring User Influence in Twitter: The Million Follower Fallacy*.
- Conover, D. (2011). *Predicting the Political Alignment of Twitter Users*.
- Gayo-Avello, D. (2011). *A warning against converting Twitter into the next Literary Digest*.
- Golbeck Jen, D. L. (2010). *Computing Political Preference among Twitter Followers*.
- Jungherr, e. a. (2011). *Why the Pirate Party won the german election of 2009*.
- Mislove, A. L. (2011). *Understanding the Demographics of Twitter Users*.
- Panagiotis, e. a. (2011). *How (Not) To Predict Elections*.
- Redondo, e. a. (2007). "The Spanish adaptation of ANEW (affective norms for English words)". *Behavior Research Methods*, 39(3), 600-605. Psychonomic Society Publications.
- Tumasjan, e. a. (2010). *Predicting Elections with Twitter: What 140 characters reveal about political sentiment*.