#140caracteres al sillón presidencial opinion mining para la predicción de resultados en las elecciones presidenciales Chile 2017

Gerardo Lecaros, Claudia Martínez-Araneda, José Abreu Salas
Computer Science Department
Universidad Católica de la Santísima Concepción
Concepción, Chile
gglecaros@ing.ucsc.cl, cmartinez@ucsc.cl, joseabreu@ucsc.cl

Alejandra Segura Navarrete, Christian Vidal-Castro, Clemente Rubio-Manzano
Computer Science Department
Universidad del BíoBio
Concepción, Chile
asegura@ubiobio.cl, cvidal@ubiobio.cl, clrubio@ubiobio.cl

Abstract*— La predicción de los resultados en una elección parlamentaria o presidencial puede ser determinante a la hora de tomar ciertas acciones correctivas tanto paras las coaliciones políticas como para los votantes. Entre los instrumentos clásicos de predicción en una elección están las encuestas o sondeos de opinión, los que son considerados medidas estadísticas que se aplican para determinar la intención de voto de los electores. La historia reciente habla de tasas de error no despreciables que se atribuyen entre otras causas a voto voluntario, voto oculto, dificultades en sondeos vía teléfono fijo, zonas rurales alejadas, alta tasa de no-respuesta en celulares, indecisos, censo no actualizado, entre otros. A nivel mundial, los pronósticos que daban por perdedor a Trump en EEUU, o que Reino Unido se mantendría en Brexit, fueron emblemáticos. El caso más reciente ocurrió en Chile, en cuyas últimas elecciones presidenciales de 2017 las tres encuestas más reconocidas tuvieron un error de 9,1 puntos entre la votación escrutada de la candidata Beatriz Sánchez y lo predicho. El objetivo de este trabajo fue determinar si los tweets resultaban ser buenos predictores de la intención de voto y si eran una herramienta efectiva para definir un perfil basado en atributos para un candidato presidencial. Se seleccionaron 66217 tweets para ocho candidatos en el periodo de mayo a octubre de 2017. Por otra parte, se aplicaron técnicas de sentiment analysis para determinar las polaridades de los tweets en el tiempo y mediante el método de co-ocurrencia y análisis descriptivo de palabras detectar qué atributos están más relacionados con un candidato. Los resultados muestran una baja tasa de aparición de los atributos del perfil en los microtextos con un error de predicción de 7,8 puntos porcentuales, valor comparable con el obtenido en las encuestas de opinión en Chile.

Keywords-component; Opinion mining, co-ocurrence words, text mining.

I. INTRODUCCION

Desde la aparición de Twitter el año 2006, éste ha llegado a ser uno de los servicios de microblogging más utilizados en internet considerando que en el año 2016 tenía 310 millones usuarios activos al mes enviando en promedio 6000 tweets por segundo. Un dato que resulta muy interesante es que el 83% de los líderes mundiales tiene una cuenta de Twitter. Según un estudio de Consumer Lab de Ericksson (2016), Chile se encuentra entre los países latinoamericanos más hiperconectados (33%) superando a Brasil (28%) y a Colombia (23%). A partir de este fenómeno aparecen conceptos como netizens y networkers, los primeros usuarios de la red que usan al menos siete servicios digitales al día y los segundos al menos tres. Ambos segmentos equivalen al 62% de los usuarios de internet en Chile, y en conjunto impulsan el uso de la red y el consumo de aplicaciones móviles en diferentes áreas como la comunicación, búsqueda de información, viajes, entretenimiento, educación y salud. La Encuesta Nacional Bicentenario 2016 indica que los usuarios de Twitter declaran seguir a los medios de comunicación (56%), líderes de opinión (53%) y políticos e instituciones de gobierno (21%), entre otros.

Este panorama no puede pasar desapercibido por la clase política nacional, mientras algunos analistas políticos hablan de que Twitter se ha convertido en un indicador importante de considerar al momento de ver tendencias otros lo siguen menospreciando como se menciona en Tumasjan et al. [1].

Este artículo tiene como foco determinar la capacidad predictiva de Twitter en las elecciones presidenciales chilenas de 2017.

El resto del artículo se organiza de la siguiente manera: el capítulo II incluye conceptos relacionados con opinión mining y trabajos relacionados con experiencias eleccionarias en distintos países del mundo. El capítulo III describe la metodología de trabajo aplicado en este estudio para continuar a partir del capítulo IV describiendo los principales resultados obtenidos. El capítulo V intenta explicar y plantear una discusión de dichos resultados. Finalmente, en el capítulo VI se presentan las conclusiones y trabajo futuro.

II. CONTEXTO

Sentiment analysis u opinion mining es un área del procesamiento de lenguaje natural encargada de clasificar documentos, frases o palabras en las categorías de positivo, negativo o neutro [2]. En términos generales en sentiment

analysis se habla de un texto bien estructurado de acuerdo a la sintaxis de cada idioma, sin embargo, los microtextos o tweets presentes en Twitter suelen requerir de un trabajo especial a la hora de ser preprocesados dado que los usuarios intentan volcar todo su sentir en sólo 140 caracteres.

Elecciones previas como la de Barack Obama el año 2008, las de Mariano Rajoy y Donald Trump en el año 2016 y la de Angela Merkel en el 2009 y 2017, entre otras, han mostrado que los ciudadanos comunican intensivamente sus opiniones a través de las redes sociales ya sea para mostrar adherencia o aversión. Estos escenarios políticos han sido analizados en estudios como el de O'Connor et al. [3] en donde se relacionaron medidas de opinión extraídas de encuestas con otras de polaridad extraídas desde Twitter y descubrieron que las encuestas telefónicas tradicionales aplicadas que se hacen durante las campañas electorales podrían ser complementadas e incluso reemplazadas con las opiniones vertidas en Twitter. Por su parte, Tumasjan et al. [1] reveló la utilidad de Twitter vista como una plataforma para la deliberación política, como una herramienta de reflexión sobre sentimiento político y como un buen predictor de resultados de elecciones. También descubrieron que el 40% de las opiniones vertidas en la plataforma eran escritas por sólo el 4% de los usuarios que opinaban. Entre estos estudios también se encuentra el de Kouloumpis et al. [4] cuyo hallazgo dice que podría no ser tan útil en este dominio el uso de características de parth-of-speech para el sentiment analysis de estos micromensajes, a diferencia de características específicas de los microbloggings como son los hashtags, emoticones e intensificadores.

III. MÉTODOS Y HERRAMIENTAS

Como se observa en la Fig.1, la metodología de trabajo para el proceso de opinion mining en tweets consideró cinco etapas:

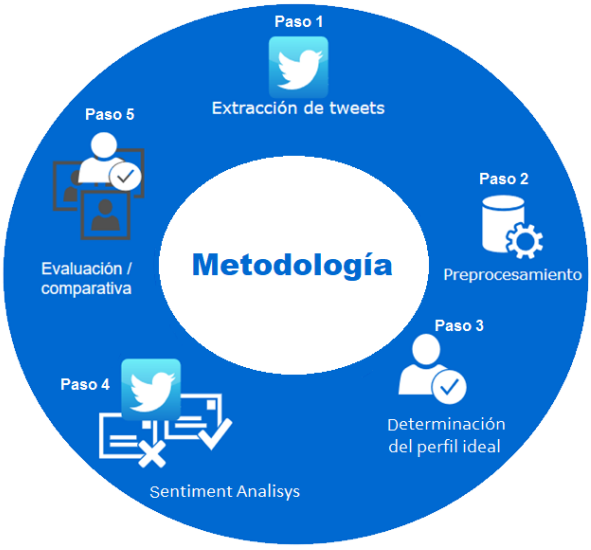


Figura 1. Metodología de trabajo

A. *Extracción y preprocesado de tweets (paso 1 y paso 2)*

62217 tweets fueron recolectados entre mayo y octubre de 2017 considerando a los ocho candidatos presidenciales que llegaron a las elecciones de noviembre 2017. Durante la etapa de extracción se utilizaron herramientas como R versión 3.4.2 (64bit), el paquete TwitterR de R-package y la API de *Twitter* versión gratuita[1]. La **Tabla I** caracteriza los tweets una vez preprocesados considerando la eliminación de aquellos que fueron retweets:

TABLA I. CARACTERIZACIÓN DE LOS TWEETS PREPROCESADOS[2]

Candidatos	Conglomerado	Número de tweets	%
Carolina Goic	Convergencia democráitca	8226	18,19
José Antonio Kast	Independiente	2620	5,79
Sebastián Piñera	Chile Vamos	8169	18,06
Alejandro Guillier	Fuerza de la Mayoría	7166	15,84
Beatriz Sánchez	Frente Amplio	8670	19,17
Marco Enríquez-Ominami (MEO)	Partido Progresista	6081	13,45
Eduardo Artés	Unión Patriótica (UPA)	2644	5,85
Alejandro Navarro	Partido Amplio de Izquierda Socialista	1650	3,65
	Total tweets	45226	

En la fase de preprocesamiento se utilizaron las prestaciones de la función nativa gsub del lenguaje de programación R, considerando la eliminación de puntuaciones, dígitos, tabulaciones, saltos de línea, tweets duplicados, retweets, caracteres especiales, y una lista de stopwords del idioma español. Además, se hizo necesaria la construcción de expresiones regulares para la extracción de algunas URL generadas por Twitter para acortar links y no exceder sus 140 caracteres.

B. *Determinación del perfil ideal (paso 3)*

1) Aplicación de encuesta: Se consideró la determinación de un perfil de candidato ideal a partir de una encuesta en línea aplicada en mayo del 2017, a través de redes sociales, misma forma que se utilizó para la recolección de datos. Se consideró un muestreo aleatorio simple sin reposición. Para el diseño, se revisaron dos encuestas a nivel nacional, la del Centro de Estudios Públicos (CEP)[3] periodo septiembre-octubre de 2017 que en su parte V sondea la opinión acerca de los personajes políticos y la encuesta CADEM[4] con el estudio electoral No. 6 y la encuesta No.199 orientada a determinar el panorama

[1] https://developer.twitter.com/

[2] Se utilizó el mismo orden de aparición en la papeleta de votación

[3] Disponible en https://www.cepchile.cl/estudio-nacional-de-opinion-publica-septiembre-octubre-2017/cep/2017-10-25/105022.html

[4] Disponible en https://www.cadem.cl/encuestas/encuesta-n-199-03-de-noviembre-de-2017/

nacional en varias dimensiones como percepciones políticas y económicas, visión del país y principales problemas y el informe de septiembre de 2017 generado por CERC-Mori[5].

2) Atributos del perfil: El resultado obtenido a partir de las encuestas permitió definir los atributos del candidato ideal a partir de los seis mejor rankeados (Tabla II). Para enriquecer dichos atributos se consideró una lista de sinónimos obtenidos desde WordReference [6] y la Real Academia de la Lengua Española (RAE)[7].

TABLA II. ATRIBUTOS PERFIL IDEAL DE PRESIDENTE DE LA REPÚBLICA

Atributo perfil	Atributos relacionados
1 liderazgo	líder; representante; guía, jefe
2 transparencia	transparente
3 honestidad	honesto; honesta; honrado honrada; honradez; recto; recta; rectitud; íntegro integra; integridad; probo proba; probidad
4 coherencia	coherente; congruente consistencia; consistente
5 credibilidad	creíble; veraz; verosímil
6 confiabilidad	confiable; fiabilidad; fiable

Este perfil, denominado ideal fue contrastado con las opiniones vertidas en los tweets, denominado perfil real, considerando ocurrencia de los atributos y sus sinónimos para determinar luego la distancia entre ambos. Para esta fase se utilizó R.TeMiS (R Text Mining Solution)[5][6] que es un paquete de R (RcmdrPlugin.temis) desarrollado como un plugin de R Commander, que permite analizar, manipular y crear corpus de textos. De este paquete se utilizaron las funciones frequentTerms para la determinación de los términos más frecuentes y termChisqDist, esta última función permite determinar los términos que están más asociados o que co-ocurren con uno o varios términos dados, de acuerdo con la matriz de documentos y términos del corpus.

C. *Sentiment analysis (paso 4)*

Esta etapa fue desarrollada utilizando el filtro para TweetToSentiStrengthFeatureVector para WEKA [4], que permitió determinar intensidades de opiniones positivas y negativas en textos cortos usando enfoque basado en SentiStrength lexicón [7][8][9] y que se ha utilizado en numerosos experimentos con información extraída desde la web[10][11][12][13]. Esta herramienta asigna en forma simultánea una valoración positiva en el rango [1,5] y una negativa en [-5, -1], su algoritmo se basa en un lexicón de 2608 palabras con valencias y una colección de métodos como corrección ortográfica, manejo de negaciones (no; nunca; nada), letras repetidas (triiiiiste), lista de emoticones (☺; ☹; ☺) y amplificadores (super; demasiado; muy).

D. Evaluación y comparativa (paso 5): Esta etapa incluye representaciones de las co-ocurrencias y asociaciones entre palabras vertidas en los tweets y los atributos del perfil ideal. Con el fin de disponer de una medida de comparación entre la predicción y los resultados de las votaciones, se consideró la opinión de un experto para definir el peso asociado a cada atributo del perfil ideal y obtener así una medida a partir de los perfiles reales de cada candidato (2).

$$score_{candidato} = \sum_{1}^{6} weight * atributo \quad (2)$$

IV. PRINCIPALES RESULTADOS

Después de efectuadas cada una de las actividades presentadas en la metodología se encontraron los siguientes hallazgos:

A. *Sentiment analysis*

Los resultados del proceso de polarización se visualizan desde la Fig. 2 hasta la Fig. 9 que representan las frecuencias ponderadas a partir de las valencias positivas y negativas obtenidas en las 11 instancias de medición. La TABLA III detalla los 5 hitos importantes en el periodo de estudio que pueden explicar las fluctuaciones de las polaridades en las opiniones recogidas de cada candidato en el periodo mayo a octubre de 2017.

TABLA III. HITOS PRIMERA Y SEGUNDA VUELTA (2017)

Hito	Descripción	Fecha
H1	Debate ANATEL primarias Chile Vamos	26 junio
H2	Primarias Chile Vamos y Frente Amplio	2 julio
H3	Fin plazo inscripción candidatos en SERVEL[8] e inicio de campaña	21 agosto
H4	Debate presidencial ANP	28 septiembre
H5	Debate presidencial UCH y Radio Cooperativa	3 octubre

ANP = Asociación Nacional de la Prensa; ARCHI = Asociación de Radio difusores de Chile; ANATEL = Asociación Nacional de Televisión de Chile; UCH = Universidad de Chile

Es importante mencionar que en general, el análisis de polaridad para los 8 candidatos mostró que el 90% de los tweets presentan polaridades en los rangos [1, 2] y [-1, -2].

[5] Disponible en http://morichile.cl/wp-content/uploads/2017/10/INFORME-DE-PRENSA-BAROMETRO-POL-SET-20171.pdf
[6] http://wordreference.com
[7] http://rae.es
[8] Servicio electoral de Chile

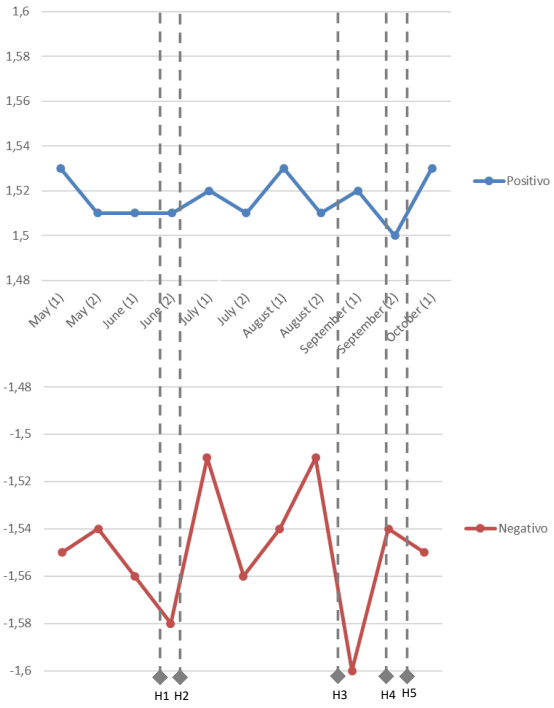


Figura 1. Polarización para candidato Piñera

Para el caso del candidato Piñera (Fig.2) se observó que las mediciones con mayor variabilidad y mayor intensidad están en las opiniones negativas, existiendo un máximo negativo en el mes de septiembre.

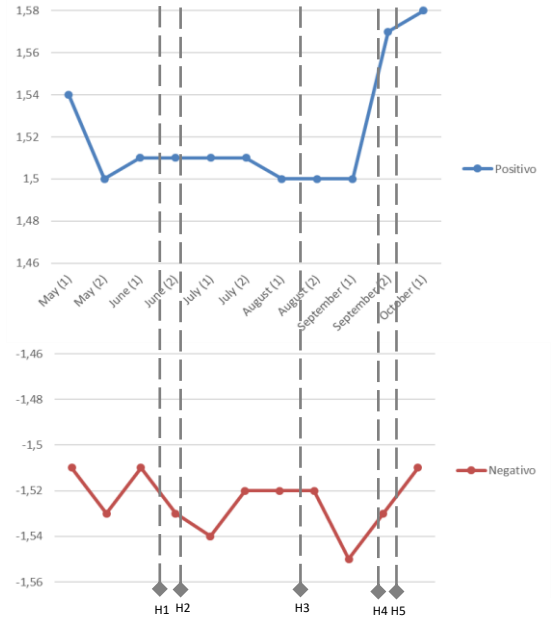


Figura 2. Polarización para candidato Guillier

Los resultados para el candidato Guillier (Fig. 3) mostraron, al igual que el anterior, una baja oscilación en las polaridades positivas y una mayor en las negativas. El máximo positivo para este candidato ocurrió en la primera medición del mes de octubre. Cabe destacar que este tipo de variación es importante considerando la cercanía las elecciones primarias.

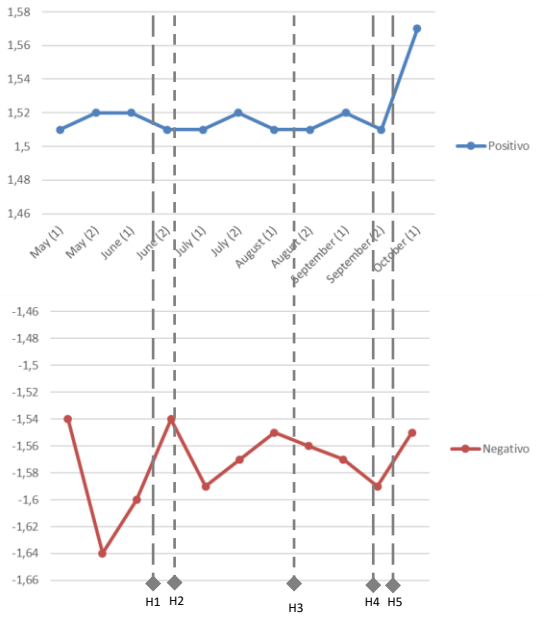


Figura 3. Polarización para candidata Sánchez

La Fig. 4 mostró baja fluctuación en la polaridad positiva de los tweets de la candidata Sánchez siendo más fluctuantes las negativas en especial al principio de la toma de datos. La dimensión positiva aparece con valor punta en la primera quincena de octubre.

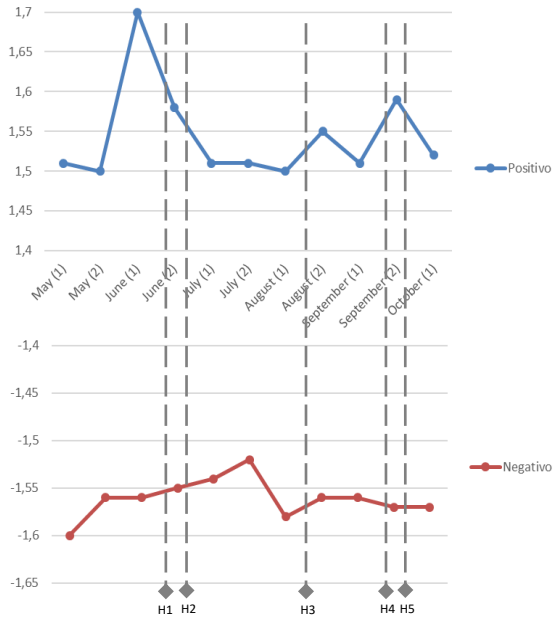


Figura 4. Polarización para candidato Enríquez-Ominami

En el caso del candidato Enríquez-Ominami (Fig. 5) se observa una mayor fluctuación en las valencias positivas en especial en la primera medición del mes de junio, no

obstante, en la medida que avanza el tiempo las intensidades de los tweets son menores.

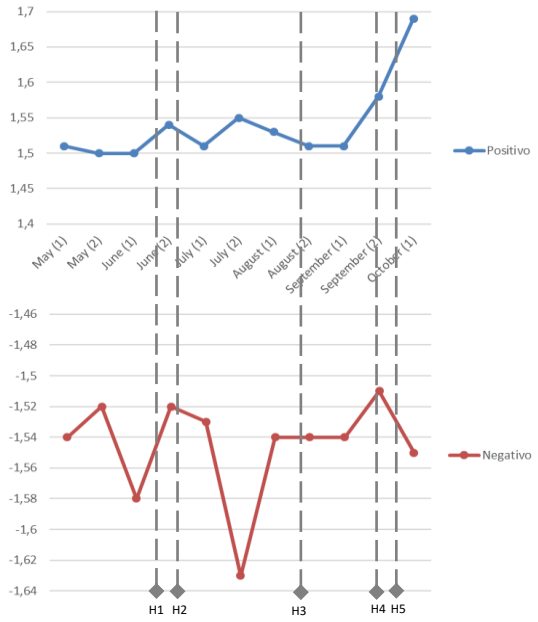


Figura 5. Polarización para candidata Goic

La candidata Goic representada en la Fig. 6 presenta fluctuaciones en ambas polaridades siendo destacable el valor punta de la segunda quincena del mes de julio para las valencias negativas y la primera de octubre para las positivas.

Las Fig. 7, Fig. 8 y Fig. 9 sólo consideran 4 mediciones dado que los candidatos Kast, Navarro y Artés se sumaron a la carrera presidencial el mes de agosto de 2017. En este contexto, se observa en la Fig. 7 que el candidato J.A. Kast presenta valencia positiva sin fluctuaciones significativas, observándose un máximo negativo en la segunda quincena de septiembre.

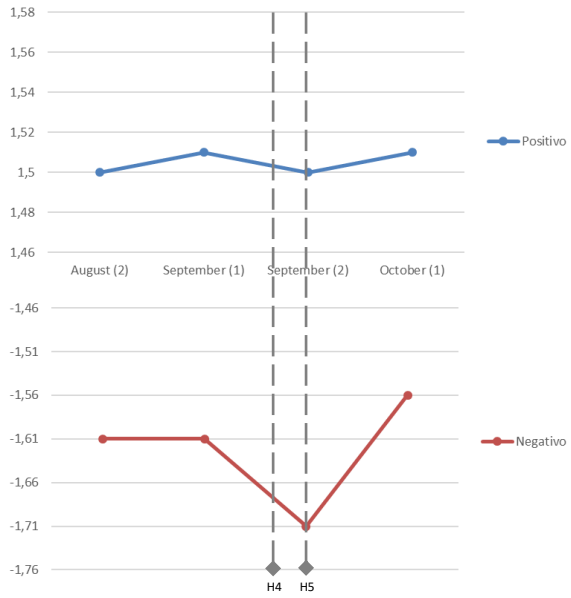


Figura 6. Polarización para candidato J. A. Kast

Navarro (Fig. 8) por su parte muestra mayor fluctuación para las valencias positivas alcanzando su punto más alto en la medición de la primera quincena de octubre después del hito 5 donde los candidatos fueron interrogados por los Premios Nacionales de Chile. Se observa baja fluctuación para el candidato Artés tanto en las polaridades positivas como negativas (Fig. 9).

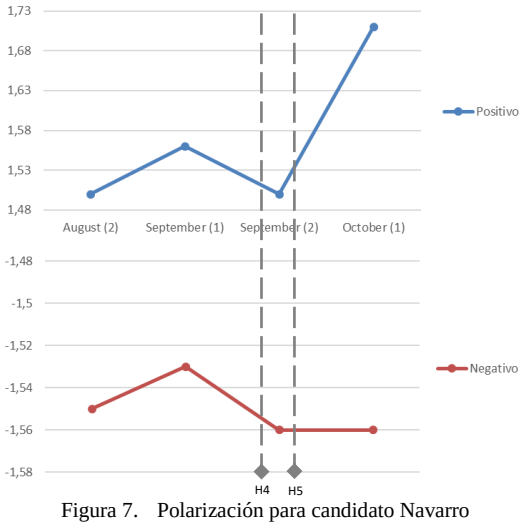


Figura 7. Polarización para candidato Navarro

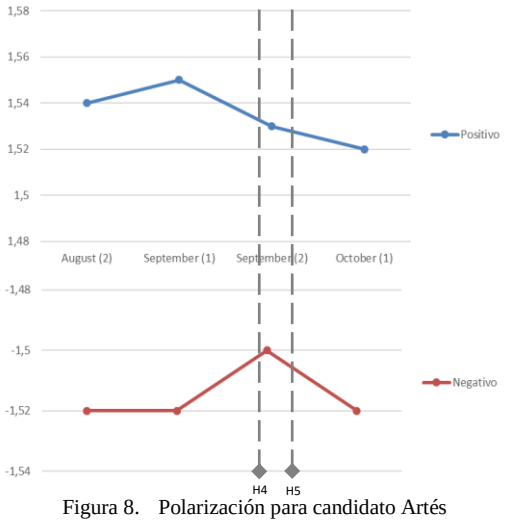


Figura 8. Polarización para candidato Artés

B. Evaluación candidato ideal vs real

La primera parte de esta fase corresponde a un análisis de los 10 términos más frecuentes del corpus, entre los que se encuentran 5 de 8 candidatos como se observa en la Tabla IV no aparecen Marco Enríquez-Ominami, Alejandro Navarro y Eduardo Artés. Hay que mencionar que en este análisis se utilizaron los nombres, apellidos, nombres + apellidos y en algunos casos apodos, como por ejemplo Tatán denominación común dada en redes sociales al candidato Piñera.

TABLA IV. PRESENCIA DE CANDIDATOS EN CORPUS

Candidato	ocurrencias	% del corpus
Carolina Goic	14032	4,00
Alejandro Guillier	13664	3,89
Sebastian Piñera	12321	3,51
Beatriz Sánchez	12223	3,49
J.C. Kast	2526	0,72

A continuación, se muestran los resultados del análisis descriptivo de vocabulario para cada candidato considerando las ocurrencias porcentuales de los atributos del perfil ideal por candidato en cada uno de los puntos de control a partir del mes de mayo hasta octubre de 2017. En primer lugar, la Fig. 10 permite observar cómo los atributos del perfil están presentes en los tweets asociados al candidato Piñera.

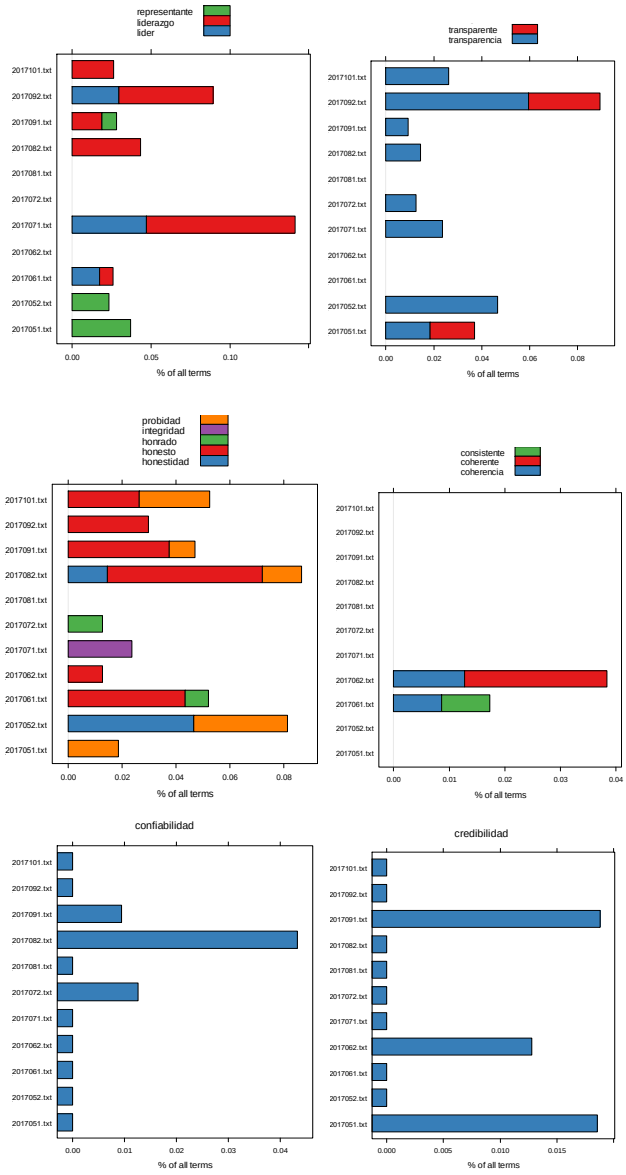


Figura 9. Análisis descriptivo de conceptos en el tiempo (perfil real de Piñera)

Se observa que aquellos mayormente presentes están relacionados con el liderazgo, la transparencia y la honestidad y muy por debajo se encuentran los relacionados a los atributos de consistencia, confiabilidad y credibilidad.

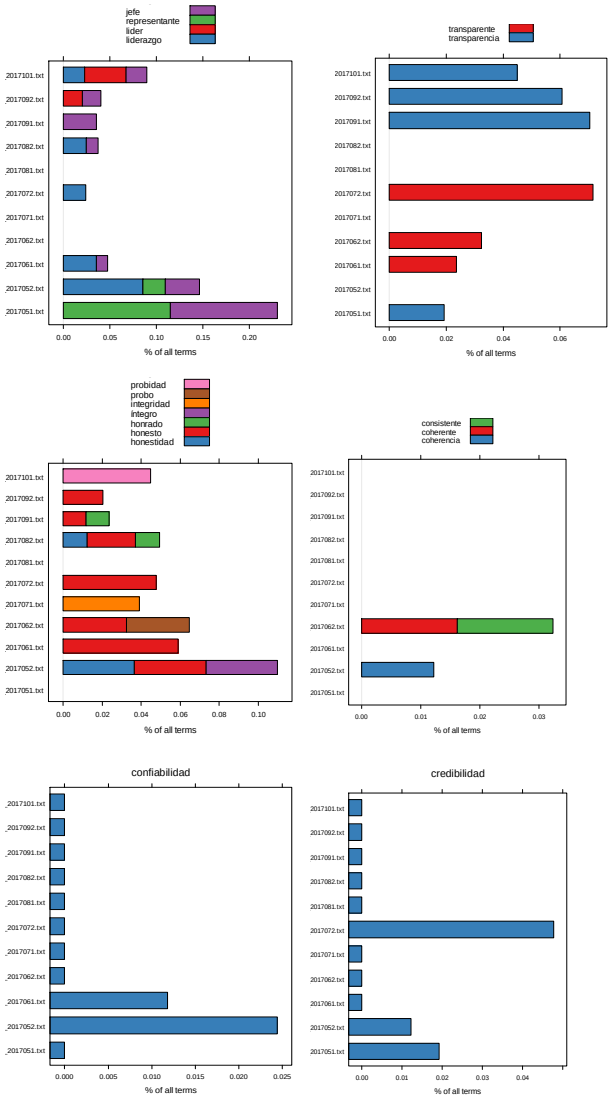


Figura 11. Análisis descriptivo de atributos (perfil real de Guillier)

La Fig. 11 muestra que los conceptos asociados al liderazgo transparencia y honestidad están más presentes en todo el periodo de muestra, por otra parte, los relacionados a consistencia, confiabilidad y credibilidad se presentan en menor proporción y más al comienzo de la campaña, a excepción de credibilidad que tuvo un punto más alto en la segunda muestra del mes de julio.

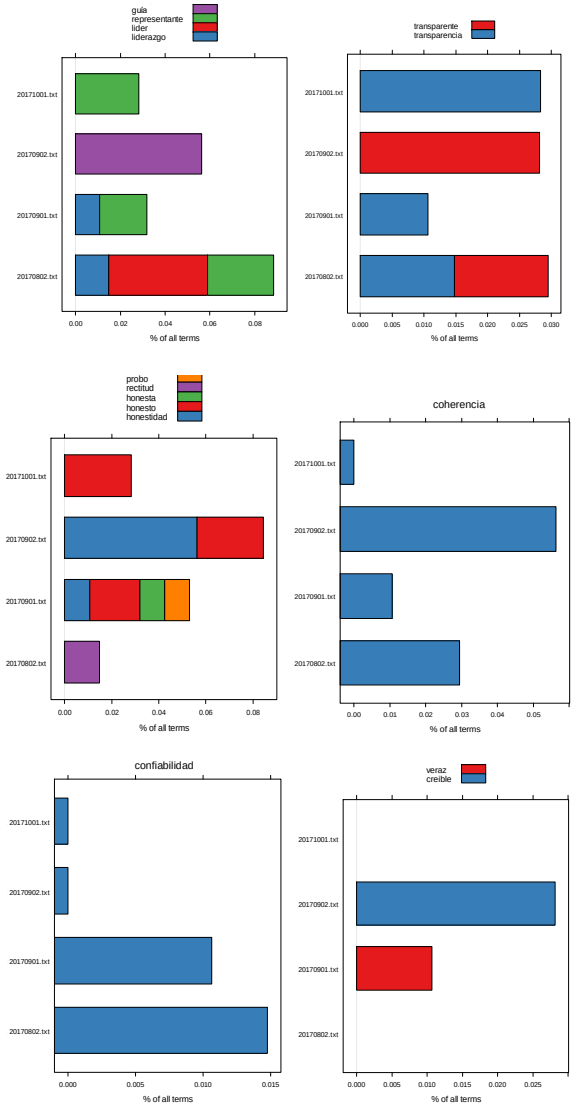


Figura 12. Análisis descriptivo de atributos (perfil real de J.C.Kast)

Se observa para el caso del candidato J.C. Kast que los atributos con mayor ocurrencia corresponden al liderazgo, transparencia y honestidad (Fig. 12).

En la Fig. 13 se marca la presencia de los atributos transparencia, liderazgo y credibilidad, muy por debajo se encuentran honestidad que tiene apariciones al principio del periodo (mayo y segunda quincena de agosto) y coherencia en la mitad del periodo (segunda quincena de julio). No se observan conceptos relacionados con el atributo confiabilidad.

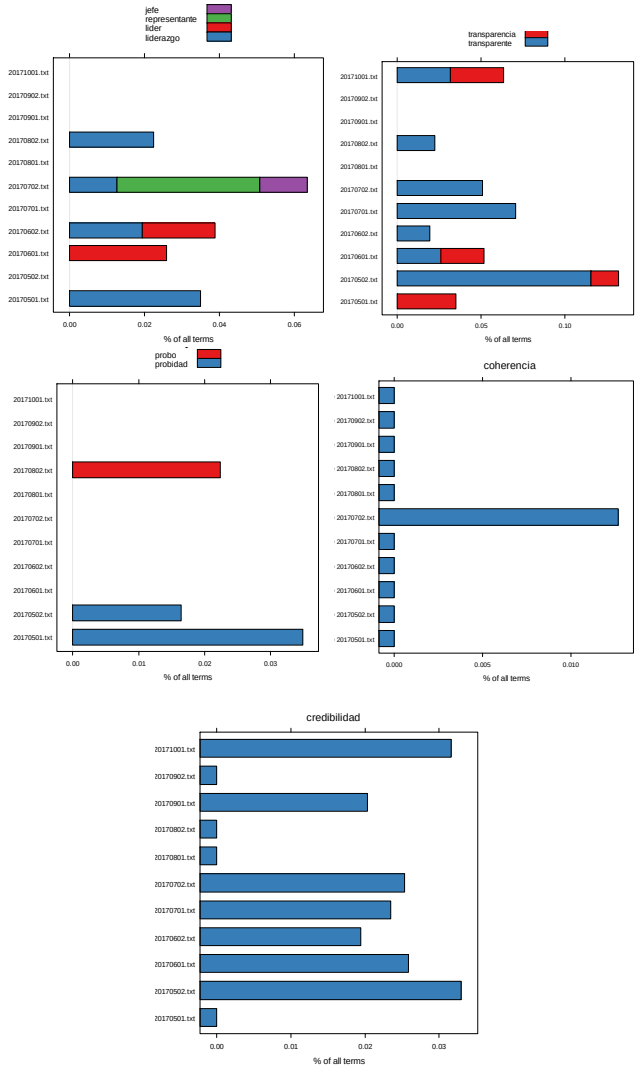
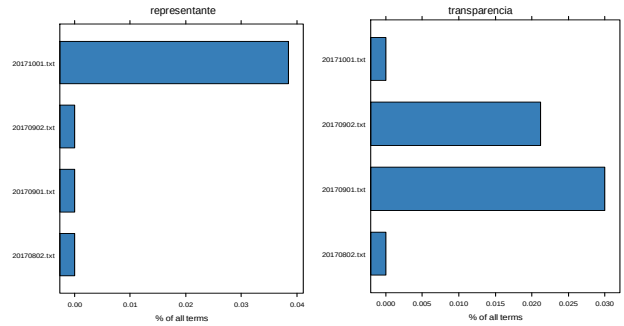


Figura 13. Análisis descriptivo de atributos (perfil real de MEO)

Hay que mencionar que el candidato Navarro sólo tiene datos asociados a cuatro quincenas, dado que éste asumió su calidad de candidato en agosto de 2017. El análisis descriptivo (Fig. 14) indica una baja aparición atributos, sólo 4 de 6 atributos, además de una ausencia de aquellos asociados a la coherencia y confiabilidad. Así, Alejandro Navarro se presenta como uno de los candidatos con menos ocurrencias de los atributos del perfil ideal.



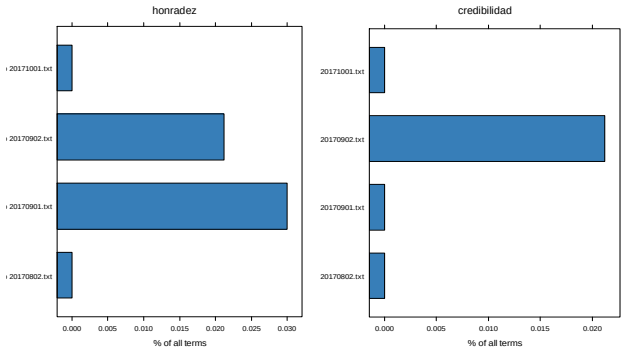


Figura 14. Análisis descriptivo de atributos (perfil real de Navarro)

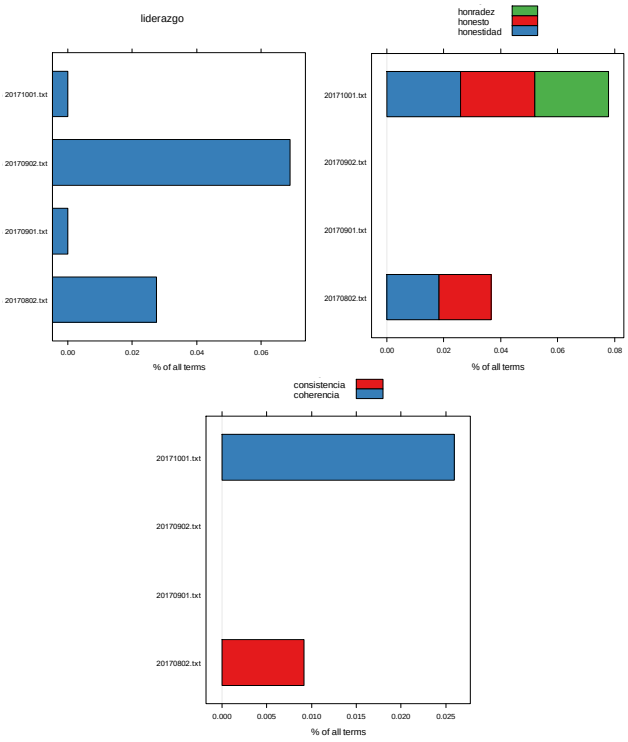


Figura 15. Análisis descriptivo de atributos (perfil real de Eduardo Artés)

Para el candidato Artés (Fig. 15), la aparición de los atributos asociados a liderazgo ocurre preferentemente en 2 de las 4 mediciones de preferencia en la segunda quincena de septiembre, por su parte honestidad tiene un punto más alto al inicio y final del periodo. De la misma forma, el atributo coherencia se presenta al inicio y al final, pero en menor proporción. En este caso no se visualizan atributos relacionados a transparencia, credibilidad y confiabilidad.

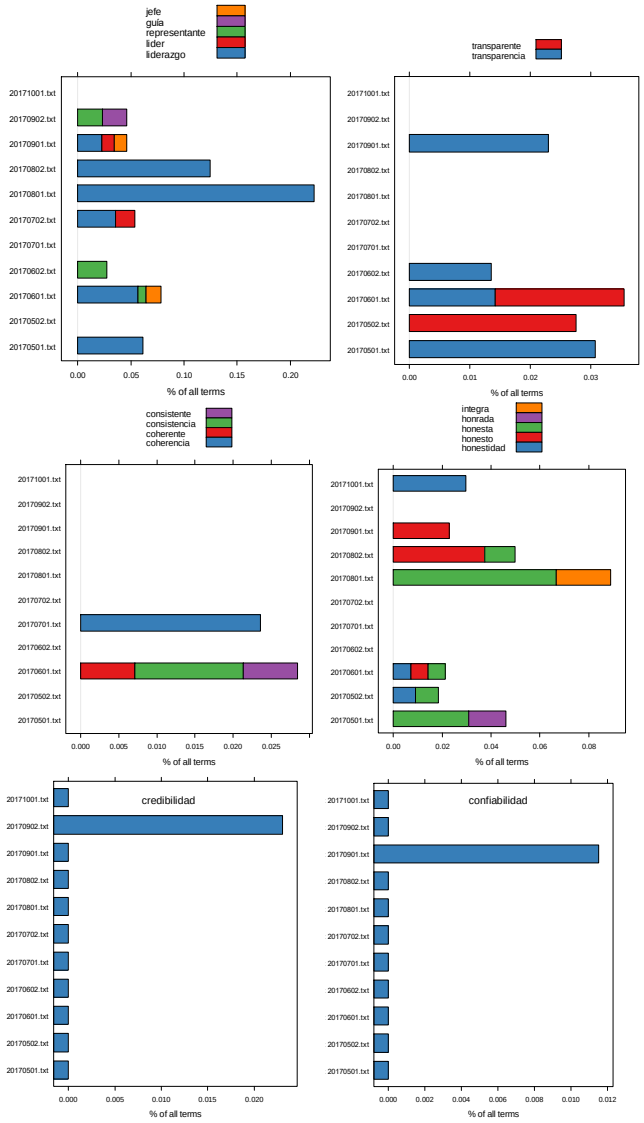


Figura 16. Análisis descriptivo de atributos (perfil real de Beatriz Sánchez)

La Fig. 16 muestra que la candidata Beatriz Sánchez tiene presenta mayor ocurrencia en los conceptos relacionados a liderazgo, transparencia y honestidad. A diferencia de lo anterior sólo en 2 de las mediciones se tiene presencia de los atributos asociados a la coherencia y menor aún a los relacionados a la credibilidad y confiablidad.

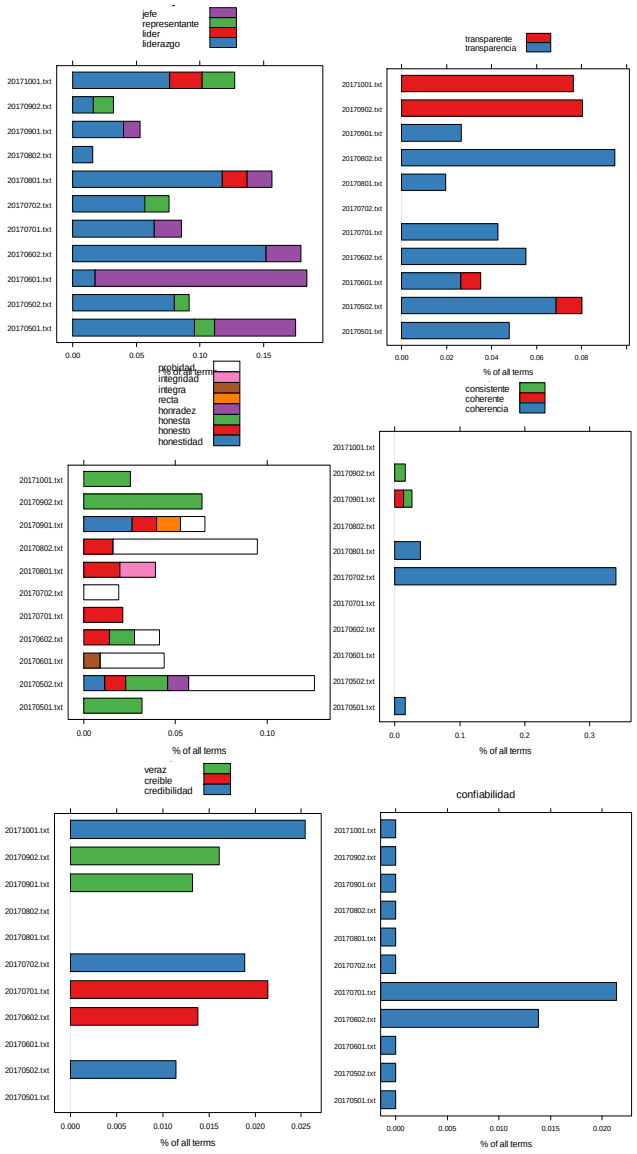


Figura 17. Análisis descriptivo de atributos (perfil real de Carolina Goic)

Se observa en la Fig. 17 que los conceptos asociados a liderazgo, transparencia y honestidad tienen una presencia homogénea durante todo el periodo, con menor aparición se encuentran aquellos relacionados con la credibilidad, coherencia y la confiabilidad. El atributo coherencia y sus relacionados aparecen preferentemente en la segunda quincena de julio y el de confiabilidad la última quincena de junio y primera de julio. Para complementar la obtención de información relevante se confeccionaron las nubes de palabras [14] por cada uno de los candidatos como se observa en la Fig. 18. Dentro de las frases que se pudieron reconstruir a partir de las co-ocurrencias de palabras que aparecen están:

"Presidente de la gente", para el caso de Guillier.

"Será presidente de Chile", para el caso de Piñera.

"Tercera candidatura presidencial", para el caso de Enríquez-Ominami.

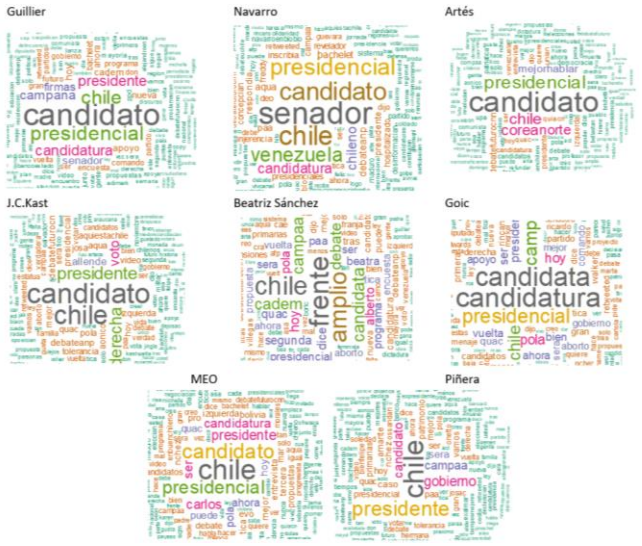


Figura 18. Tag cloud por candidato

A la luz de los resultados de la Fig. 19 se puede observar la conformación del perfil real de cada candidato, también que los atributos con menor presencia fueron la coherencia, credibilidad y confiabilidad y los candidatos que más carecen de atributos son Artés y Navarro con 3 y 4 atributos de un total de 6 respectivamente. La Fig. 20 por su parte permite comparar el score obtenido y la votación real de la primera vuelta ocurrida el 19 de noviembre de 2017 considerando la diferencia en puntos porcentuales correspondiente a cada candidato.

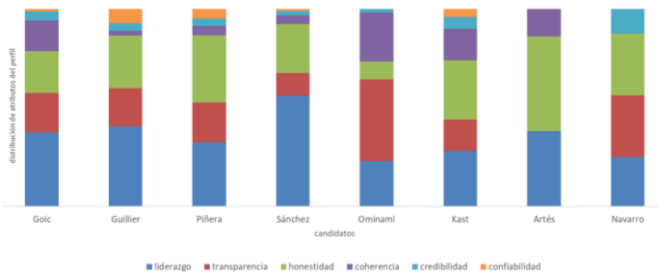


Figura 19. Contribución de los atributos del perfil a cada candidato

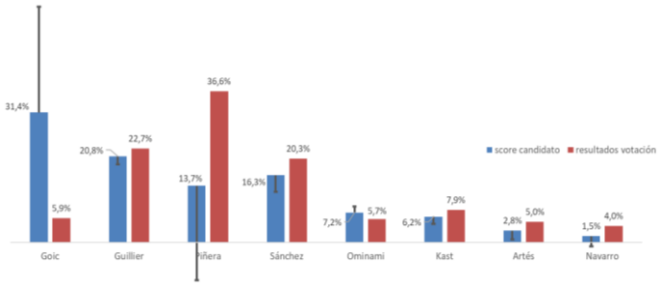


Figura 20. Scores candidatos versus resultados en primarias

V. DISCUSIÓN

Hay que iniciar la discusión indicando que la frecuencia de aparición de los atributos del perfil en los tweets de cada candidato es baja puesto que no supera al 1%. Esta débil presencia podría deberse a que las personas usan twitter para opinar sobre hechos, eventos o personas, pero no necesariamente para describir atributos del objeto de interés.

Con respecto al periodo de muestreo y su contenido, se pueden apreciar algunos comportamientos homogéneos a la hora de hablar de los candidatos presidenciales, tanto positiva como negativamente. Así, la polaridad encontrada en Sebastián Piñera marca un punto más alto negativo durante el primer periodo de septiembre, días posteriores al hito 3, esto puede explicarse debido a que en dicho espacio de tiempo comenzó a circular una polémica fotografía del momento en que hacia un gesto con el dedo medio y que supuestamente iba dirigido a un grupo de pescadores, quienes se habrían manifestado en contra de la Ley de Pesca, la cual fue impulsada bajo su administración. Este hecho generó muchos cuestionamientos al abanderado de Chile Vamos.

Por otra parte, fue notoria el alza en polaridad positiva del candidato Alejandro Guillier durante la primera quincena de octubre (hito 5), quizás por la difusión por parte de su comando de una minuta llamada Motivos para creer, en la que se alinea con avances sociales y reformas, además de desacreditar al sector privado y a su más cercano competidor, que según las encuestas de opinión sería Sebastián Piñera. Esta alza ya se había iniciado a partir del hito 4 en donde manifestó su divergencia con la situación de Venezuela indicando que "nuestro país no necesita aprender de elecciones ajenas. Haremos nuestras reformas en democracia y con amplia participación ciudadana", también habló del tema delincuencia donde mencionó que "hay una sensación de impunidad en las familias chilenas. Hay excesiva facultad de los fiscales para cerrar causas sin investigar (...) Debemos hacer más atractiva la carrera policial. Hay 5 mil vacantes en Carabineros porque hoy nadie quiere serlo".

Por su parte, la candidata Sánchez, alcanzó su mayor valoración negativa la segunda mitad de mayo, pues se criticó al sector que representa por mostrar una visión utópica del país y una posición ambigua sobre ciertos temas, por ejemplo, al tratar de democracia en crisis el régimen de Nicolás Maduro en Venezuela y no de presentar una posición más crítica frente al tema. Su punto positivo más alto, coincidió con los días posteriores al debate al debate UCH (hito 5), lo que podría explicarse con una promesa programática que corresponde a terminar con el Crédito con Aval del Estado (CAE) y la propuesta sobre la creación de un plan para las personas que contrajeron deudas tras financiar sus estudios y finalizar con la publicación de deudas educativas en DICOM y registros comerciales, táctica usada también por el candidato MEO, con el que alcanza su máximo positivo durante la primera quincena de junio al asegurar que parte de su programa incluiría "estatizar las deudas del CAE y condonar las multas e intereses". Respectos a los tweets negativos, este último candidato mantiene una intensidad cuyas fluctuaciones no superan un punto porcentual.

Por su parte Carolina Goic, enfrentó numerosos cuestionamientos durante la segunda quincena de julio tras la publicación de una carta realizada por su esposo Christian Kirk, donde criticó a parlamentarios de la Democracia Cristiana, cuyo contenido conminó a "dar la cara y sus razones" por respaldar al diputado Ricardo Rincón, lo que terminó por poner en duda su carrera presidencial y que la obligó a disculparse públicamente, este hecho podría explicar en parte el máximo negativo alcanzado en ese periodo de medición. La misma candidata presentó un alza en las valoraciones positivas y reducción de las negativas coincidentemente con los días posteriores al debate ANP en donde se mencionó en algunos medios que "ha ganado solidez en el discurso" además de sus propuestas y pronunciamientos sobre invertir en tecnología, desarrollo científico, crecimiento económico y hablar de una nueva constitución.

José Antonio Kast mantuvo relativamente constantes las opiniones positivas presentando un punto más alto en las opiniones negativas en la segunda quincena de septiembre, donde se le criticó en diversos medios y redes sociales por afirmar abiertamente que, si el actual Gobierno cerraba el centro penitenciario Punta Peuco, él lo reabriría de inmediato y que retiraría la estatua de Salvador Allende de la Plaza de la Constitución. Estas opiniones negativas llegan a su punto más alto de intensidad en los días posteriores al hito 3.

La mayor variación positiva en el caso del candidato Alejandro Navarro se presentó en la medición de los primeros quince días de octubre con respecto a su emplazamiento a los demás candidatos a aclarar ciertos temas pendientes para la opinión pública y participar en los debates (hito 4), que habían sido escasos hasta aquella fecha y de poca convocatoria, el haber anunciado en el hito 5 la expropiación de SQM también pudo haber influido en esa alza.

El candidato de la UPA, Eduardo Artés, no presentó mayores variaciones durante todo el periodo de medición tanto de las opiniones positivas como negativas.

Las nubes de palabras coinciden con algunas de las razones expuestas que explicarían las fluctuaciones en las polaridades positivas o negativas, pues los términos con mayor observación dentro de ellas coinciden con las temáticas abordadas en el transcurso del período electoral y que obtienen sus momentos de acentuación en las fechas asociadas a los hitos contemplados en este estudio, como ocurrió en el caso del senador Navarro quien declaró simpatía por Nicolás Maduro en reiteradas ocasiones.

Con respecto a los resultados de la métrica (2) determinada a partir del perfil real de cada candidato y representado en la Fig. 21, si bien es cierto, la candidata Goic fue la que resultó mejor valorada en los distintos atributos, esto no necesariamente se reflejó en la intención de voto final en las elecciones primarias que permitió que Guillier y Piñera pasaran a una segunda vuelta. Una situación similar sucedió en las últimas elecciones norteamericanas, donde las intensidades de las opiniones negativas fueron superiores para el caso de Trump que para Hillary Clinton (Fig. 21), aun así, los resultados favorecieron al candidato republicano.

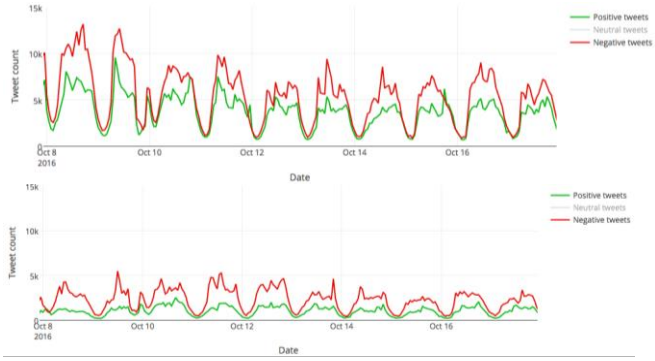


Figura 21. Sentiment analysis Trump v/s Clinton

En la TABLA V se pueden observar los puntos porcentuales de diferencia de las tres encuestas más reconocidas en Chile para los tres candidatos más competitivos en las elecciones primarias. Se observa que el promedio de las diferencias es mayor que el obtenido usando la métrica basada en Twitter propuesta en este artículo, aun cuando las expectativas de un buen instrumento predictivo pudieran ser mucho mayores.

TABLA V. DIFERENCIAS ENTRE ENCUESTAS DE OPINION Y RESULTADOS COMICIOS 2017

Encuesta	Pronóstico (%)					
	Guillier (22,7)	Dif	Piñera (36,6%)	Dif	Sánchez (20,3%)	Dif
CEP	19,7	3 (-)	44,4	7,8(+)	8,5	11,8(-)
CERC-Mori	30	7,3 (+)	44	7,4(+)	11	9,3(-)
CADEM	23	0,3 (-)	45	8,4(+)	14	6,3(-)
Dif $\overline{X}$		3,53		7,86		9,13

Otro aspecto interesante de mencionar fue la aparición de chilenismos, palabras o frases que corresponden a una variante del español propio de nuestro país que presenta ciertas diferencias según distribución geográfica y nivel cultural [15] [16]. En el diccionario de la lengua española (2014), se registran 2214 chilenismos o términos propios del español de Chile. En general, la incorporación de estos términos o expresiones podría mejorar la calidad del análisis considerándolo dentro de la fase de preprocesamiento. En este caso y considerando los 10 chilenismos más usados en las redes sociales se encontró un volumen de 2,3KB para el total de 3,56MB tweets analizados.

VI. CONCLUSIONES Y TRABAJO FUTURO

Pocos días después de conocer los resultados de la elección presidencial 2017 de Chile, y a la luz de los resultados ya descritos la capacidad predictiva del modelo propuesto a partir de los datos de Twitter se puede considerar baja pero muy comparable con el desempeño de los instrumentos tradicionales. De la misma forma, las diferencias obtenidas mediante encuestas oficiales y los resultados reales hace pensar que se debieran diversificar las fuentes de datos y constituir una plataforma consolidada de predicción donde se recoja la opinión de todos los actores y grupos etarios. Dentro de los trabajos futuros se podría considerar un análisis de aquellos tweets emitidos por los propios candidatos y del contenido de sus agendas

programáticas relacionados con aquellos conceptos de interés para los electores como son: educación, gratuidad universal, seguridad, sanidad, afp, isapre, trabajo, inclusión, entre otras. Atendiendo a la premisa que el 4% de los usuarios de Twitter genera el 40% de los tweets se debiera considerar en la métrica un factor de corrección que dependa de la cantidad de seguidores que tenga quien publica un mensaje, esto con el fin hacer más representativa la muestra de tweets. Así entonces, Twitter ha sido capaz de posicionarse como una fuente primaria de información, aunque en el detalle de su contenido, sino se consideran parámetros correctivos, de proporción y representatividad es muy probable generar tendencias que no necesariamente se ajustan a la realidad, terminando en equivocaciones como las cometidas por sondeos de alta reputación. Por lo tanto, hemos de considerar la red social de tweets como una poderosa herramienta para entregar datos e información relevante para la toma de decisiones, teniendo la precaución de darle un uso con un modelo al menos probado anteriormente o maduro.

AGRADECIMIENTOS

Este artículo es el resultado del trabajo del grupo de investigación SOMOS (SOftware - MOdelling - Science), financiado por la Dirección de Investigación and Facultad de Ciencias Empresariales of the Universidad del Bío-Bío, Chile. Se agradece también el aporte de Francisco Gatica Neira, Doctor en Economía y Gestión de la Innovación y Política Tecnológica del Departamento de Economía y Finanzas de la Facultad de Ciencias Empresariales de la Universidad del Bío-Bío y a la Facultad de Ingeniería de la Universidad Católica de la Santísima Concepción.

REFERENCES

[1] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Election Forecasts With Twitter," *Soc. Sci. Comput. Rev.*, vol. 29, no. 4, pp. 402–418, 2011.

[2] B. Pang, L. Lee, and others, "Opinion mining and sentiment analysis," *Found. Trends® in Inf. Retr.*, vol. 2, no. 1--2, pp. 1–135, 2008.

[3] B. O 'connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," *Fourth Int. Conf. Weblogs Soc. Media, ICWSM 2010, May 23-26, 2010*, 2010.

[4] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," *Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media (ICWSM 11)*, pp. 538–541, 2011.

[5] M. Bouchet-Valat and G. Bastin, "RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R.," *R J.*, pp. 188–196, 2013.

[6] B. Garnier, "R.TeMiS. Une approche intégrée et libre de l'analyse de données textuelles.," 2014.

[7] A. Reyes and P. Rosso, "On the difficulty of automatically detecting irony: beyond a simple case of negation," *Knowl. Inf. Syst.*, vol. 40, no. 3, pp. 595–614, 2014.

[8] D. Vilares, M. Thelwall, and M. A. Alonso, "The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets," *J. Inf. Sci.*, pp. 1–16, 2014.

[9] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *JASIST*, vol. 63, no. 1, pp. 163–173, 2012.

[10] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and Sentiment in Tweets," *Spec. Sect. ACM Trans. Internet Technol. Argumentation Soc. Media*, vol. 17, no. 3, 2017.

[11] Svetlana Kiritchenko Xiaodan Zhu and S. M. Mohammad, "Sentiment Analysis of Short Informal Texts," vol. 50, pp. 723–762.

[12] F. Bravo-marquez, M. Mendoza, and B. Poblete, "Knowledge-Based Systems Meta-level sentiment models for big social data analysis," *Knowledge-Based Syst.*, vol. 69, pp. 86–99, 2014.

[13] M. Thelwall and K. Buckley, "Topic-based sentiment analysis for the social web: The role of mood and issue-related words," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. 8, pp. 1608–1617, 2013.

[14] Y. Hassan-montero, "Usabilidad de los tag-clouds: estudio mediante eye-tracking," *Scire, Represent. y Organ. del Conoc.*, pp. 15–33, 2010.

[15] A. Rabanales, "El Español De Chile: Presente Y Futuro," *Onomázein*, no. 5, pp. 135–141, 2000.

[16] C. Wagner, "Sincronía y diacronía en el habla dialectal chilena," *Estud. Filol.*, vol. 41, pp. 277–284, |2006.