

Análisis de la reacción ciudadana en Twitter a través del modelamiento de tópicos

Analysis of the citizen reaction on Twitter to through topic modeling

Carlos Vílchez-Román (*)

Pontificia Universidad Católica del Perú
<https://orcid.org/0000-0002-6802-053X>

Fernando Palomino Villanueva (*)

Asociación de Comunicadores Sociales Calandria
<https://orcid.org/0000-0003-1706-8821>

Prisea Georgina Vílchez Samanez (*)

Universidad Jaime Bausate y Meza
<https://orcid.org/0000-0003-4563-7753>
pvilchez@bausate.edu.pe

Recibido: 10 de noviembre de 2021

Aceptado: 10 de diciembre 2021

Resumen

En el estudio se analiza la reacción ciudadana en la red social Twitter frente a la vacancia presidencial en el Perú, a partir del uso y validación del modelamiento de tópicos, particularmente del algoritmo *Latent Dirichlet Allocation* (LDA). En noviembre del 2021 se descargaron 17,944 tweets usando el paquete *rtweet* para el programa R. El modelo con los mejores indicadores de bondad de ajuste estuvo formado por ocho tópicos que permitieron caracterizar la reacción ciudadana frente al incidente antes mencionado. Una de las conclusiones de este estudio es que, pese a la alta fragmentación de los discursos políticos en Twitter, el uso del modelamiento de tópicos y del algoritmo LDA permite reconstruir “sentidos comunes” que orientan la opinión en los tweets y retweets analizados. Se puede afirmar, por lo tanto, que su aplicación en el análisis del hashtag #VacanciaPedroCastillo ha validado la utilidad de esta metodología. Además, la información proporcionada en Métodos y Resultados posibilita que el enfoque pueda repetirse en otros estudios, enfocados en la dimensión política o cualquier otro tema de interés para la ciudadanía.

Palabras clave:

Vacancia presidencial, modelamiento de tópicos, *Latent Dirichlet Allocation*.

Abstract

The study analyzes the citizen reaction on the social network Twitter regarding the presidential vacancy in Peru based on the use and validation of the Topic Modeling. For this, the algorithm was used the Latent Dirichlet allocation (LDA). In November 2021, 17,944 tweets were downloaded using the *rtweet* package

for the R program. The model with the best indicators of goodness of fit was made up of eight topics that allowed us to characterize the citizen's reaction to the aforementioned incident. One of the conclusions of this study is that, despite the high fragmentation of political discourses on Twitter, the use of topic modeling and the LDA algorithm allows us to reconstruct "common meanings" that they guide the opinion in the analyzed tweets and retweets. It can be stated, therefore, that its application in the analysis of the hashtag #VacanciaPedroCastillo has validated the usefulness of this methodology. What's more, the information provided in Methods and Results enables the approach to be replicated in other studies, focused on the political dimension or any other topic of interest to citizens.

Keywords:

Topic modeling, Latent Dirichlet allocation, Presidential vacancy

Introducción

Tradicionalmente, la reacción ciudadana frente a hechos importantes se expresaba en foros públicos, como las plazas, o en movilizaciones, también conocidas como marchas. Sin embargo, en los últimos 20 años, dicha reacción suele canalizarse de forma creciente a través de las redes sociales, lo que da lugar –muchas veces– a verdaderas batallas digitales, sobre todo cuando el tema de la confrontación es de naturaleza política o ideológica.

En el caso del Perú, la gestión de Pedro Castillo –elegido presidente en la segunda vuelta de las elecciones generales en junio de 2021– ha sido duramente cuestionada y sus decisiones han generado diversos debates en las redes sociales. De hecho, las críticas empezaron antes de que asuma la presidencia: 1) durante la segunda vuelta de la campaña electoral sus opositores le acusaron de terrorismo, 2) luego de conocerse los resultados de la segunda vuelta le acusaron de fraude electoral. Ya durante su gobierno, se le ha denunciado por diferentes escándalos en los que se vieron envueltos sus ministros y funcionarios nombrados. Los grupos opositores han apuntado a desacreditarlo y a cuestionar no solo su “representatividad” sino también su “legitimidad”, lo cual ha dado lugar a que en las redes sociales se generen corrientes de opinión en torno a la vacancia presidencial cada vez que se hace una denuncia contra Castillo.

Esto ocurre en un contexto político en el que la mayoría de expresidentes o candidatos a presidentes de los últimos 40 años están investigados, encarcelados o con arresto domiciliario por diversos delitos (desde casos de corrupción hasta violaciones de derechos humanos). Asimismo, en los últimos 4 años, el Perú ha tenido 5 presidentes. De ellos, dos renunciaron (Pedro Pablo Kuczynski 2018, Manuel Merino 2020) y otro fue vacado e inhabilitado (Martín Vizcarra 2020).

Todo eso da una idea de la fragilidad del sistema político democrático en el Perú y la importancia creciente que van adquiriendo las redes sociales en la configuración de la agenda pública, en las decisiones políticas claves y en el mantenimiento de la gobernabilidad democrática.

La presente investigación busca analizar la reacción ciudadana en la red social Twitter frente a la propuesta de vacancia presidencial de Pedro Castillo (expresada a través del *hashtag* #Vacancia-PedroCastillo) a partir del uso y validación del modelamiento de tópicos.

Se debe precisar que la muestra analizada abarca el período del 17 al 24 de noviembre del 2021. Esto es importante tener en cuenta ya que el *hashtag* #VacanciaPedroCastillo es uno de los tantos que los grupos opositores han circulado en las redes desde el inicio de su gestión gubernamental y, por las fechas seleccionadas, los *tweets* y *retweets* analizados están relacionados específicamente con el caso del exsecretario de Palacio de Gobierno Bruno Pacheco ¹.

Para comprender la respuesta de los usuarios de esta red social se usó la minería de texto, en particular un algoritmo para modelamiento de tópicos conocido como *Latent Dirichlet allocation* (LDA). En las siguientes secciones se revisan algunos estudios que emplearon el modelamiento de tópicos para analizar el contenido de los *tweets*, luego se detalla el procedimiento empleado para la recolección y análisis de datos, y se concluye con la presentación y discusión de los resultados.

¹ Bruno Pacheco fue secretario de Palacio de Gobierno entre agosto y noviembre del 2021. En ese lapso fue denunciado por influir en los ascensos militares y favorecer a empresarios en licitaciones del gobierno (tráfico de influencias). En noviembre fue denunciado nuevamente por ocultar 20 mil dólares en un baño de Palacio y actualmente es investigado por lavado de activos.

Marco Teórico

En América Latina, la participación ciudadana en redes sociales ha sido estudiada previamente ya sea para analizar cómo fomentar dicho activismo en diarios alternativos (Harlow, 2015) o para explorar la construcción de narrativas políticas (Rodríguez et al., 2017).

El uso de las redes sociales, como Twitter, durante las campañas electorales ha sido también objeto de estudio, sea mediante el análisis de las estrategias utilizadas por los candidatos (Carreazo, 2018; Angulo et al., 2018), la forma como construyen su relación con los votantes (Alfaro, 2018; Dueñas et al., 2016) o el comportamiento de los electores (Deltell, et al., 2013).

Asimismo, en contextos no electorales, se han estudiado las actitudes o reacciones de los usuarios de Twitter frente a determinados hechos de coyuntura, como la pandemia del Covid-19 (Manfredi et al., 2021), la construcción de la agenda pública desde las redes y los medios (Rubio, 2014), el debate acerca de la violencia de género (Fernández, 2018) y los discursos del odio en línea (Del Valle, 2021), o la forma en que se relacionan las autoridades con la ciudadanía durante las crisis (Pérez, 2020). Es decir, el rango de temas políticos abordados en la investigación del uso de las redes sociales como Twitter es bastante amplio y no se circunscribe a la coyuntura electoral o poselectoral, sino que incorpora temas claves de la gestión pública.

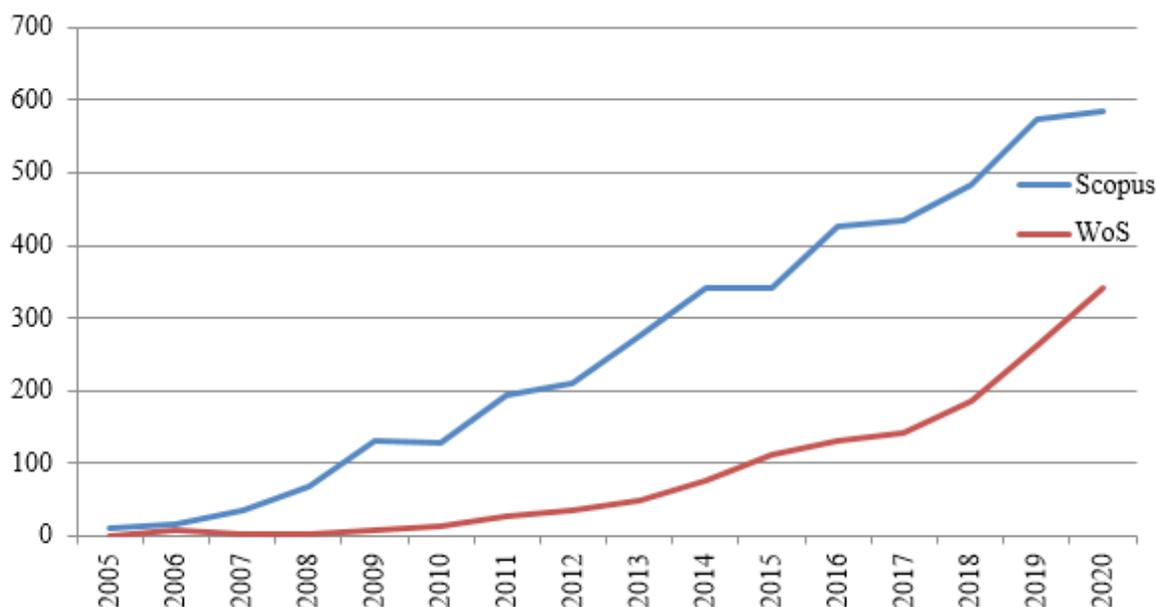
Modelamiento de tópicos con Latent Dirichlet allocation (LDA)

Tradicionalmente, para analizar el contenido de los mensajes de forma automática (enfoque también conocido como análisis de contenido por computadora) se suele usar un diccionario de códigos, con el cual se calcula un puntaje en base a la cantidad de veces que los términos incluidos en el diccionario aparecen en el conjunto de documentos analizados. En líneas generales, es la forma como trabajan los programas General Inquirer o el Linguistic Inquiry and Word Count (LIWC), que se han empleado en varios análisis de contenido por computadora. Por ejemplo, los trabajos más recientes que han empleado el diccionario LIWC han permitido medir la respuesta emocional a la música (Yinger & Springer, 2019), examinar la efectividad del liderazgo de los gerentes estadounidenses durante las crisis financieras (Josef & Helena, 2019) o analizar el estilo lingüístico durante las negociaciones diplomáticas (Bayram & Ta, 2019), entre otras aplicaciones.

Una de las principales limitaciones del enfoque analítico basado en el uso de un diccionario de códigos es que las categorías son fijas. Si bien es cierto que las palabras pueden asignarse a más de una categoría, una vez que se obtiene un puntaje, este ya no cambia, porque el diccionario de códigos es de naturaleza estática. Para obtener nuevos puntajes deben diseñarse y validarse nuevos diccionarios de términos, con nuevos conjuntos de documentos y seguir afinándolos hasta que los puntajes se estabilicen.

La aproximación a los textos cambia cuando se emplea el modelamiento de tópicos, un enfoque analítico que permite analizar grandes volúmenes de datos textuales para identificar patrones latentes que revelen criterios de organización. Este enfoque ha adquirido una importancia creciente en los últimos años, en particular el algoritmo conocido como Latent Dirichlet allocation (LDA). Una búsqueda en las bases de datos Scopus y Web of Science (WoS) muestra que el número de publicaciones en las que se empleó el algoritmo LDA ha aumentado significativamente en los últimos 15 años (ver Figura 1).

Figura 1
Publicaciones en las que se usó el algoritmo LDA entre los años 2005-2020



Nota: Para la búsqueda se usó el término “Latent Dirichlet allocation”. En Scopus y WoS se utilizó el campo Resumen (Abstract). Fecha de la búsqueda: 24 de noviembre del 2021.

LDA tiene una aproximación analítica diferente porque asume que un documento o colección de documentos puede abordar varios temas a la vez, los cuales se expresan en términos con mayor o menor probabilidad de pertenecer a un tópico; es decir, tienen un carácter probabilístico. Para el modelamiento de tópicos, LDA analiza la ocurrencia simultánea de raíces de palabras que sean similares (Bleu et al., 2003).

Análisis de tweets usando LDA

Al ser un enfoque analítico que permite identificar patrones y tendencias latentes en una colección de documentos, el algoritmo LDA se ha aplicado en diversos campos y usado con múltiples idiomas. Dado que LDA identifica la ocurrencia simultánea (o coocurrencia) de las raíces de las palabras y que los modelos analizados se basan en términos normalizados, es un algoritmo flexible que puede emplearse con diversos idiomas, aunque la mayoría de las aplicaciones se han desarrollado para el idioma inglés.

Esta versatilidad para analizar un corpus documental y extraer los temas o tópicos latentes ha llevado a que LDA se use para analizar las opiniones de los usuarios de servicios de salud (Islam, 2019; Puranam et al., 2017), las preferencias de los consumidores por productos competitivos (Wang et al., 2017), las noticias deportivas (Hidayatullah et al., 2018), la clasificación de sentimientos de los clientes de un hotel ecoturístico (Calheiros et al., 2017), entre otras aplicaciones. La posibilidad de trabajar con grandes volúmenes de texto hace que LDA sea muy útil para analizar los comentarios difundidos en redes sociales, porque estos se producen rápidamente, su costo de almacenamiento es barato y existen herramientas gratuitas para descargar datos textuales publicados en Twitter o YouTube.

Con relación a la aplicación del LDA para analizar el contenido de los *tweets*, un grupo de investigaciones dentro de la ciencia de la computación se ha enfocado en la mejora del algoritmo (Ingle & Emmanues, 2016; Islam, 2019; Megala et al., 2018; Zhao et al., 2011), mientras que un segundo grupo orientado a la comunicación se ha orientado al análisis político y la percepción de los servicios de salud (Ryo & Bendle, 2017; Kwon et al., 2019).

En el primer grupo encontramos estudios que han comparado Twitter con los medios informativos tradicionales (Zhao et al., 2011) o que han examinado cómo separar los temas de fondo de los temas de actualidad en base a la popularidad de los tópicos extraídos (Ingle & Emmanues, 2016), cómo combinar el análisis de sentimiento con el modelamiento de tópicos usando datos de la empresa Uber (Alamsyah et al., 2018) o para recomendar *hashtags* relevantes a partir de los comentarios de los usuarios (Kannan et al., 2018).

El segundo grupo de investigaciones se ha orientado hacia el contenido de los *tweets*. Con relación a la estrategia de los partidos políticos en redes sociales, se examinaron 10,602 *tweets* publicados en la cuenta Twitter de los candidatos demócratas y republicanos para las elecciones primarias en Estados Unidos el año 2016. Los tópicos preferidos por los demócratas fueron: derechos de la mujer, persona, familia y justicia. Por el lado de los candidatos republicanos, destacaron los siguientes temas: gobierno, nación, política y comercio (Ryo & Bendle, 2017). En un estudio reciente se analizaron casi 420,000 *tweets* emitidos a raíz del tiroteo en la mezquita de Québec, ocurrido el 19 de enero del 2017. El modelamiento de tópicos identificó los tópicos relevantes frente a este incidente violento: víctimas, conflicto entre supremacistas y el islam, crimen de odio y desinformación, inmigración y fobia al islam, comunidad musulmana y solidaridad (Kwon et al., 2019).

Asimismo, se ha usado el LDA para determinar los temas abordados por los distintos grupos de la clase política peruana (de los poderes Ejecutivo y Legislativo) a través del análisis de los contenidos compartidos por sus miembros en sus cuentas de Twitter, en el periodo de enero a setiembre del 2018. Esto permitió detectar que la principal prioridad de los políticos peruanos ha sido compartir las labores de gestión que conllevan su cargo y que el tema de corrupción judicial ha sido uno de los de mayor difusión en el año. Al comparar las agendas por grupos políticos, se observa que los temas son disímiles: mientras que el Poder Ejecutivo otorga importancia a las actividades de inclusión social, Fuerza Popular y las bancadas de izquierda dan más espacio a sus actividades partidarias (Gamboa, 2019).

En cuanto a la reacción ciudadana en los países latinoamericanos sobre la vacancia presidencial, en la revisión de la literatura no se encontraron estudios previos que hubiesen empleado LDA para analizar *tweets* referidos a este tema. Por ello, al no contar con suficiente evidencia empírica para plantear una hipótesis, en ese estudio compararemos el modelamiento de tópicos, basado en LDA, entre los mensajes originales y aquellos con contenido no original (*retweets*).

Métodos

Todo el proceso de descarga, preprocesamiento y modelamiento se realizó en el ambiente de análisis R, aunque también se generaron archivos Excel con los datos descargados para permitir una rápida exploración visual de dichos datos. A continuación, describimos la fuente de información empleada, los criterios de inclusión y exclusión, el proceso de recolección de datos, las tareas de preprocesamiento y los parámetros empleados para el modelamiento de tópicos con el algoritmo LDA.

Fuente de información

Elegimos la red social Twitter por dos razones. Primero, por ser una de las pocas plataformas que permite descargas masivas de información usando un conector para bases de datos, conocido como *Application Programming Interface* (API). En segundo lugar, por el volumen de información difundido a través de esta red. En noviembre del 2021, Twitter contaba con 322.4 millones de usuarios activos, es decir un número grande de personas publicando y compartiendo mensajes a cada momento. Según el portal *statista.com*, en todo el mundo, en octubre del 2021 el 70.4% de usuarios fueron varones y el 29.6% mujeres.

Criterios de inclusión y exclusión

Solo se incluyeron los *tweets* asociados con un *hashtag* que daba cuenta de la reacción de los ciudadanos y los partidos políticos frente al pedido de vacancia presidencial para el presidente Castillo, luego de los primeros 100 días de su mandato como máxima autoridad del Ejecutivo. Las medidas más criticadas fueron las referidas al nombramiento de funcionarios con escasa o nula trayectoria en gestión pública. No se excluyó ningún *tweet* ya que el propósito fue mostrar el abordaje de este tema de la manera más amplia posible.

Recolección de datos textuales

El 24 de noviembre del 2021 se descargaron los *tweets* publicados en el *hashtag* #VacanciaPedroCastillo (la muestra abarcó el período del 17 al 24 de noviembre). Para la descarga usamos el programa *rtweet*, versión 0.7.0 (Kearney, 2019), una API desarrollada para el programa R. Se descargaron dos conjuntos de datos, en el primero se consideraron los *retweets* ($n = 17,944$) y en el segundo estos se omitieron ($n = 17,611$). Empleamos la configuración predeterminada de la función de búsqueda que en la API estándar permite descargar hasta 18,000 *tweets* publicados en los últimos siete días.

Preprocesamiento de los datos textuales

Antes del modelamiento de tópicos, se filtraron las palabras no significativas de los dos conjuntos de datos. Para ello se empleó el filtro para artículos y preposiciones en español (o lista *stopword*), usando el parámetro de idioma español del paquete *textmining*™ para R. Además, se eliminaron los signos de puntuación y los números. Luego, el texto se transformó a palabras en minúsculas para que no hubiera diferencias en la forma como estaban escritos los términos. De esta forma, palabras como SOCIALISTA o Socialista se convirtieron en socialista. El mismo procedimiento se aplicó a términos similares como PATRIOTA, COMUNISTA, MOVADDEF o VACANCIA. En esta fase de preprocesamiento solo se consideraron los términos formados por una o dos palabras (unigramas o bigramas). Con estos parámetros se creó la matriz documento-término, la cual lista los términos candidatos para cada documento y la presencia o ausencia de ellos dentro de cada documento. Para generar esta matriz usamos el programa *textmineR* para R, versión 3.0.5 (Jones, 2019). Para analizar la efectividad del algoritmo de lematización (es decir, la identificación de la raíz de las palabras) de Porter se generó una matriz documento-término para el conjunto de datos textuales por ser analizados.

Modelamiento de tópicos con LDA

Empleamos la función *FitLdaModel* del paquete *textmineR* con los siguientes parámetros: tópicos = 5, $\alpha = 0.10$, $\beta = 0.05$, iteraciones = 200. Además, calculamos los indicadores de bondad de

ajuste de cada modelo (R^2 y logaritmo de verosimilitud) y el coeficiente de coherencia, cuyo rango de valores va de 0.01 a 0.99 y su magnitud indica la fuerza de asociación entre los principales términos de cada tópico. Como el R^2 aumenta a medida que crece el número de tópicos, el criterio de demarcación para determinar el número de tópicos fue el valor más alto para el logaritmo de verosimilitud y el coeficiente de coherencia. Como el logaritmo se expresó como número negativo, el valor más alto correspondió al número menor, porque los números negativos más grandes son aquellos cercanos a uno (ver Tabla 1).

Tabla 1
Indicadores de bondad de ajuste de las matrices documento-término

Tópicos con LDA	Matriz con retweets		
	R^2	Log V	Coherencia
3	0.09036	-233578	0.26253
4	0.12993	-233927	0.30174
5	0.13270	-232723	0.25890
6	0.15307	-232623	0.27580
7	0.18221	-232641	0.19065
8	0.18804	-231778	0.31016
9	0.22659	-231575	0.27714
10	0.23855	-232307	0.28268

Fuente: Elaborado por los autores con los resultados del estudio

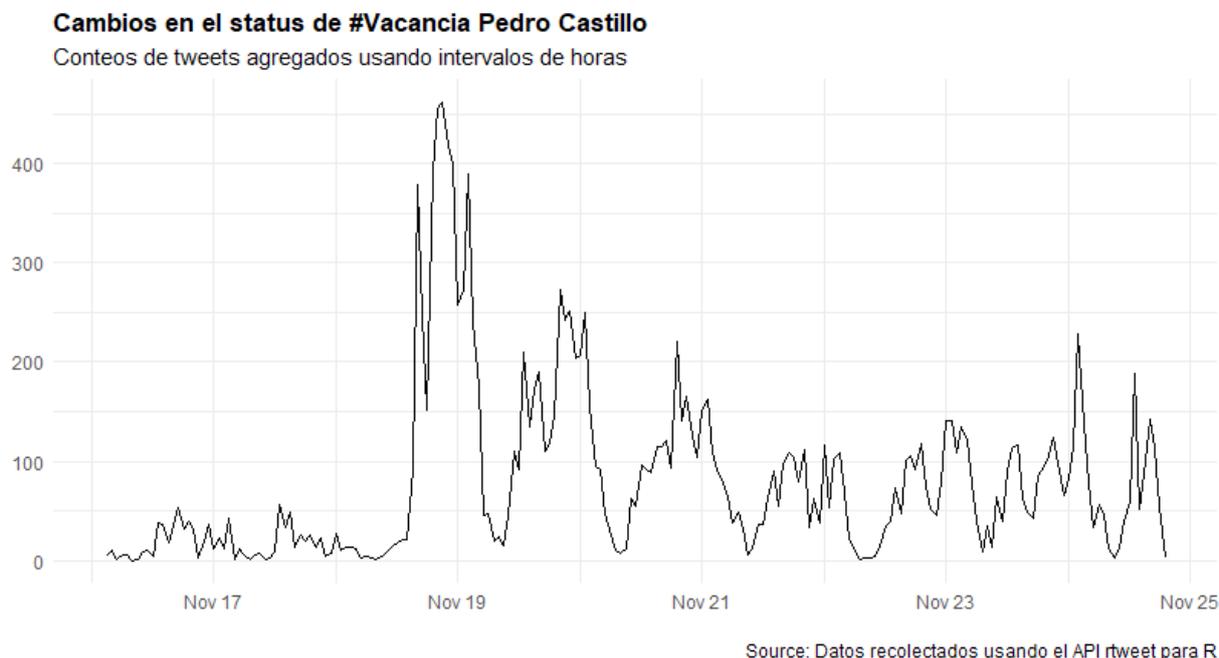
Nota: LogV = Logaritmo de verosimilitud, expresado como valor negativo

Resultados

La mayoría de *tweets* asociados con el *hashtag* #VacanciaPedroCastillo se publicaron después de que el Congreso de la República anunciara que iba a presentar una moción para sustentar la solicitud de vacancia presidencial.² Luego de ese período inicial, el tema de la vacancia presidencial todavía concita el interés en la red Twitter, aunque no con la misma intensidad de los primeros días (ver Figura 2).

2 El 18 de noviembre del 2021 la vicepresidenta del Congreso, Patricia Chirinos, pidió durante la sesión del pleno del Congreso de la República que se firme una moción de vacancia presidencial contra Pedro Castillo por incapacidad moral. El 25 de noviembre se presentó la moción de vacancia contra Pedro Castillo por incapacidad moral permanente. De acuerdo con la Constitución peruana (artículo 113), el presidente puede ser vacado por varias causales, entre las que está “su permanente incapacidad moral o física, declarada por el Congreso”. Finalmente, el 7 de diciembre la mayoría del Congreso de la República rechazó admitir la moción de vacancia presidencial contra Pedro Castillo (46 votos a favor, 76 en contra y 4 en abstención).

Figura 2
Cambios en el estatus de #VacanciaPedroCastillo



Fuente: Elaborado por los autores con los resultados del estudio

Nota: El conteo de tweets se hizo usando conteos con intervalos por hora

Al comparar los indicadores de bondad de ajuste de los ocho modelos de tópicos, encontramos que el modelo con ocho tópicos para el conjunto de datos con *retweets* y lematización de Porter obtuvo un coeficiente de coherencia = 0.31016, cuando el logaritmo de verosimilitud alcanzó un alto valor (ver Tabla 1). Por esa razón, la extracción de términos y tópicos más relevantes se hizo a partir de dicha matriz de documento-término.

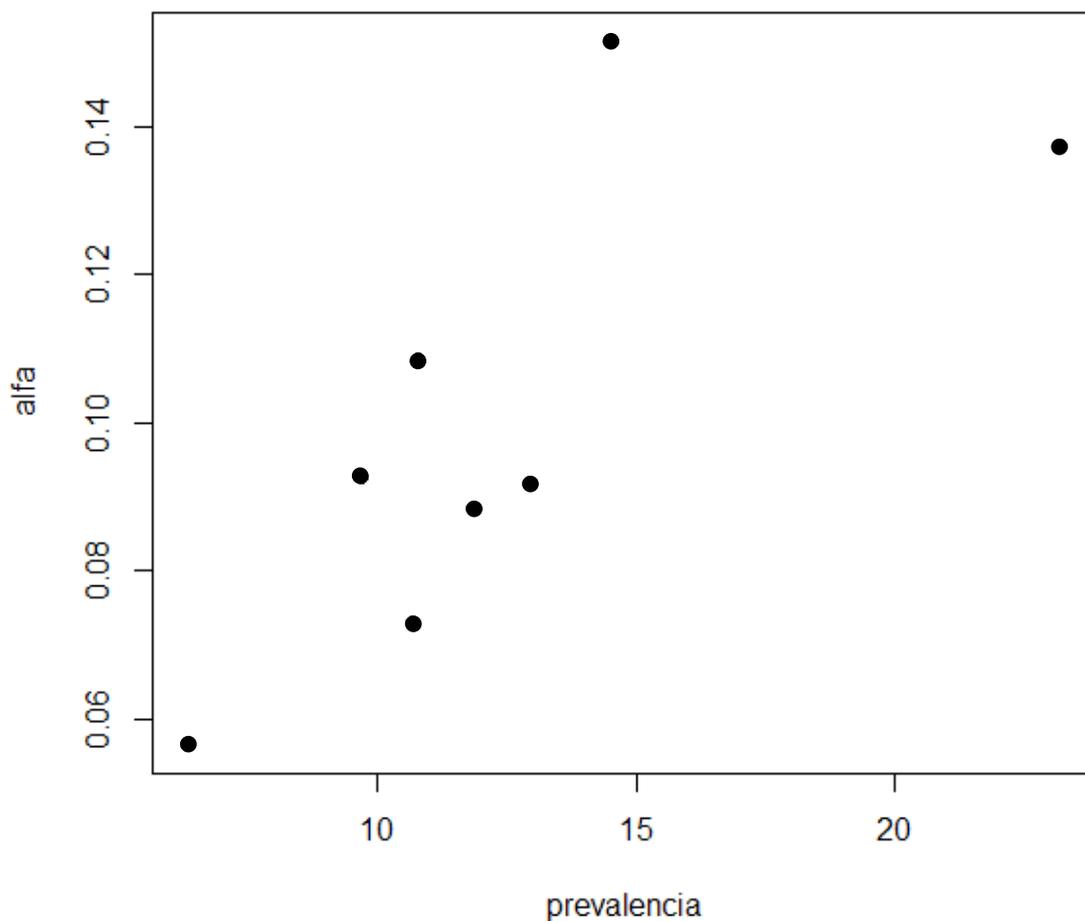
Con relación a la calidad de los tópicos, la coherencia probabilística mostró una asociación moderada entre las palabras que forman cada tópico (0.31 ± 0.22 , intervalo de confianza al 95% [0.1276 – 0.4928]). Es decir, las palabras están relacionadas con los tópicos y son dependientes entre sí. Luego se identificaron los términos que aparecieron con más frecuencia en cada tópico (ver Tabla 2) y se verificó la proporcionalidad entre el tópico y su probabilidad de representar cada documento (ver Figura 2).

Tabla 2
Términos más relevantes para los ocho tópicos del modelo seleccionado

Tópico	Término 1	Término 2	Término 3	Término 4	Término 5
1	Día	Gobierno	Castillo	Pueblo	Voz
2	Dicen	Vacancia	Pedro	Castillo	Dio
3	Pueblo	Arequipa	Pedro Castillo	Palacio gobierno	Marcha
4	Castillo	Vacancia	Dólares	Sábado	Dólar
5	Campesino	Palacio	Carajo	Mil	Pueblo
6	Baño	Perú	Vacancia Carajo	Mil dólares	Pueblo voz
7	Bueno	Pueblo	Secretario	Sacará	Voz dio
8	Fiscalía	Gobierno	Pacheco	Castillo	Gobernabilidad

Fuente: Elaborado por los autores con los resultados del estudio

Figura 3
Prevalencia de cada tópico



Fuente: Elaborado por los autores con los resultados del estudio

Finalmente, se asignaron de forma automática etiquetas para cada tópico y se integraron los indicadores de calidad de los tópicos y los términos que tuvieron la mayor probabilidad de representar cada tópico (ver Tabla 3). Si bien es cierto, la identificación de etiquetas y términos se basó en técnicas de procesamiento de lenguaje natural, los autores revisaron el 10% de los *tweets* para verificar que tenían un fraseo para las etiquetas y los términos.

Tabla 3
Prevalencia y coherencia de los tópicos extraídos

Tópico	Etiqueta	Coherencia	Prevalencia	Términos top
4	Miles de dólares	0.506	14.524	Gobierno, palacio de gobierno, miles de dólares, secretario, Pacheco
3	Pedro Castillo	0.389	12.937	Pedro Castillo, vacancia, carajo, vacancia carajo, gobierno
5	Este Castillo	0.365	23.212	Pueblo, Castillo, marcha, sábado, viene, sacará, este Castillo
7	Hoy nunca	0.348	10.774	Palacio, baño, secretario, corrupción, gobierno, presidente
6	Voz del pueblo	0.238	9.698	Voz, dólar, voz del pueblo, pueblo, gobernabilidad
1	Tanque del inodoro	0.180	6.335	Día, dicen, pueblo, miles, campesino, bueno, tanque del <u>baño</u>
2	Cometen acto	0.067	11.850	Gobierno, vacancia, Arequipa, Castillo, palacio, pueblo, fiscalía
8	Fiscalía encontró	0.036	10.670	País, Castillo, fiscalía, baño, Bruno Pacheco, Pacheco

Fuente: Elaborado por los autores con los resultados del estudio.

Discusión

Relevancia de los resultados obtenidos

En la Tabla 3 se puede observar que los tópicos más comentados en Twitter fueron los vinculados al tópico 4 (miles de dólares), tópico 3 (Castillo) y tópico 5 (este Castillo), pero también que los distintos tópicos se articulan en torno a dos ejes que son a la vez distintos y complementarios, tal como explicaremos a continuación.

El tópico 4 (miles de dólares) hace referencia específicamente al caso Bruno Pacheco, exsecretario general de Palacio de Gobierno, acusado de esconder 20 mil dólares en un baño de palacio y es investigado actualmente por lavado de activos. Se debe destacar que los términos top de esta etiqueta se repiten y se relacionan con los tópicos 7, 6, 1, 2 y 8, en los que también se aborda el caso Bruno Pacheco. Como se puede observar, los términos top de todas estas etiquetas son palabras que se asocian a: 1) el responsable del hecho denunciado, 2) la corrupción en el gobierno, 3) la responsabilidad y/o complicidad de Castillo en los hechos de corrupción. Ejemplos de estos *tweets* y *retweets* son los siguientes:

durante la incursión a palacio de gobierno se hallaron mil dólares en efectivo al secretario del presidente por qué no se ha pedido la detención preliminar de bruno pacheco querían pruebas aquí están
fiscalía encontró en el baño del despacho de bruno pacheco secretario del presidente pedro castillo
deberían preservar videos de las visitas que tuvo bruno pacheco antes de que castillo las desaparezca deben pedir a telefónica todos sus contactos esos verdes
que más pruebas quieren para sacar a pedro castillo del gobierno su asesor mano derecha bruno pacheco encontrado con las manos en la masa
bruno pacheco tenía escondidos en su baño cuánto tendrá pedro castillo en el suyo cuánto esconderá debajo de su sombrero

En cambio, los tópicos 3 (Castillo) y 5 (este Castillo) están referidos específicamente al pedido explícito de vacancia del presidente Pedro Castillo, y los términos top que aparecen en estas etiquetas se asocian con los tópicos 7, 2 y 8. Como se puede observar, los términos top de todas estas etiquetas son: 1) palabras orientadas a la acción (vacar), 2) consignas para participar en acciones de protesta (marcha por la vacancia), 3) adjetivos e insultos contra Castillo (incapaz, corrupto, narcoterrorista, mentiroso, ocioso, borrico, sinvergüenza). Ejemplos de estos *tweets* y *retweets* son los siguientes:

castillo se compró todos los boletos de la vacancia por incapacidad moral en defensa del orden constitucional y democrático debe vacarlo
todos a palacio fuera castillo y cerrón par de delincuentes terroristas lescano acuña miserables corruptos todo el esú somos la resistencia
es el pueblo quien sacará a castillo este sábado todos a la marcha me uno a la delegación que viene de piura
después de vacar a castillo tenemos que ir por la corrupta fiscalía
este sábado todos a la calle no te quedes en casa y defendamos al Perú vamos por la vacancia de pedro castillo nos vemos a partir de las en la av peruanidad de Jesús María
por ocioso y mentiroso por tráfico de influencias por sembrar el caos fuera pedro castillo y su pandilla de narcoterroristas
bien paso a paso es una primera batalla la siguiente será convencer y que el país está ante todo y el borrico castillo no debe estar un día más en palacio
de frente a palacio a sacar al sinvergüenza de castillo

Al comparar los resultados asociados en estos dos ejes, lo primero que destaca es el trato diferenciado a Pacheco y Castillo. Pese a que el tema del *hashtag* es el mismo (#VacanciaPedroCastillo) y que la mayoría de los comentarios u opiniones giran en torno al caso Bruno Pacheco, se les da un tratamiento distinto a estos dos actores en los discursos de Twitter. En tal sentido, pese a que el denunciado por un posible caso de corrupción es Bruno Pacheco, se constata que hay un discurso violento contra Castillo caracterizado por el uso de insultos y adjetivos personales. Es decir, las expresiones utilizadas contra Castillo en el *hashtag* no buscan promover un amplio debate en torno a ideas, propuestas o temas de agenda pública-política, sino más bien de “descalificar y agredir al otro.

También se constata que en el caso del *hashtag* #VacanciaPedroCastillo hay determinados encuadres o *frames* predominantes en los *tweets* y *retweets*. En este caso, hay un predominio del encuadre de responsabilidad individual o también conocido como “búsqueda del culpable”, un encuadre que individualiza las responsabilidades y busca la sanción o castigo del culpable.

Semejanzas y diferencias con estudios previos

El estudio establece semejanzas con otras investigaciones que utilizan el modelamiento de tópicos y el LDA para el análisis de otros temas políticos (electorales y de gestión pública) en Twitter, aunque no con relación a la vacancia presidencial, sobre lo cual no hemos encontrado estudios específicos. Con relación al uso de esta metodología en el tratamiento de otros temas políticos, hay una mayor afinidad con aquellos estudios vinculados a contextos electorales y poselectorales.

Aporte a la práctica profesional

Los resultados obtenidos nos permiten señalar varios niveles de aplicación. Por un lado, la metodología utilizada puede repetirse en otros estudios similares sobre temas políticos, pero también ampliarse con la combinación de otras técnicas cuantitativas o cualitativas, lo que permitiría una mayor profundidad en el análisis de los fenómenos sociales. Eso abre nuevas posibilidades de investigación de tipo inter y transdisciplinario.

Por otro lado, los resultados obtenidos permiten entender mejor la naturaleza y características del Twitter y el uso del *hashtag* en la discusión de temas polémicos en contextos de “alta polarización” y “fragmentación política”, así como sus posibilidades y límites para promover un debate público a partir de propuestas y/o generar consensos en torno a una agenda pública inclusiva. En esa línea, resulta de utilidad para el diseño de estrategias de comunicación pública-política, marketing, propaganda o incidencia política a través de las redes sociales.

Limitaciones del estudio

Como todo estudio, el nuestro tuvo algunas limitaciones, las cuales explicamos a continuación. En primer lugar, el período de análisis no fue muy amplio ya que el *hashtag* tuvo una corta duración. Una forma de superar esta limitación sería analizar el contenido de varios *hashtags* para ver si la pérdida de popularidad, frente a otro que se vuelve tendencia, aún permite hacer modelamiento de tópicos, o si la volatilidad de los *hashtags* es tan fuerte que solo permite análisis a corto plazo.

En segundo lugar, al no contar con estudios previos sobre cómo modelar la reacción ciudadana expresada en *tweets*, la elección del número de tópicos se basó en criterios estadísticos únicamente, y no en supuestos teóricos. Para futuras investigaciones, es recomendable que la elección del número de tópicos responda a un criterio teórico, por ejemplo, los tópicos extraídos al analizar los comentarios políticos en Facebook.

En tercer lugar, debido a las restricciones de tiempo, no se comparó la efectividad del modelamiento de tópicos en programas basados en R frente a otros programas basados en Python, porque es posible que la manera como se implementan los algoritmos para los programas en R y Python responda a criterios distintos en cada caso. Por ello, sería importante hacer esa comparación en estudios futuros.

Finalmente, pese a que la cantidad de información es significativa, el análisis de un solo *hashtag* presenta algunas limitaciones, en la medida que solo permite mirar el fenómeno desde un lado. Ello no permite compararlo con otras situaciones similares (ej. otros hashtags sobre el tema de la vacancia presidencial que han estado circulando en los últimos meses) ni tampoco analizar las diferencias en las posiciones de los usuarios de Twitter frente a un mismo tema. Por lo que se ha podido observar en los *tweets* y *retweets*, la gran mayoría de las opiniones están a favor de la vacancia. Es decir, no hay mayores diferencias en la posición, solo en lo que se dice y cómo se dice. A ello se debe agregar la circulación de noticias falsas o *fake news* y la presencia de troles en las redes sociales que diseminan información y opinión en los procesos electorales o en determinadas crisis políticas, lo que no permite garantizar que las opiniones que circulan en las redes provengan de ciudadanos que se expresan en forma libre y abierta, sino de personas contratadas que repiten determinadas consignas políticas prefabricadas. De allí la necesidad de que en próximos estudios se analicen y comparen opiniones diferentes frente a un mismo tema y se tengan en cuenta estos elementos.

Conclusiones

La identificación de dos ejes distintos y complementarios al mismo tiempo, que recorren y articulan los distintos tópicos, etiquetas y palabras top del estudio, evidencian que, pese a la alta fragmentación de los discursos políticos en Twitter, el uso del modelamiento de tópicos y del algoritmo LDA permite reconstruir “sentidos comunes” que orientan la opinión en los *tweets* y *retweets* analizados.

En tal sentido, la aplicación del modelamiento de tópicos y del algoritmo LDA al análisis del *hashtag* #VacanciaPedroCastillo ha permitido validar la utilidad de esta metodología. La información proporcionada en Métodos y Resultados posibilita que el enfoque pueda repetirse en otros estudios, enfocados en la dimensión política o cualquier otro tema de interés para la ciudadanía.

Si bien la aplicación del modelamiento de tópicos y del algoritmo LDA responde a un método y enfoque cuantitativo de la investigación, ello no significa que no pueda usarse en combinación con otros métodos y técnicas cuantitativas o cualitativas (ej. análisis del discurso, análisis de encuadres o *frames*). En tal sentido, se debe recordar que todos los métodos y técnicas de investigación tienen sus propias ventajas y desventajas, por lo que es recomendable usar una metodología mixta que permita una mejor comprensión de los fenómenos estudiados.

Referencias

- Alamsyah, A., Rizkika, W., Nugrobo, D. D. A., Renaldi, F., & Saadah, S. (2018). Dynamic large-scale data on Twitter using sentiment analysis and topic modeling. En: *6th International Conference on Information and Communication Technology* (pp. 254-258), Bandung, Indonesia. <https://doi.org/10.1109/IColCT.2018.8528776>.
- Alfaro, M. S. (2019) - *Cómo hacen campaña los candidatos a diputados - El estudio del vínculo entre candidato y votante en las elecciones de 2018 en El Salvador*. Memoria del Seminario de Difusión de Investigaciones 2019. Universidad Centroamericana “José Simeón Cañas”. San Salvador, El Salvador.
- Angulo, N.A., Osorio, E., Virgilio, A. & López, M.(2018). La política en Twitter. Un estudio comparativo de las estrategias discursivas de los candidatos finalistas a la Presidencia de Ecuador en 2017. *Comunica*, n°16. pp. 25-44. DOI: <http://dx.doi.org/10.6035/2174-0992.2018.16.3>

- Bayram A. B. & Ta, V. P. (2019). Diplomatic chameleons: Language style matching and agreement in international diplomatic negotiations. *Negotiation and Conflict Management Research*, 12(1), 23-40. <https://doi.org/10.1111/ncmr.12142>.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993-1022.
- Calheiros, A. C., Moro, S., & Rita, P. (2017). Sentiment classification of consumer-generated online reviews using topic modeling. *Journal of Hospitality Marketing & Management*, 26(7), 675-693. <https://doi.org/10.1080/19368623.2017.1310075>.
- Carreazo, D.I. (2020). Anatomía política de Twitter en Colombia Elecciones Presidenciales 2018 (Trabajo de investigación de Magíster), Universidad Nacional de Colombia. <https://repositorio.unal.edu.co/handle/unal/79257>
- Deltell, L., Claes, F. & Osteso, J. M. (2013). Predicción de tendencia política por Twitter – Elecciones Andaluzas 2012. En: Ámbitos, Revista Internacional de Comunicación N° 22. Grupo de Investigación en Estructura, Historia y Contenidos de la Comunicación, pp. 1-13. Universidad de Sevilla, España.
- Del Valle Cano, G. (2021). Detección de mensajes de odio en Twitter. Un estudio basado en perfiles dentro de la red social. Universidad Autónoma de Madrid. Escuela Politécnica Superior. Madrid, España.
- Fernández, P. (2018). Violencias de género en twitter: análisis desde el Trabajo Social. (Tesis de Doctorado), Universidad Complutense de Madrid Madrid, España.
- Gamboa, J.E (2019). Topic modeling en twitter: determinación de la agenda política peruana en el periodo de enero a setiembre del 2018. En: Anales Científicos 80(2): 308-327. Universidad Nacional Agraria La Molina. Lima, Perú.
- Harlow, S. (2015). Beyond representation: Exploring participation in and through technology at an alternative newspaper in El Salvador. *Cuadernos.info*, 37, 35-47. <https://doi.org/10.7764/cdi.37.822>.
- Hidayatullah, A. F., Pembrani, E. C., Kurniawan, W., Akbar, G., & Pranata, R. (2018). Twitter topic modeling on football news. En: *Proceedings of the 3rd International Conference on Computer and Communication Systems* (pp. 467-471). Nagoya, Japón: IEEE. <https://doi.org/10.1109/CCOMS.2018.8463231>.
- Ingle, M. & Emmanuel, M. (2016). Evaluations on sentiment analysis of micro blogging site using topic modeling. En: *International Conference on Signal Processing, Communication, Power and Embedded System* (pp. 1267-1272), Paralakhemundi, India. <https://doi.org/10.1109/SCOPES.2016.7955645>.
- Islam, T. (2019). *Ex-Twit: Explainable Twitter mining on health data*. Recuperado de [arXiv.org: 1906.02132](https://arxiv.org/abs/1906.02132).
- Jones, T. (2019). *textmineR: Functions for text mining and topic modeling*. Recuperado de <https://cloud.r-project.org/package=textmineR>.
- Josef, S. & Helena, C. (2019). Leadership and effectiveness: A content analysis of letters to shareholders during the financial crisis. *International Journal of Business Communication*, En prensa. <https://doi.org/10.1177/2329488418804044>.
- Kearney, M. W. (2019). *rtweet: Collecting Twitter data*. Recuperado de <https://cran.r-project.org/package=rtweet>.
- Kwon, K. H., Chadha, M., & Wang, F. (2019). Proximity and networked news public: Structural topic modeling of global Twitter conversations about the 2017 Quebec Mosque Shooting. *International Journal of Communication*, 13, 2652-2675.
- Manfredi, J. L., Amado-Suárez, A. & Waisbord, S. (2021). Twitter presidencial ante la COVID-19: Entre el populismo y la política pop. En: *Comunicar*, N° 66, v. XXIX, 2021 j Revista Científica de Educomunicación, pp. 83-94. España.
- Marín, P. P. & Díaz, A. (2016). Uso de Twitter por los partidos y candidatos políticos en las elecciones autonómicas de Madrid 2015. En: *Ámbitos, Revista Internacional de Comunicación*, N° 32, 2016, pp. 1-16. Universidad de Sevilla. Sevilla, España.
- Megala, S. K., Smitha, E. S., Mahalakshmi, G. S., Sendhilkimar, S. (2018). A word embedding model for topic recommendation. En: *Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies* (pp. 1307-1311), Coimbatore, India. <https://doi.org/10.1109/ICICCT.2018.8473055>.
- Pérez, P. V. (2020). Framing en Twitter - Análisis de los tweets enviados por parte de Lenin Moreno durante las protestas 2019. (Trabajo de Investigación Formativa previo a la obtención del Título de Licenciado en Comunicación Social) Universidad Casa Grande, Facultad de Comunicación Mónica Herrera. Guayaquil, Ecuador.
- Puranam, D., Narayan, V., & Kadiyali, V. (2017). The effect of calorie posting regulation on consumer opinion: A flexible Latent Dirichlet allocation model with informative priors. *Marketing Science*, 36(5), 726-746. <https://doi.org/10.1287/mksc.2017.1048>.
- Rodríguez Fidalgo, M. I., Ruiz Paz, Y., Paíno Ambrosio, A., & Jiménez Iglesias, L. (2017). Use of Facebook as narrative tool to articulate the new political dialogue. The case study of Vistalegre. *Revista Latina de Comunicación Social*, 72, 849-860. <https://doi.org/10.4185/RLCS-2017-1196-46en>.
- Ryoo, J. & Bendle, N. (2017). Understanding the social media strategies of U.S. primary candidates. *Journal of Political Marketing*, 16(3-4), 244-266. <https://doi.org/10.1080/15377587.2017.1338207>.

- Rubio García, Raquel (2014). Twitter y la teoría de la Agenda Setting: mensajes de la opinión pública digital. *Estudios sobre el Mensaje Periodístico*, Vol. 20, Núm. 1 (2014), pp. 249-264. Universidad Complutense. Madrid, España.
- Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using Latent Dirichlet allocation. *Electronic Commerce Research and Applications*, 29, 142-156. <https://doi.org/10.1016/j.elerap.2018.04.003>.
- Yinger, O. S. & Springer, D. G. (2019). Using psycholinguistic inquiry to measure emotional response to music: A feasibility study. *Psychology of Music*, 47(4), 606-614. <https://doi.org/10.1177/0305735618765292>.
- Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. En: *ECIR 2011 Advances in Information Retrieval* (pp. 338-349), Dublin, Irlanda. https://doi.org/10.1007/978-3-642-20161-5_34.

* **Carlos Vilchez-Román.**

Investigador Concytec, docente de Centrum Católica Graduate Business School de la PUCP. Magíster en Comunicación Social con mención en Investigación en Comunicación.
Correo electrónico: cvilchez@pucp.edu.pe

* **Fernando Palomino Villanueva.**

Especialista en Comunicación para el Desarrollo, consultor de la Asociación de Comunicadores Sociales Calandria.
Correo electrónico: luisp143@yahoo.es

* **Prisea Georgina Vilchez Samanez.**

Jefe de Capacitación y de Gestión de Proyectos del Instituto de Investigación de la Universidad Jaime Bausate y Meza. Magíster en Comunicación y Doctoranda en Sociología.
Correo electrónico: pvilchez@bausate.edu.pe