

# MINERÍA DE OPINIONES EN ELECCIONES PRESIDENCIALES

LA OPINIÓN PÚBLICA ES EL RESULTADO DE LA COMPLEJA INTERACCIÓN ENTRE PERSONAS QUE TIENEN DISTINTAS CREENCIAS Y PUNTOS DE VISTA ACERCA DE LA REALIDAD. ES LA AGREGACIÓN DE PUNTOS DE VISTA PARTICULARES SOBRE TEMAS ESPECÍFICOS QUE AFECTAN NUESTRAS VIDAS.



## MARCELO MENDOZA

Instructor académico de la Universidad Técnica Federico Santa María (UTFSM), desde 2011. Doctor en Ciencias mención Computación, Universidad de Chile (2007); Magister en Ciencias de la Ingeniería Informática, UTFSM (2001) e Ingeniero Civil Electrónico, UTFSM (2001). Áreas de investigación: Minería de Texto, Recuperación de Información, Análisis de Redes Sociales en Línea.

[mmendoza@inf.utfsm.cl](mailto:mmendoza@inf.utfsm.cl)

La opinión pública puede ser estudiada usando encuestas. Una amplia discusión acerca de la efectividad de las encuestas ha sido abordada en el estado del arte. Una de las principales falencias que las encuestas revelan es que el diseño del instrumento sesga la opinión del encuestado.

La opinión pública puede ser estudiada *in vivo*, sin necesidad de someter a las personas a encuestas. El surgimiento de la Web Social permite observar lo que la gente opina con respecto a marcas, celebridades y temas de interés público en general. La Computación ha tenido un papel relevante en este tema. Una amplia gama de métodos de minería de opiniones, análisis de reacciones emocionales frente a determinados eventos, así como la consolidación de recursos léxicos para estos fines han perfilado a la minería de opiniones como uno de los grandes desafíos para la comunidad de *text mining* en la última década [Pang07].

En este artículo vamos a abordar el problema específico de minería de opiniones en elecciones políticas. Comenzaremos relatando los primeros hallazgos de la minería de opiniones en este dominio, para luego mostrar nuestros resultados al analizar las elecciones de Estados Unidos de 2008. Luego, abordaremos las elecciones presidenciales de Chile, analizando los resultados entregados por nuestro observatorio político de Twitter recientemente liberado para monitorear el proceso electoral 2013. Finalmente, discutiremos resultados recientes y perspectivas para trabajo futuro.

## MINERÍA DE OPINIONES: EL ENTUSIASMO INICIAL

La posibilidad de recolectar opiniones desde las redes sociales y poder procesarlas para construir estudios de opinión pública, impulsó un sinnúmero de estudios con resultados asombrosos. Entre los resultados más resonantes destacan Asur & Huberman [Asur10] quienes usaban tweets para predecir el éxito de ventas de los próximos estrenos del cine; Bollen *et al.* [Bollen11] que encontraron correlación entre la emocionalidad de los usuarios de Twitter y las variaciones de índices bursátiles; O'Connor *et al.* [O'Connor10] quienes encontraron correlación entre encuestas de aprobación presidencial y polaridad de tweets; o bien Tumasjan *et al.* [Tumasjan10] quienes encontraron correlación entre el número de menciones en tweets de los candidatos a elecciones federales de 2009 en Alemania y los resultados de las elecciones. Estos trabajos hacían pensar que era posible realizar estudios de opinión pública directamente desde las redes sociales, en particular desde Twitter, con resultados precisos y representativos.

---

## PRIMERAS SEÑALES DE ALERTA

---

Daniel Gayo-Avello [Gayo11] reveló diferencias significativas entre las muestras de votantes que definieron las elecciones de Estados Unidos de 2008, y quienes emitían opiniones acerca de las elecciones en Twitter. Las principales diferencias detectadas mostraban sesgo demográfico, evidenciando que los usuarios de Twitter tienden a ser jóvenes o adultos jóvenes que habitan áreas urbanas de gran extensión. También fue posible detectar que una importante fracción de los votantes no emitía opiniones en forma pública, lo que conocemos como votante silencioso y, más aún, muchos de ellos decidían su voto poco antes de las elecciones.

González-Bailón *et al.* [González12] estudiaron foros de opinión de Usenet por un período que comprendió varios eventos políticos de impacto, como por ejemplo el 11/S y las elecciones presidenciales de Estados Unidos de 2000 y 2004. Usando métodos de minería de opiniones encontraron que la polaridad con respecto a la aprobación presidencial sufría cambios abruptos que correlacionaban con las reacciones emocionales ante eventos de shock. Ellos encontraron evidencia en los foros de Usenet que permitía afirmar que el 11/S había generado un *topic shift* que favoreció la aprobación del presidente Bush.

Podemos concluir del trabajo de Gayo-Avello [Gayo11] que existen importantes diferencias demográficas que deben ser consideradas para estudiar la opinión pública a partir de opiniones recolectadas desde redes sociales. Más aún, González-Bailón *et al.* [González12] muestran que las opiniones recolectadas desde foros de la Web usando métodos de minería de opiniones correlacionan con *topic shifts*, los que a su vez correlacionan con encuestas de opinión pública acerca de aprobación presidencial. Esto indica dos hechos: 1) **Los métodos de minería de opiniones miden algo, ese algo puede estar ses-**

**gado, puede ser poco representativo, pero refleja cambios y tendencias en opinión.** 2) **La opinión pública es permeable a eventos de alto impacto.** Luego, es de interés caracterizar el escenario político de una elección en términos del riesgo, para cuantificar cuán permeable está la opinión pública ante cambios en la escena política.

---

## ESTADOS UNIDOS 2008

---

Decidimos abordar el caso de las elecciones presidenciales de Estados Unidos de 2008 desde una perspectiva distinta, proveniente de la siguiente analogía: la aprobación o rechazo de un candidato presidencial sigue un proceso análogo al de la fluctuación de índices bursátiles. Para estudiar la opinión pública es necesario caracterizar sus tendencias al alza o a la baja y, sobre todo, cuantificar el riesgo, es decir, la volatilidad a la que están sujetas las intenciones de voto. En este contexto, decidimos caracterizar la variabilidad de las series de tiempo de opinión. La **Figura 1** muestra los datos sobre los que trabajamos.

La **Figura 1** muestra una alta variabilidad durante los primeros meses de campaña, donde ocasionalmente la serie McCain–Palin superó a Obama–Biden. Después del tercer mes de campaña, luego de las nominaciones oficiales de los candidatos, Obama–Biden siempre superó a McCain–Palin. Podemos observar también que las menciones a Obama–Biden aumentan significativamente el día de la elección, con motivo de su triunfo. Esto correlaciona con la aprobación de Obama–Biden y el rechazo a McCain–Palin.

En [Bravo-Márquez12] estudiamos la variabilidad de ésta y otras series de opinión, caracterizando el riesgo por presencia de GARCH, de acuerdo con la metodología propuesta por el economista Robert Engle [Engle82]. Uno de los hallazgos del análisis es que la serie de Obama–Biden era menos volátil que la serie de

McCain–Palin y que ambas tenían alta volatilidad durante los primeros meses de campaña. Por tanto, en este escenario, el problema de proyección se encontraba mal condicionado en los primeros meses y mostraba buenas propiedades predictivas sólo para Obama–Biden en los últimos dos meses de campaña. Dicho en otras palabras, cualquier método predictivo del estado del arte habría fallado en sus proyecciones sobre Twitter durante los primeros meses de campaña, y en el caso de McCain–Palin, no podría haber realizado proyecciones confiables nunca.

---

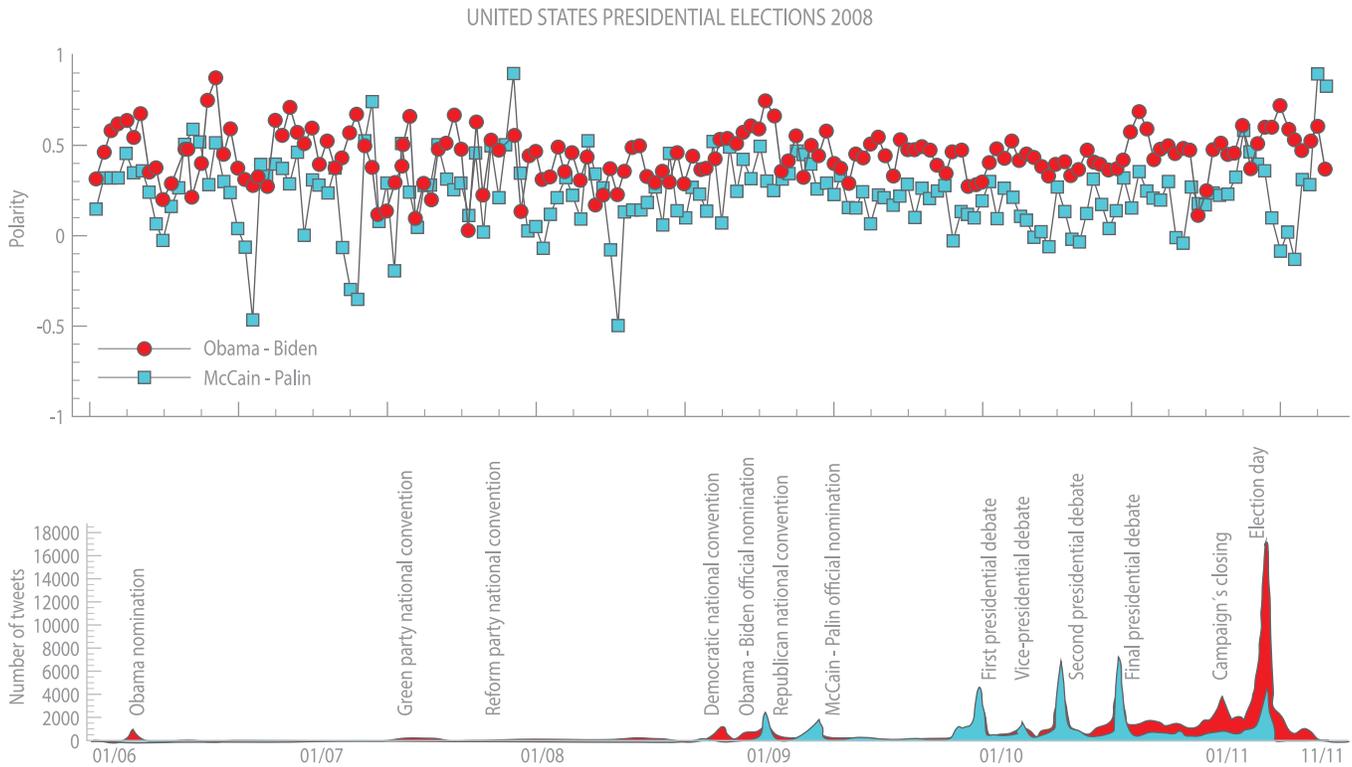
## OBSERVATORIO POLÍTICO EN CHILE

---

El escenario electoral de 2013 es sumamente interesante en Chile, lo cual nos motivó a desarrollar una herramienta que permitiera monitorear las elecciones presidenciales en Twitter. Esta plataforma web se encuentra disponible en <http://www.observatoriopolitico.cl>. La **Figura 2** nos muestra su *front*.

Observatorio Político [Marín13] recupera los tweets que mencionan a los candidatos y estima su polaridad usando un método de minería de opiniones acondicionado para el español. Haciendo agregación diaria por candidato, se construyen series de tiempo de menciones, aprobación (polaridad [Gayo-Avello11, Bravo-Márquez12]) y alcance, ésta última considerando el número de retweets para cada candidato. Las series de la **Figura 3** muestran la actividad para el mes de junio.

La **Figura 3** nos muestra la actividad del observatorio durante el mes de junio, mes que finaliza con las primarias. Podemos observar que tanto el número de menciones como el alcance aumentan a medida que se acerca el 30 de junio. El índice de aprobación muestra una alta variabilidad, al igual que lo observado en las presidenciales de Estados Unidos 2008, indicando que la campaña se encuentra en una primera fase



**FIGURA 1.** SERIE DE TIEMPO DIARIA DE OPINIÓN SOBRE CANDIDATOS A LAS PRESIDENCIALES DE ESTADOS UNIDOS DE 2008. CADA PUNTO INDICA LA DIFERENCIA ENTRE APROBACIÓN Y RECHAZO ENTRE LOS TWEETS QUE HAN MENCIONADO AL CANDIDATO. LA FIGURA DEL PIE MUESTRA MENCIONES TOTALES A LOS CANDIDATOS.

con un escenario abierto para la mayoría de los candidatos. Las series permiten distinguir dos grupos de candidatos, los que presentan mayor actividad tanto en menciones como en alcance (Bachelet, Longueira, Enríquez, Claude y Parisi) y otros candidatos con muy poca presencia en Twitter (Sfeir, Ruz, Miranda y Jocelyn-Holt). Nuestro Observatorio muestra también que las series no correlacionan con el tamaño de las comunidades de seguidores de cada candidato. Parisi y Longueira registran mucha actividad a pesar de tener comunidades de seguidores relativamente pequeñas en comparación con Enríquez, quien mantiene desde la elección 2009 más de 338.000 seguidores. A pesar de ello, los índices de actividad de Enríquez, tanto en menciones como en alcance, son comparables a los de sus contendores.

## ELECCIONES PRESIDENCIALES

17 de Noviembre del 2013

Primarias **Presidenciales**

**Acerca de...**

Observatorio Político es una plataforma Web enfocada en el monitoreo y análisis de la opinión pública acerca de candidatos a cargos políticos en Chile, usando como fuente de información la red social Twitter. Esta herramienta permite analizar los mensajes en dicha red social mediante técnicas tales como clustering, análisis de palabras, análisis de sentimientos en texto y otras técnicas computacionales de análisis de opinión.

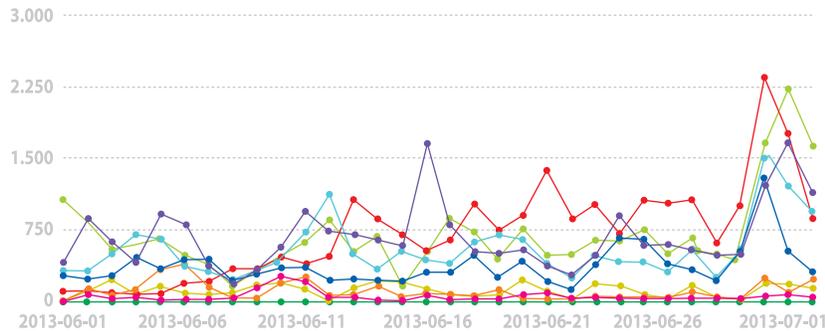
**LO ÚLTIMO @PoliticObs**

“ @PoliticObs mañana conversaremos con @puohringer acerca del rol de las #politicospublicas en la salud mental <http://l.co/?VVWbHmAsGQ> ”

© POLITICAS PUBLICAS

**FIGURA 2.** OBSERVATORIO POLÍTICO PERMITE MONITOREAR LAS ELECCIONES PRESIDENCIALES 2013 DE CHILE EN TWITTER.

## MENCIONES

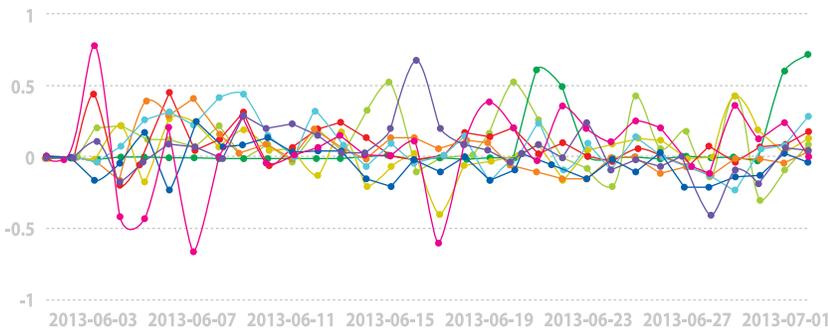


¿Qué es?

El gráfico muestra, para cada político, el número de tweets que ha hecho mención de alguna de sus cuentas a lo largo del tiempo.

EJEMPLO: "@presi2014 es mi candidato".

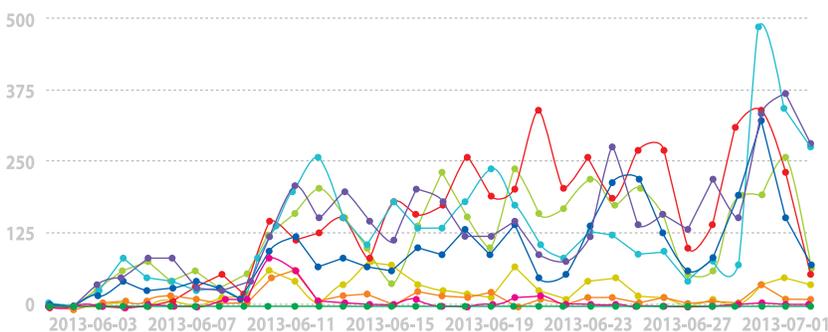
## APROBACIÓN



¿Qué es?

Este índice indica la diferencia entre aprobación y rechazo entre las personas que han mencionado al candidato. Si el número es positivo, significa que el candidato tiene más aprobación que rechazo. Si el número es negativo, significa que el candidato tiene más rechazo que aprobación. Mientras más cercano a 1 o -1 mayor es la diferencia.

## ALCANCE



¿Qué es?

El gráfico muestra el número de retweets que han tenido los mensajes de cada político a lo largo del tiempo.

MICHELLE BACHELET JERIA  
ALFREDO SFEIR YOUNIS  
GUSTAVO RUZ ZAÑARTU

ROXANA MIRANDA  
TOMÁS JOCELYN HOLT  
PABLO LONGUEIRA

MARCEL CLAUDE  
MARCO ENRÍQUEZ OMINAMI  
FRANCO ALDO PARISI

FIGURA 3.

OBSERVATORIO POLÍTICO PERMITE MONITOREAR LAS ELECCIONES PRESIDENCIALES 2013 EN TWITTER, CARACTERIZANDO EL ESCENARIO POLÍTICO EN EL TIEMPO.

## TOPIC SHIFTS

Recientemente hemos abordado el problema de detección de cambios en series de tiempo para proyección de mediano y largo plazo. Uno de los métodos explorados consiste en segmentar la serie de tiempo de acuerdo con una noción de coherencia de modelos, esto es, buscar el punto en el tiempo alrededor del cual el ajuste de modelos distintos maximiza un criterio de optimalidad. En [Mendoza13] mostramos que la segmentación de la serie de tiempo acorde a un enfoque divisivo permite distinguir entre períodos convulsionados y períodos calmos, y que la transición entre ambos indica un *topic shift*. La Figura 4 muestra los resultados de este método sobre la serie Obama-Biden 2008.

## DESAFÍOS

Finalizamos este artículo enumerando los principales desafíos de este problema.

### REPRESENTATIVIDAD

Los usuarios de las redes sociales se concentran en segmentos específicos de la población. Si bien no existe evidencia de correlación entre estos segmentos y una orientación política específica, es necesario corregir el sesgo que éste pueda introducir en el análisis.

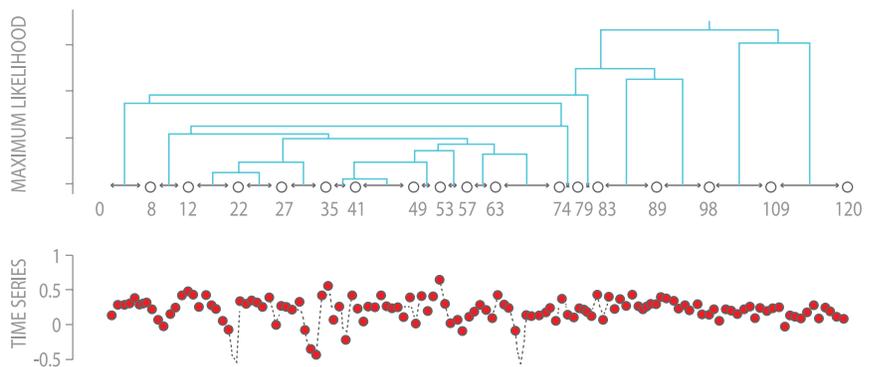
### OPINIÓN AMBIGUA

Existen situaciones en las cuales es difícil estimar la polaridad de un mensaje con respecto a una entidad nombrada. Los métodos de minería de opiniones son poco eficaces tratando con la ironía u otras expresiones de ambigüedad. Es necesario desarrollar métodos de minería de opiniones robustos a la ambigüedad.

## INTENCIÓN DE VOTO

El voto voluntario configura un escenario complejo para el análisis. No sólo es importante inferir la orientación política sino también la disposición a sufragar.

Finalmente algunas palabras acerca de aspectos éticos. Los límites de la minería de opiniones se encuentran donde comienzan los derechos de privacidad de datos de cada individuo. Los análisis deben realizarse a nivel agregado. Lo contrario hará que los usuarios sientan reticencia a expresar sus opiniones. Además, la información generada debe aportar valor a la sociedad en su conjunto. Entendemos que el desarrollo de plataformas de libre acceso a los datos es un camino valorable en este sentido. Lo contrario, la transacción de datos públicos en ambientes que buscan redituarse a intereses específicos nos acerca a un "Orweliano" *Big brother is watching you*, del cual no queremos ser parte. ■



**FIGURA 4.**

DETECCIÓN DE *TOPIC SHIFTS* USANDO COHERENCIA DE MODELOS. SE OBSERVA UN *TOPIC SHIFT* ALREDEDOR DEL DÍA 79 DE LA SERIE OBAMA-BIDEN 2008.

## REFERENCIAS

[Engle82] Engle, R. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation, *Econometrica*, 50(4):987-1007.

[Pang07] Pang, B. and Lee, L. (2007), Opinion mining and sentiment analysis, *Foundations and trends in information retrieval*, 2(1-2):1-135.

[Asur10] Asur, S. and Huberman, A. (2010), Predicting the future with social media, In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, Canadá, Aug. 2010.

[O'Connor10] O'Connor, B., Balasubramanyan, R., Routledge, B. and Smith, N. (2010), From tweets to polls: linking text sentiment to public opinion time series, In *Proceedings of the 4th International Conference on Weblogs and Social Media*, Washington, DC, USA, May 2010.

[Tumasjan10] Tumasjan, A., Sprenger, T., Sandner, Ph., and Welpe, M. (2010), Predicting elections with Twitter: what 140 characters reveal about political sentiment, In *Proceedings of the 4th International Conference on Weblogs and Social Media*, Washington, DC, USA, May 2010.

[Bollen11] Bollen, J., Mao, H. and Zeng, X. (2011), Twitter mood predicts the stock market, *Journal of Computational Science*, 2(1):1-8.

[Gayo11] Gayo-Avello, D. (2011), Don't turn social media into another 'Literary Digest' poll, *Communications of the ACM*, 54(10):121-128.

[González12] González-Bailón, S., Banchs, R. and Kaltenbrunner, A. (2012), Emotions, public opinion, and U.S. Presidential approval rates: a 5-year analysis of online political discussions, *Human Communication Research*, 38(2):121-143.

[Bravo-Márquez12] Bravo-Márquez, F., Gayo-Avello, D., Mendoza, M. and Poblete, B. (2012), Opinion dynamics of elections in Twitter, In *Proceedings of the 8th Latin American Web Congress*, Cartagena de Indias, Colombia, Oct. 2012.

[Mendoza13] Mendoza, M., Bravo-Márquez, F., Poblete, B. and Gayo-Avello, D. (2013), Long-memory time series ensembles for concept shift detection, In *Proceedings of the 2nd International workshop on Big Data, Streams and Heterogeneous Source Mining*, KDD workshop, Chicago, IL, USA, Aug. 2013.

[Marín13] Marín, M., Mendoza, M., Rodríguez, A., Contreras, A., Caro, D., Cabrera, L. and Sanhueza, C. (2013) Observatorio político, <http://www.observatoriopolitico.cl>