Anales Científicos 80(2): 308-327 (2019)

Análes Científicos

ISSN 2519-7398 (Versión electrónica)
Website: http://revistas.lamolina.edu.pe/index.php/acu/index

Topic modeling en twitter: determinación de la agenda política peruana en el periodo de enero a setiembre del 2018

Topic modeling on twitter: determination of the peruvian political agenda in the period from January to September 2018

Jesús Eduardo Gamboa Unsihuay[1*]

[1] Universidad Nacional Agraria La Molina, Lima, Perú. Email: jgamboa@lamolina.edu.pe

Recepción: 14/01/2018; Aceptación: 05/06/2019

Resumen

El objetivo de esta investigación fue determinar los temas abordados por los distintos grupos de la clase política peruana 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, haciendo uso de la técnica de minería conocida como modelo de temas (Topic Modeling) y el modelo de asignación latente de Dirichlet. Se encontró que tres cuartas partes de los contenidos textuales se refieren a la gestión del Poder Ejecutivo y Legislativo, las actividades partidarias de Fuerza Popular y los grupos parlamentarios de izquierda, la corrupción judicial, las funciones de representación parlamentaria y eventos que sucedieron durante los meses de verano; asimismo, se encontraron diferencias en los temas de mayor divulgación entre los grupos políticos. Luego de contrastar dichos hallazgos con los acontecimientos ocurridos en la realidad, se concluyó que la metodología propuesta permite efectivamente detectar los tópicos de la agenda política a partir de un gran volumen de textos.

Palabras clave: minería de texto; segmentación; modelamiento de temas; Twitter; agenda política; Perú.

Abstract

The objective of this research was to determine the topics addressed by the different groups of the Peruvian political class, through the analysis of the content shared by its members on their Twitter accounts in the period from January to September 2018 making use of the mining technique known as Topic Modeling and the Dirichlet's Allocation Latent model. The research found that three quarters of the textual content refers to the management of the Executive and Legislative Power, the party activities of Fuerza Popular and the parliamentary

Forma de citar el artículo: Gamboa, J. 2019. Topic modeling en twitter: determinación de la agenda política peruana en el periodo de enero a setiembre del 2018. Anales Científicos 80 (2):308 -327 (2019).

DOI: http://dx.doi.org/10.21704/ac.v80i2.1446
Autor de correspondencia (*): Gamboa, J. Email: jgamboa@lamolina.edu.pe

groups of the left, judiciary corruption, the parliamentary representation functions and events that took place during the summer season (January and February). In addition, there are differences in the most discussed topics among political groups. After contrasting these findings with the real events, it was determined that the proposed methodology allows to effectively finding the topics of the political agenda from a large volume of texts.

Keywords: text mining; clustering; topic modeling Twitter; political agenda; Perú.

1. Introducción

La minería de textos comprende el uso de modelos y algoritmos que permiten extraer el conocimiento implícito de datos textuales. Dichos patrones pueden referirse al comportamiento (qué hacen o sobre qué temas se manifiestan) o sentimiento (qué sienten u opinan) de los autores (Mateo, 2016). Por otro lado, la información en formato textual, especialmente aquella que es compartida en redes sociales, viene creciendo rápidamente en los últimos años. En particular, Twitter es un servicio de *microblogging* (envío de mensajes cortos) cuyo alcance e impacto es mucho mayor que los medios de comunicación tradicionales (Farías, 2017).

En lo que respecta al uso de datos textuales para análisis político, Grimmer (2009) emplea el "modelo de Agenda Expresa" con la finalidad de identificar los tópicos expuestos por senadores estadounidenses a partir de sus comunicados de prensa, mientras que Yano *et al*. (2009) aplican modelo de temas en datos extraídos de blogs políticos. Montesinos (2014) utiliza análisis de sentimientos en datos de Twitter en el contexto de elecciones presidenciales en Chile. De manera similar, Pla & Hurtado (2014) implementan la misma técnica a fin de identificar la preferencia política de usuarios de Twitter. Fang *et al*. (2015) realiza un procedimiento semejante, pero en usuarios escoceses. En Latinoamérica, Alvarado *et al*. (2016) aplica análisis de sentimiento en datos de Twitter durante la campaña política por la alcaldía de Bogotá. Uno de los trabajos más recientes en el área política corresponde a Greene & Cross (2017), quienes hacen uso del modelamiento de temas de manera dinámica con el propósito de determinar el contenido de los discursos plenarios de los parlamentarios europeos durante el periodo de 1999 a 2014. En Perú, Sigueñas (2016) presentó una conferencia acerca de Topic Modeling aplicado a discursos presidenciales. Más aplicaciones locales de minería de texto pueden ser encontradas en el artículo de Linares *et al*. (2015) quienes exponen un caso de análisis de sentimientos basado en datos de Twitter, con el propósito de estudiar los deseos de los turistas por visitar Perú. Por otro lado, Vílchez & Alhuay (2016) usan la minería de textos para comprender, a través de cuestionarios abiertos, cómo los estudiantes y de bibliotecología perciben la formación que se les brinda, mientras que Cárdenas *et al*. (2015, 2018) presentan una aplicación de Topic Modeling en el área industrial.

Entretanto, el ambiente político peruano del año 2018 viene siendo más convulso que el de los años anteriores, constituyéndose en una crisis política a causa de la corrupción (Vollenweider, 2018), la cual se ha evidenciado en los escándalos por el caso Odebrecht (Diario Gestión, 2018), los pedidos de vacancia (Diario El Comercio, 2017; Diario Correo , 2018) y la posterior renuncia de Pedro Pablo Kuczynski a la presidencia de la República (Diario La República, 2018b), la revelación de videos en los que se negociaban votos de congresistas (Diario El Comercio, 2018a) y la difusión de audios que dejaban al descubierto la corrupción en el Consejo

Nacional de la Magistratura (Diario El Comercio, 2018c). Como consecuencia de este entorno político complejo, el contenido de texto disponible se incrementa y su lectura se hace una labor difícil, ya que además la población viene perdiendo la confianza e interés en la política (Diario Perú21, 2018c). Por este motivo, es trascendente el uso de herramientas analíticas de texto, tales como Topic Modeling, que permitan resumir el contenido textual y así conocer la agenda política de nuestros representantes. Así, el objetivo de esta investigación es extraer los principales temas abordados por los grupos políticos a partir de los textos compartidos por sus principales representantes, en sus respectivas cuentas de Twitter, en el periodo de enero a setiembre del 2018.

2. Materiales y métodos

Materiales empleados

Para el desarrollo de la presente investigación se hizo uso de una computadora portátil con procesador Intel ® Core ™ i7-7500U con 8 GB de memoria RAM. En este ordenador se trabajó con el software R en su versión 3.5.1 y la interfaz RStudio versión 1.1.456. Además, fue necesario contar con una aplicación en una cuenta de Twitter, cuyas credenciales sirvieron para realizar la conexión entre el software (R) y la API de Twitter.

Metodología

Con la finalidad de alcanzar el objetivo del estudio, se llevó a cabo las siguientes tareas: extracción de documentos, limpieza y estructuración de datos, modelamiento de temas mediante la asignación latente de Dirichlet, selección del número de temas y su interpretación.

a) Extracción de documentos

Un documento es una secuencia de *tokens*, los cuales, a su vez, son una secuencia ininterrumpida de caracteres, siendo una palabra un ejemplo representativo de *token*. Para la investigación se consideró como documento al conjunto de las publicaciones realizadas en Twitter, durante cada semana, por cada uno de los 141 políticos pertenecientes al Poder Ejecutivo (EJE) o que son integrantes de uno de los siguientes grupos parlamentarios (en orden alfabético): Acción Popular (AP), Alianza por el Progreso (APP), Célula Parlamentaria Aprista (APRA), Frente Amplio por Justicia, Vida y Libertad (FA), Fuerza Popular (FP), No agrupados (NAG), Nuevo Perú (NP) y Peruanos por el Kambio (PPK). Así, tomando en cuenta las primeras 39 semanas del año 2018 y los 9 grupos políticos, se tuvo un total de 351 documentos, los cuales conformaron el corpus de la investigación.

La extracción de los tuits se realizó en el *software* R, haciendo uso del paquete rtweet, mediante el cual se realizó la lectura de las credenciales de acceso a la aplicación en Twitter y se accedió al contenido compartido de manera pública por los 141 políticos.

b) Limpieza y estructuración de datos

Los datos textuales son considerados en la categoría de datos no estructurados ya que están almacenados en un formato que no es adecuado para su análisis, sin embargo, es posible su limpieza, eliminando y convirtiendo ciertos caracteres que pueden ser considerados como ruido en el análisis. Entre las tareas de limpieza, que fueron ejecutadas en R por los paquetes tm y topicmodels, se tuvo lo siguiente:

- Remoción de signos de puntuación tales como los puntos (.), las comas (,), los signos de exclamación (¡!), los dos puntos (:), las comillas (“ ”), etc. Así también caracteres numéricos, tildes, enlaces web, marcadores HTML, entre otros.

- Retiro de palabras que no aportan al significado del texto, conocidas como *stopwords* o palabras vacías, por ejemplo, preposiciones (a, de, en, para, por, sin, …), artículos (el, la, un, …), conjunciones (aunque, luego, ni, que, pero, y, …).
- Conversión de mayúsculas en minúsculas y uniformización del espacio entre palabras o *tokens*.

La tarea posterior a la limpieza de textos es su estructuración, la cual consiste en obtener una matriz de términos de documento. Esta matriz contiene las frecuencias de aparición de las palabras del vocabulario en cada documento y fue obtenida utilizando el paquete tm del software R.

c) Modelamiento

La Asignación Latente de Dirichlet (ALD), propuesta por Blei *et al.* (2003), es un modelo probabilístico para conjuntos de datos discretos tales como los documentos de texto. Mediante este modelo se asume que un documento es generado por una mixtura de tópicos y que cada uno de estos es construido en base a una mixtura de palabras. Además, de acuerdo con Heinrich (2008), el aprendizaje se realiza de manera no supervisada ya que los temas no son conocidos de antemano, por ello se dice que la asignación de estos temas es latente y su comportamiento probabilístico (mixtura) es explicado por una distribución Dirichlet cuyos parámetros son estimados por el modelo. A diferencia de una técnica tradicional de segmentación, la ALD no restringe la asociación de un documento a un único tema en particular.

El modelo generador de documentos funciona de la siguiente manera:

1. Se asume que el corpus está compuesto por D documentos, cada uno de los cuales es generado por una mixtura de k temas, es decir $\boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha})$, d = 1, ..., D, siendo $\alpha = \alpha_1 \ldots \alpha_K$ un hiperparámetro. Griffiths y Steyvers (2004), así como Grün y Hornik (2011), sugieren el uso de α_k = 50/k como valor inicial, el cual es establecido por defecto en R, mientras que Grimmer (2009) propone el modelamiento jerárquico, asignando la distribución Gamma a cada valor de α. Se puede señalar que $\Theta = (\theta_1, \theta_1 \ldots \theta_D)$ es la matriz de dimensión $K\,X\,D$ que contiene las probabilidades de que el k-ésimo tema genere del d-ésimo documento. Para la aplicación en datos de Twitter se consideró que la cantidad de documentos es D= 351 y que el número de temas es K=20, según los criterios de selección del número de temas que serán presentados en la siguiente sección.
2. Luego, la n-ésima palabra en el d-ésimo documento ($W_{d,n}$) debe ser generada a partir de un tema ($Z_{d,n}$), el cual se muestrea a partir de una distribución multinomial con parámetro θ_d la mixtura del paso previo mediante $z_{d,n} \sim Mult(\boldsymbol{\theta}_d), \quad d = 1,\ldots,D; n = 1,\ldots,N_d$
3. Las palabras que explican cada tema vienen dadas por la mixtura $\boldsymbol{\phi}_k \sim Dir(\boldsymbol{\beta}_k)$ k = 1, ...; K donde $\beta_K = (\beta_{1K} \ldots \beta_{VK})$ es el hiperparámetro, al que se le asigna el valor de $\beta_{\omega,K} = 0.1$ según recomendación de Griffiths & Steyvers (2004) y de Grün & Hornik (2011). Por lo tanto, la matriz $\Phi = (\phi_1, \phi_2, \ldots \phi_K)$, de dimensión $K\,X\,V$, muestra la probabilidad de que la w-ésima palabra sirva para explicar el k-ésimo tema.
4. El último paso consiste en la generación de la n-ésima palabra del d-ésimo documento dado que ya se generó el tema (paso 2) y su mixtura de palabras (paso 3). Así, se tiene que $w_{d,n} \sim Mult\left(\boldsymbol{\phi}_k \mid z_{d,n} = k\right) \quad d = 1,\ldots,D; \; n = 1,\ldots,N_d$

Al utilizar este algoritmo, el documento es generado bajo el supuesto de la ´bolsa de palabras´, el cual señala que el orden de las palabras no aporta mayor información al análisis. No obstante, a diferencia de la generación de documentos en la que a partir de los tópicos se originan los términos $W_{d,n}$ que componen los textos (corpus), en la determinación de temas se da el proceso inverso.

En efecto, esta determinación de temas se realiza mediante inferencia bayesiana, la cual consiste en la obtención de la distribución a posteriori de las cantidades de interés, es decir $f(\Theta, \Phi, \mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ la cual no puede ser obtenida analíticamente, sino que debe aproximarse mediante métodos computacionales como el algoritmo del muestreador de Gibbs, el cual fue empleado en esta investigación. Este algoritmo requiere la especificación de las distribuciones condicionales completas, sin embargo no es necesario incluir los parámetros Θ y Φ en este paso ya que su distribución a posteriori puede ser derivada a partir de $f(\mathbf{z} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Heinrich (2008) propone que el tema asignado para una palabra depende de la asignación de los temas de las demás palabras y del vocabulario. De esta manera, determina que la distribución condicional completa para la n-ésima palabra del d-ésimo documento es:

$$f(z_i = k \mid \mathbf{z}_{-i}, w_i = t, \mathbf{w}_{-i}) = \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,-i}^{(t)} + \beta_t} \times \frac{n_{d,-i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_d^{(k)} + \alpha_k - 1}$$

considerando $i = \{d, n\}$ a fin de simplificar la expresión, además $n_{k,-i}^{(t)}$ representa el número de veces que la palabra t ha sido utilizada para explicar el k-ésimo tema, excluyendo la i-ésima palabra y $n_{d,-i}^{(k)}$, el número de veces que el tema k ha sido observado a través de una palabra en el d-ésimo documento, también excluyendo la i-ésima palabra.

Luego, la distribución de los parámetros multinomiales Θ y Φ es obtenida haciendo uso del teorema de Bayes. En primer lugar, la mixtura de temas por documento resulta:

$$(\boldsymbol{\theta}_d \mid \mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \sim Dir(\mathbf{n}_d + \boldsymbol{\alpha}) \Rightarrow \boldsymbol{\theta}_{d,k} = \frac{n_d^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_d^{(k)} + \alpha_k)}$$

donde n_d es el vector de frecuencias de las palabras que se refieren a cada uno de los K temas en el d-ésimo documento y $\theta_{d,k}$ es la probabilidad de que el d-ésimo documento contenga el k-ésimo tema. Luego, para la mixtura de palabras por tema:

$$(\boldsymbol{\phi}_k \mid \mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \sim Dir(\mathbf{n}_k + \boldsymbol{\beta}) \Rightarrow \boldsymbol{\phi}_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} (n_k^{(t)} + \beta_t)}$$

donde n_k es el vector de frecuencias de cada palabra del vocabulario que se refiere al k-ésimo tema y $\phi_{k,t}$ es la probabilidad de que el k-ésimo tema sea explicado por la t-ésima palabra. Todo el proceso de inferencia detallado fue ejecutado en R mediante el paquete topicmodels.

d) Selección del número de temas

Griffiths & Steyvers (2004) utilizan el concepto de selección bayesiana de modelos, aproximando la verosimilitud marginal $f(\mathrm{w}|k)$ mediante la estimación de la media armónica de un conjunto de valores $f(\mathrm{w}|\mathrm{z},K\ k)$ cuando z es obtenido de la distribución posterior $f(\mathrm{z}|\mathrm{z},K\ k)$ mediante el muestreador de Gibbs. Luego, plantea elegir el valor de K que maximiza dicha verosimilitud marginal. Por otro lado, Cao *et al.* (2009) propone un indicador basado en el promedio de las

correlaciones en pares de los temas, debiendo ser elegido el valor de K que minimiza dicha correlación promedio. Arun *et al.* (2010) considera al modelo de ALD como un método en el que la matriz de términos de documentos se factoriza en una matriz M_1 de dimensión K X V y otra M_2 de dimensión D X K. El indicador sugerido por Arun se basa en la divergencia simétrica de Kullback-Leibler de las distribuciones de valor singular de M_1 y L_{1XD} donde L_{1XD} es un vector que contiene el número de términos de cada documento del corpus. Este indicador se minimiza para elegir el número de temas K. Finalmente, Deveaud *et al.* (2014) proponen estimar el indicador de divergencia de información entre todos los pares de temas, el cual debe ser maximizado. Estos cuatro indicadores vienen implementados en el *software* R, en el paquete ldatuning.

e) Interpretación de los temas

En principio, la interpretación de cada tema se define en función a la mixtura de palabras que lo compone. Grimmer (2009) etiqueta los tópicos mediante una verificación manual de un grupo aleatorio de documentos con alta probabilidad de contener cada tema, también propone usar documentos asociados a cada tópico a través del tiempo y contrastarlos con los acontecimientos sucedidos en la realidad. El paquete LDAvis fue una de las herramientas útiles para la interpretación de temas.

Este paquete permitió obtener una gráfica en la que se visualizaron los temas y sus mixturas, así como un indicador λ cuyo valor cercano a cero permitió señalar las palabras de gran exclusividad en el tema en análisis, es decir que su probabilidad de ser parte de una mixtura que explica otro tema es cercana o igual a cero, mientras que lo opuesto, valores de λ cercanos a uno señalaron términos bastante frecuentes, pero que no eran exclusivos del tema en cuestión. Utilizar ambos valores, así como el valor λ = 0,6, permitió interpretar y dar nombre a cada tema.

2. Resultados y discusión

Limpieza de textos

Durante el procedimiento de limpieza, el texto original se convierte en una lista de *tokens* (en minúsculas, sin tildes, números, caracteres especiales, URL, ni signos de puntuación) tal como se muestra en la Tabla 1.

Tabla 1. Ejemplo de limpieza de texto

Texto sin limpiar	Texto limpio
De los 7335 millones que grandes empresas deben por impuestos, más de 6000 millones debe #Telefónica (#Movistar). Sinvergüenzas https://t.co/rOJN0ShkJN	millones grandes empresas deben impuestos millones debe telefonica movistar sinverguenzas

Resultados descriptivos

Una vez que los tuits están limpios, un primer resultado corresponde a la nube de palabras. En las Figuras 1 y 2 se muestran estos resultados para los meses de enero y setiembre de 2018. En estas nubes, se observa que las palabras más frecuentes varían según el contexto: en enero resaltan las palabras referidas a las reacciones frente al indulto de Alberto Fujimori (Fowks, 2017) y la visita del papa Francisco (Vilcachagua, 2018), mientras que en setiembre son notorios los términos relacionados con las reformas políticas (Diario El Comercio, 2018d), sin embargo, aquellas palabras concernientes al Congreso de la República (congreso, ley, comisión) se mantienen en ambos meses.

Figura 1. Nube de palabras para enero de 2018

A partir de las nubes, se puede extraer una lista de personajes más nombrados, es decir aquellos que han sido mencionados al menos 200 veces durante cada mes, los cuales aparecen, junto a su(s) *token*(s), señalados en la Tabla 2.

También es posible indicar los grupos políticos y los usuarios más activos en Twitter para el periodo de estudio, lo cual se muestra en las Figuras 3 y 4, respectivamente. Fuerza Popular, la bancada mayoritaria del Congreso de la República, ha compartido en promedio 1134 tuits semanales, lo que equivale a 162 tuits diarios, siendo las congresistas Úrsula Letona y Lourdes Alcorta, ambas de Fuerza Popular, las más activas en Twitter, con un promedio de 12 tuits publicados por día.

Figura 2. Nube de palabras para setiembre de 2018

Tabla 2. Personalidades más mencionadas por políticos peruanos en sus cuentas de Twitter

Personalidad	*Token*	E	F	M	A	M	J	J	A	S
Papa Francisco	papa	✓								
Alberto Fujimori	fujimori	✓								
Pedro Pablo Kuczynski	ppk, ppkamigo	✓	✓	✓						
Jorge Barata	barata		✓							
Keiko Fujimori	keikofujimori			✓						
Martín Vizcarra	vizcarra, martinvizcarrac			✓	✓			✓	✓	✓
César Villanueva	cesarvperu				✓	✓				
Pedro Chávarry	chavarry							✓	✓	✓
César Hinostroza	hinostroza							✓	✓	✓

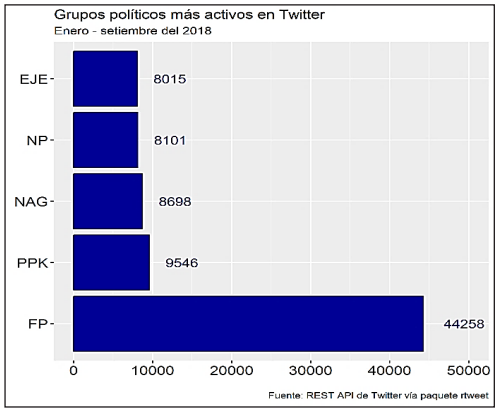

Figura 3. Grupos políticos más activos en Twitter

Estructuración de datos y modelamiento

Para iniciar con el modelamiento se requiere estructurar los datos. Así, la matriz de términos de documentos es una tabla de contingencia de 351 filas (documentos) y 72 250 columnas (vocabulario). Un extracto de esta matriz, referido a algunos términos en los documentos de la semana 2, se muestra en la Tabla 3.

A partir de este punto, la data se encuentra estructurada y lista para el análisis. Se aplicó el modelo de Asignación Latente de Dirichlet considerando los siguientes parámetros de control para el algoritmo del muestreador de Gibbs: un total de 200 000 iteraciones de las cuales "se queman" las primeras 1000 (*burn in*) y de las restantes, se conservan solo 40 000 a efectos de disminuir la autocorrelación entre dos iteraciones consecutivas. Además, los valores iniciales par a α y β fueron los propuestos por Grün & Hornik (2011). Por otro lado, los indicadores de Griffiths (2004), Cao (2009) y Arun (2010) señalan que es razonable considerar que el corpus está construido en base a 20 temas.

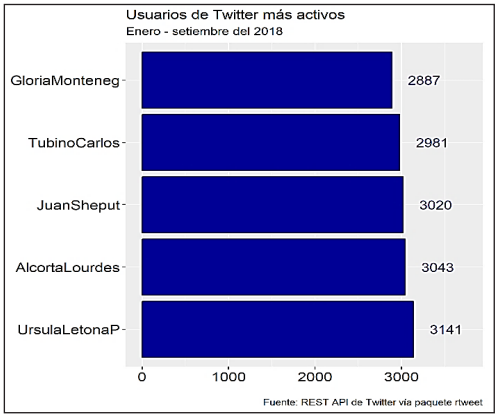

Figura 4. Políticos más activos en Twitter

Tabla 3. Extracto de la matriz de términos de documentos

	Términos						
Documento	igual	imponer	impuestos	independientes	indignante	indulto	infame
AP.2	1	0	0	0	0	1	0
APP.2	0	0	0	0	0	6	0
APRA.2	0	0	0	0	0	10	0
EJE.2	2	0	2	0	0	0	0
FA.2	0	0	0	0	0	4	2
FP.2	5	1	3	1	2	17	0
NAG.2	2	0	0	0	1	13	0
NP.2	2	0	0	0	0	57	3
PPK.2	0	0	0	2	0	2	0

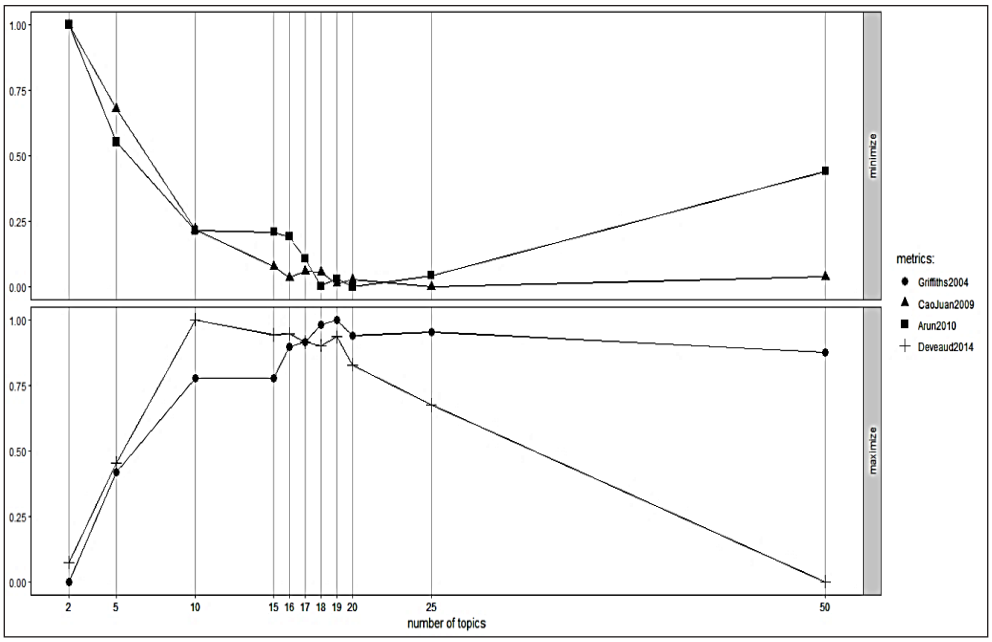

Figura 5. Indicadores para selección del número de temas

Interpretación de temas mediante mixtura de palabras y verificación de documentos

En las Tablas 4 y 5 se aprecia las palabras que caracterizan los temas para $\lambda = 0$ y $\lambda = 1$, respectivamente, ordenados de mayor a menor frecuencia de aparición. Ambas Tablas pueden ser usadas para dar nombre a los temas. Así, al primer tema, que concentra casi la mitad de los contenidos, se le dará el nombre de ´Labor del Poder Ejecutivo y Legislativo´ ya que es explicado por palabras como congreso, ley (una de las principales funciones del Congreso de la República es la aprobación de leyes), presidente y país (el presidente representa los intereses del país).

Luego, el tema 2 recibió el nombre de ´Actividades de la bancada Fuerza Popular´, dados los *tokens* asociados al tema tales como bankadafp y milagrostakayama. De manera similar, en el tema 3 resaltan los nombres de las bancadas Nuevo Perú y Frente Amplio, así como de sus integrantes, motivo por el cual se le denominó ´Actividades de bancadas de izquierda´. Por otro lado, el tema 4 trata acerca de los audios y procedimientos irregulares en el Congreso Nacional de la Magistratura (CNM), por lo que se le denominó ´Corrupción judicial´. A manera de ejemplo, un extracto del documento 251, en el que se mencionan diversas palabras asociadas al tema de la ´Corrupción judicial´, es mostrado en la Figura 6; además, los 30 términos más relevantes que componen la mixtura de este tema se muestran en el panel derecho de la Figura 7.

Enseguida, el tema 5 hace referencia a eventos que sucedieron durante el verano: los pedidos de vacancia presidencial, el caso Odebrecht, las reacciones frente al indulto de Alberto Fujimori, el accidente de un bus ocurrido en Pasamayo (Diario Perú21, 2018a) y la huelga de agricultores de papa (Diario La República, 2018a). Hasta este punto, los cinco primeros temas explican el 70,8% del corpus. La lista completa de temas se detalla en la Tabla 6.

Tabla 4. Términos de caracterización exclusiva (λ = 0)

tema	términos
1	asi, hace, bien, mejor, importante
2	descentralizacion, bankadafp, ferrenafe, milagrostakayama, grtujillo
3	nuevoperu, bancadafaperu, mov, richardarceperu, cdpueblos
4	becerril, reformajudicial, recuperarelperu, senorak, remover
5	pasamayo, keikoestalimpia, carnaval, fusiones, serpentin
6	familiamidis, education, lilarosahu, amp, school
7	terepresenta, fpentodoelperu, semanaderepresentacion, inopinada, verificando
8	cooperativas, sbs, castración, química, noallavado
9	cuestiondeconfianza, canzio, reformaya, composición, hermanitosnuncamás
10	yoshiyama, diainternacionaldelamujer, kenjivideos, anosdefuerzapopular, uif
11	robertovieirap, nicolaslucar, snp, econterno, amariateguiblog
12	cumbreperu, chavin, huantar, comandos, síria, chavindehuantar
13	salaverry, patrias, mesadirectiva, fiestaspatrias, aurelio
14	jibaja, comonotevoyaquerer, felizdiadelcampesino, juntos, cosechando
15	sheput, salvadorheresi, penitenciario, homofobia, bancadappk
16	cavassa, cupula, murodelima, aironnelson, servidora
17	unidosporlavida, votodeconfianza, conmishijosnotemetas, puentes, marchaporlavida
18	eyvi, juanita, detenerse, diadelmaestro, simulacro
19	pontifex, franciscoenperu, unidosporlaesperanza, polo, modopapa
20	siempreadelante, dmorazevallos, nadietelodicenosotrostelodecimos, laeducserespeta, gorjeda

Tabla 5. Términos recurrentes (λ = 1)

tema	términos
1	pais, congreso, ser, ley, presidente
2	comision, congresoperu, sesion, pleno, bankadafp
3	mujeres, nuevoperu, violencia, ley, congresoperu
4	cnm, justicia, chavarry, hinostroza, fiscal
5	ppk, barata, odebrecht, pasamayo, vacancia
6	familiamidis, midis, ministra, educacion, education
7	terepresenta, fpentodoelperu, semanaderepresentacion, distrito, obra
8	cooperativas, sbs, supervisión, pleno, ahorro
9	constitución, reforma, congreso, cnm, reformas
10	ppk, vacancia, presidente, ppkamigo, renuncia
11	robertovieirap, rppnoticias, exitosape, canaln, nicolaslucar
12	lista, menores, teestamosbuscando, completa, conoce
13	directiva, mesa, periodo, vizcarra, referendum
14	rusia, medios, gracias, publicidad, mundial
15	juansheput, sheput, bruce, juan, carlos
16	vizcarra, cavassa, ppk, fiscal, chavarry
17	cesarvperu, unidosporlavida, lima, votodeconfianza, conmishijosnotemetas
18	violencia, eyvi, agreda, feminicidio, maestros
19	papa, francisco, esperanza, pontifex, franciscoenperu
20	siempreadelante, universidades, reconstruccionconcambios, piura, mtc

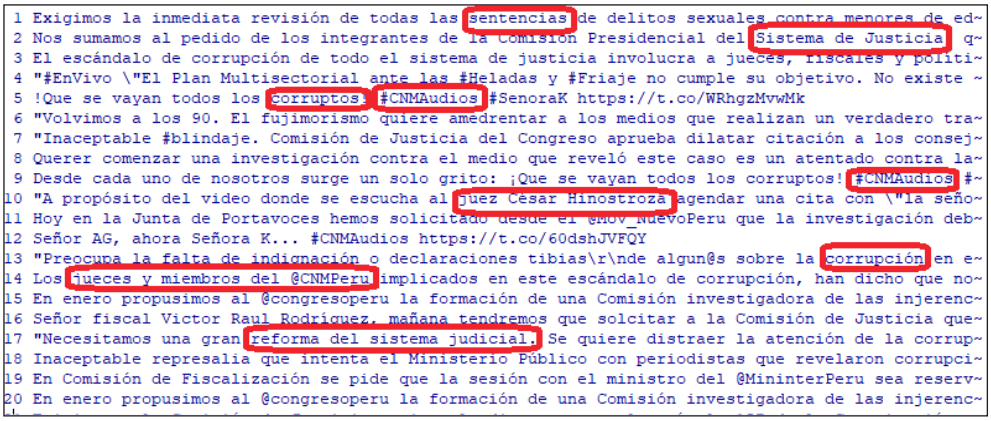

Figura 6. Extracto del documento 251

En la Figura 7 se puede visibilizar, además de la mixtura de palabras que explica el tema 4, la relación existente entre los temas, formándose ciertas agrupaciones. Los contenidos de la ´Labor del Poder Ejecutivo y Legislativo´ (tema 1) guardan relación con los tuits acerca de los temas 2 y 3: ´Actividades de la bancada de Fuerza Popular´ y ´Actividades de las bancadas de izquierda´, respectivamente. Esta asociación natural se da ya que el Poder Ejecutivo, la bancada fujimorista y las de izquierda son los principales actores del debate político durante el periodo de estudio.

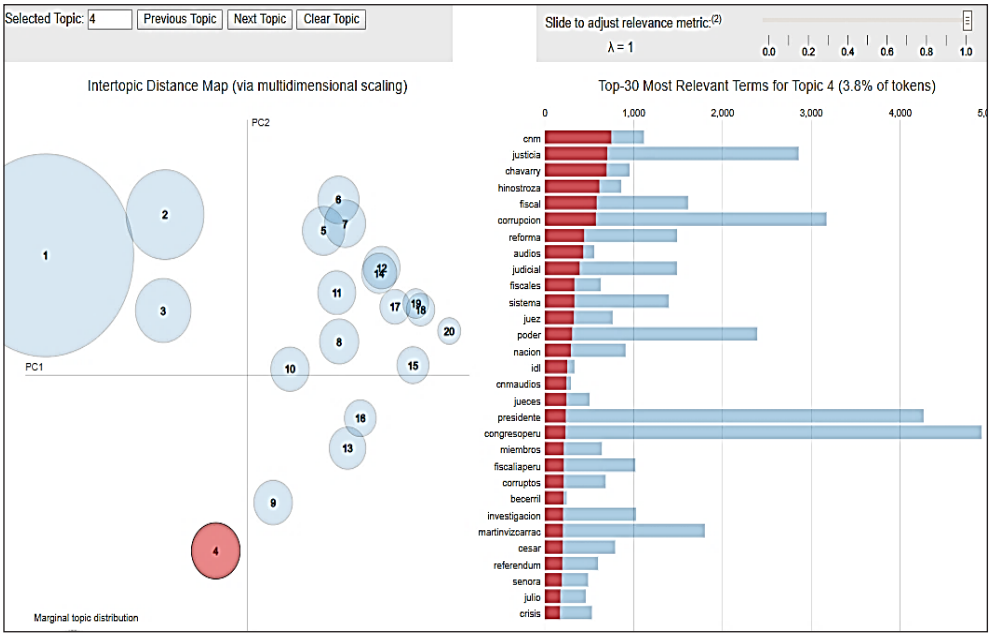

Figura 7. Relación entre temas (izquierda) y mixtura de términos para el tema 4 (derecha)

Tabla 6. Temas encontrados

N°	Nombre	Porcentaje
1	Labor del Poder Ejecutivo y Legislativo	49,5 %
2	Actividades de la bancada Fuerza Popular	9,7 %
3	Actividades de las bancadas de izquierda	4,9 %
4	Corrupción judicial	3,8 %
5	Eventos de verano	2,9 %
6	Inclusión social	2,7 %
7	Actividades de representación parlamentaria	2,7 %
8	Supervisión de cooperativas	2,5 %
9	Cuestión de confianza	2,4 %
10	Vacancia presidencial	2,4 %
11	Contenido periodístico	2,3 %
12	Búsqueda de personas desaparecidas y terrorismo	2,2 %
13	Referéndum	2,1 %
14	Mundial Rusia 2018 y ley de publicidad estatal	2,0 %
15	Actividades de bancada Ppk	1,6 %
16	Relación de José Cavassa con el partido Ppk	1,6 %
17	Actividades del Primer Ministro y polémica por enfoque de género	1,5 %
18	Violencia contra la mujer	1,3 %
19	Visita del papa Francisco	1,1 %
20	Educación superior	0,9 %

Un segundo grupo de temas está conformado por el tema 4 (´Corrupción judicial´) y 9 (´Cuestión de confianza´) ya que el último puede considerarse como consecuencia directa del primero (Diario El Comercio, 2018e). Luego, los temas 6 (´Inclusión social´) y 7 (´Representación parlamentaria´) se centran en actividades específicas que se dan generalmente fuera de la capital, por lo que resultan muy poco distanciadas; asimismo, los acontecimientos sucedidos en verano (tema 5), como el accidente de bus en Pasamayo o la huelga de agricultores de papa.

Un siguiente grupo se refiere a los actos de corrupción ligados a personalidades del Poder Ejecutivo y el partido de gobierno (PpK). Así, el tema 13 (´Referéndum´) es una iniciativa contra la corrupción (Ramos, 2018), mientras que el tópico 16 (´Cavassa y el partido PpK´) es una acusación, por parte de la oposición, de que José Luis Cavassa habría trabajado para el partido PPK en campaña de 2016, según informó La República (2018d).

Interpretación de temas mediante verificación temporal

Si bien el primer tópico domina casi el 50% de los contenidos, cuando se realiza el análisis semanalmente, los porcentajes referidos a los demás temas van variando. En la Figura 8 se detalla la distribución semanal de temas para el primer trimestre de 2018 en la que destacan tres coincidencias

temporales con lo acontecido en la realidad:

a) La visita del papa Francisco a Perú durante la tercera semana del año.

b) Los acontecimientos que sucedieron en verano (accidente de bus en Pasamayo, huelga de agricultores, etc.), los cuales van perdiendo terreno en cuanto a contenido textual a medida que avanzan las semanas.

c) La renuncia de Pedro Pablo Kuczynski en la semana 12, la cual vino antecedida de algunas semanas por los pedidos de vacancia presidencial

A continuación, en la Figura 9, se representan los temas para el segundo trimestre del año 2018. De manera similar al primer trimestre, se observan las coincidencias temporales:

a) El arresto domiciliario de Osmán Morote (Diario El Comercio, 2018b), ordenado por el Poder Judicial durante la semana 13 y hecho efectivo en la semana 16, lo cual conlleva a un mayor contenido textual referente al tema ´Terrorismo y búsqueda de personas desparecidas´, entre esas semanas.

b) La ´Marcha por la Vida´ realizada en la semana 18 (Diario Perú21, 2018b) origina que el tema de ´Polémica por temas de enfoque e ideología de género´ fuera resaltante durante esa semana y la inmediata posterior.

c) El fallecimiento de Eyvi Ágreda (Diario La República, 2018c), ocurrido en la semana 22, se refleja en el tema ´Violencia contra la mujer"´

d) El debate por la supervisión de cooperativas de ahorro y crédito por parte de la Superintendencia de Banca y Seguros (2018), cuya aprobación se dio finalmente en la semana 24, está asociado al tema ´Supervisión de cooperativas´ en las semanas previas.

e) El ´Mundial de fútbol Rusia 2018´ y la ´Ley que regula la publicidad estatal´ en medios fueron los temas de agenda política a partir de la quincena de junio (semanas 24, 25 y 26).

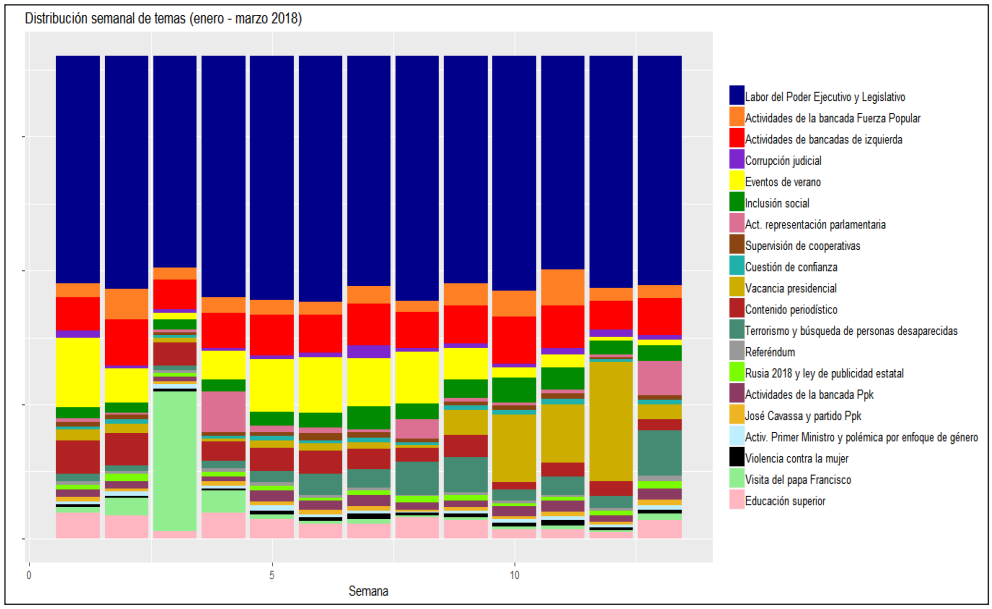

Figura 8 . Distribución semanal de temas (enero a marzo de 2018)

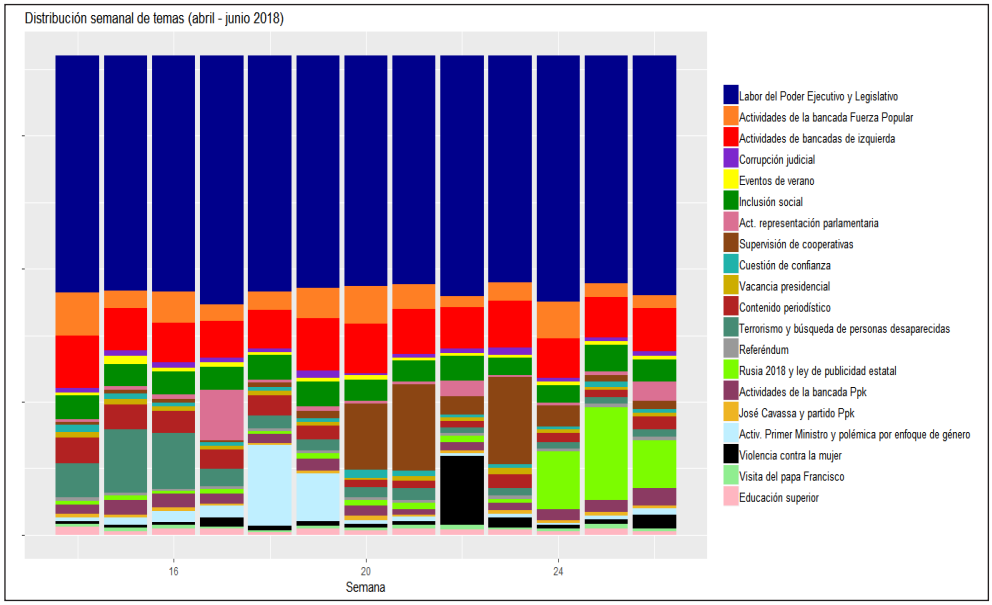

Figura 9. Distribución semanal de temas (abril a junio del 2018)

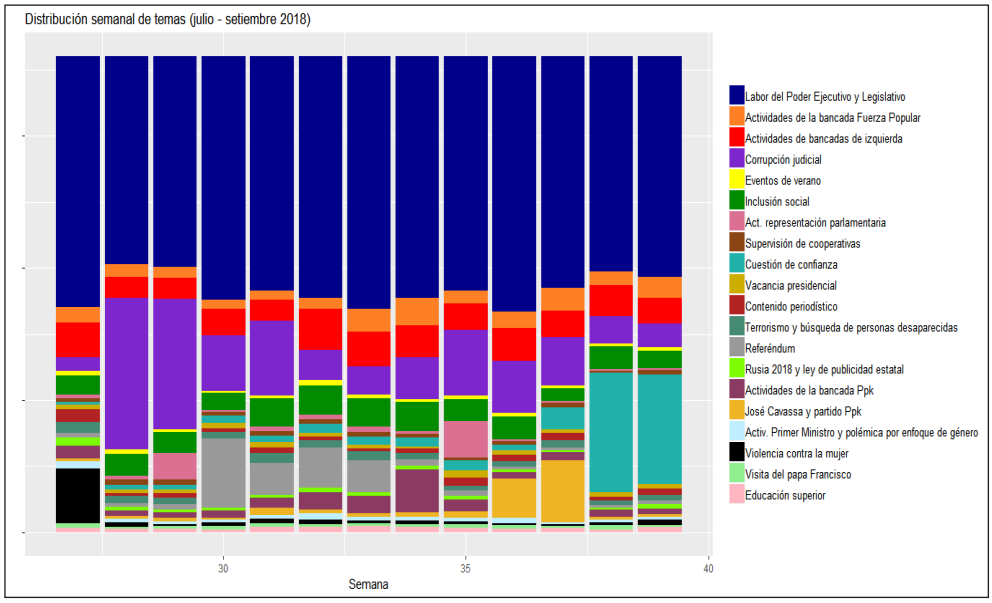

Figura 10. Distribución semanal de temas (julio a setiembre del 2018)

Finalmente, en la Figura 10 resaltan tres temas, en la agenda política del país, del tercer trimestre del año 2018:

a) Las denuncias de corrupción judicial y política que salieron a la luz la semana 28 mediante los denominados ´CNM Audios´.
b) El pedido de referéndum durante el mensaje a la nación del presidente Martín Vizcarra, en la semana 30.
c) El pedido de cuestión de confianza para el referéndum, en la semana 38.

Agenda por grupo político

Se muestra, en las Figuras 11 y 12 las agendas políticas del Poder Ejecutivo y Fuerza Popular, respectivamente, grupos políticos que se han venido enfrentando durante varios meses, según lo señalado en el diario El Comercio (2018d). En el primer caso, es notoria la división natural en tres grandes temas: labor del Poder Ejecutivo y Legislativo, inclusión social y educación superior. Además, nótese que disminuye la cantidad de contenido textual de este último tema una semana después de que Vizcarra asuma la presidencia, pasando a tener una agenda más variada (a partir de la semana 13).

En contraste al Poder Ejecutivo, la agenda política de la bancada de Fuerza Popular se ha concentrado en 3 o 4 temas por semana, dos de los cuales se refieren a su labor como Congresistas de la República y como miembros fujimoristas, y los restantes sobre un tema coyuntural y de corta duración, por ejemplo, los acontecimientos de verano, la vacancia presidencial, el terrorismo, la cuestión de confianza, etc. También resalta de manera periódica el tema de ´actividades de representación parlamentaria´.

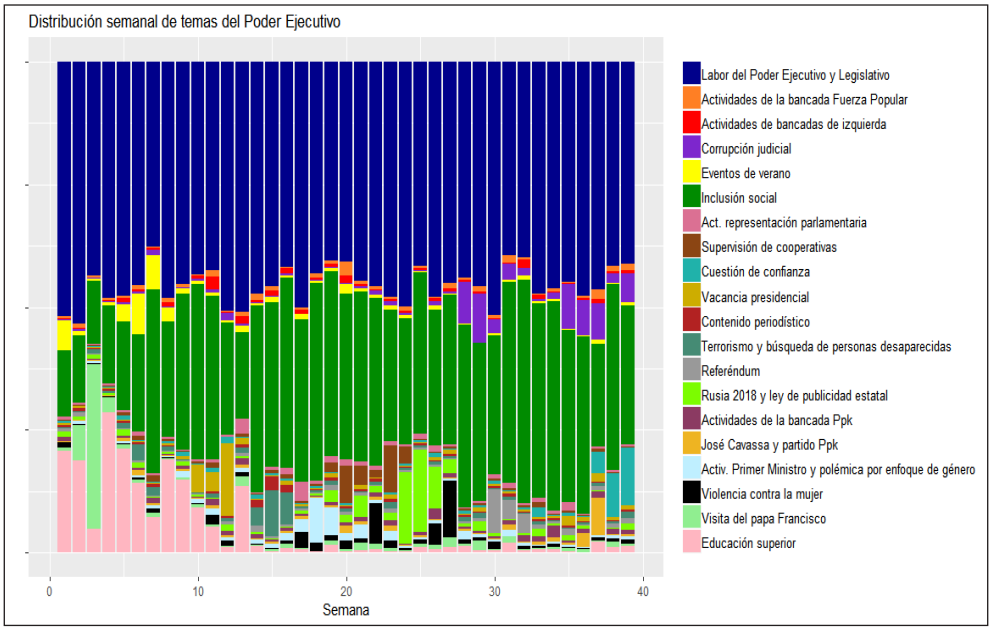

Figura 11. Agenda política del Poder Ejecutivo

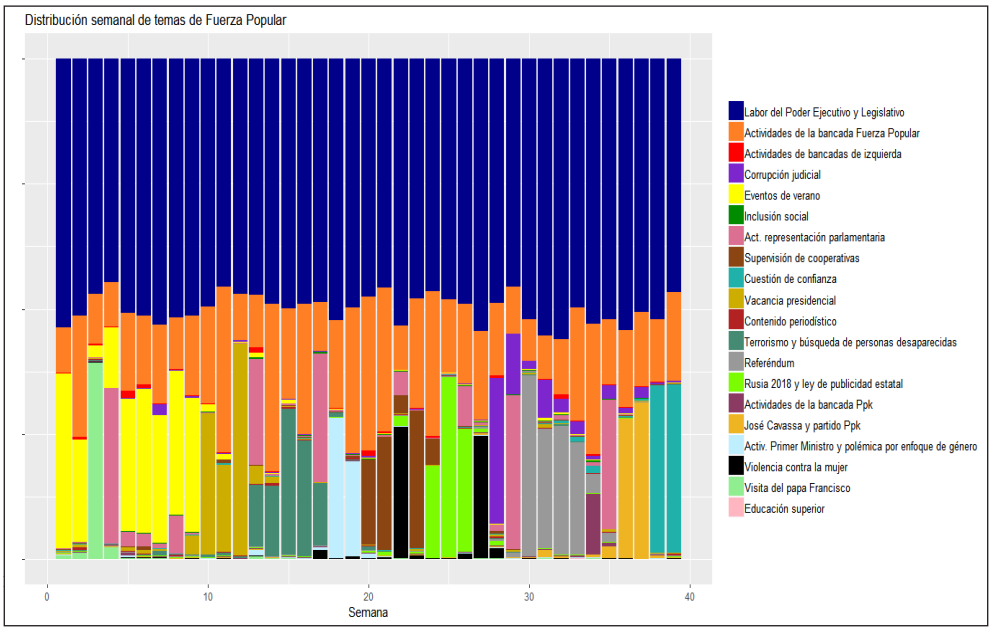

Figura 12. Agenda política del grupo parlamentario Fuerza Popular

Finalmente, en la Figura 13 se tiene la agenda política de los 9 grupos políticos para todo el periodo en estudio:

1. Los contenidos de mayor recurrencia entre los políticos del Poder Ejecutivo responden a sus labores regulares, así como los de inclusión social y, en menor medida, la educación superior del país. Puede observarse también siendo este patrón de contenidos textuales bastante disímil con los grupos políticos parlamentarios.
2. La agenda política de la bancada Fuerza Popular prioriza, además de sus actividades regulares, los eventos de verano y las actividades de representación parlamentaria. Es la bancada con mayor contenido respecto a este último tema.
3. Las bancadas de izquierda (Nuevo Perú y Frente Amplio por la Justicia, Vida y Libertad) presentan agendas políticas similares: además de comunicar sus labores cotidianas, prima el contenido acerca de la corrupción judicial.
4. Las bancadas de Acción por el Progreso y Peruanos por el Kambio también se inclinan por compartir contenidos acerca de corrupción en el sistema de justicia. Este último grupo parlamentario, de manera similar a Fuerza Popular y a las bancadas de izquierda, suele compartir sus actividades partidarias, aunque en menor medida.
5. La Célula Parlamentaria Aprista opta por informar acerca del terrorismo y la búsqueda de personas desaparecidas. Fuera de ello, su agenda de temas difundidos es bastante variada.
6. Los congresistas no agrupados anteponen el contenido periodístico. A excepción de los temas de ´Terrorismo y búsqueda de personas desparecidas´ y ´Contenidos periodísticos´, las agendas políticas de la Célula Parlamentaria Aprista y los congresistas no agrupados es muy similar.
7. La bancada Acción Popular muestra una agenda bastante variada, dando pesos no muy distintos a todos los temas, a excepción de las labores del Poder Ejecutivo y Legislativo que resaltan al igual que en los otros grupos políticos.

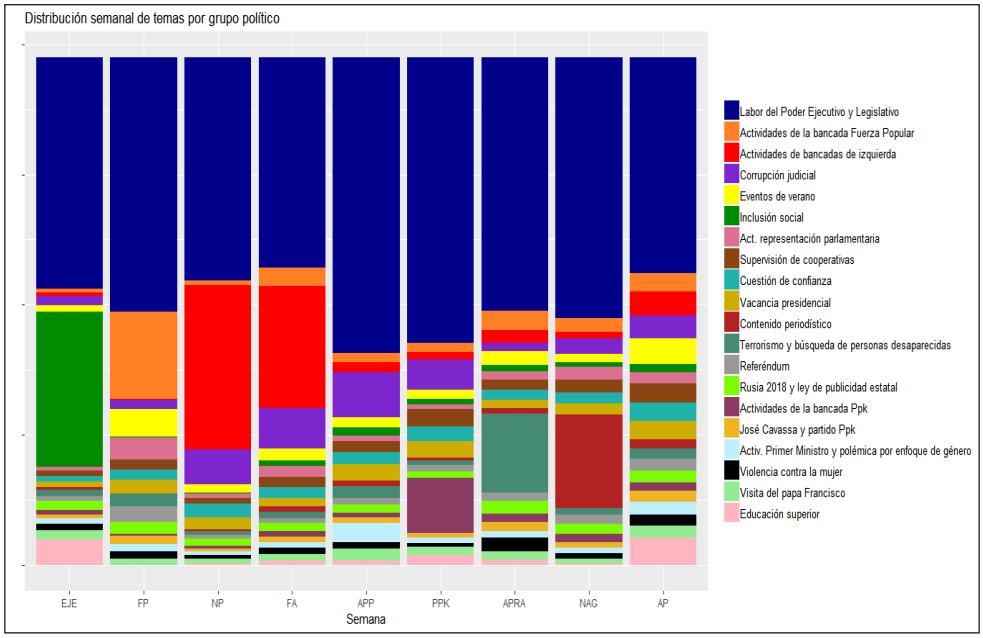

Figura 13. Agenda por grupo político

4. Conclusiones

La técnica de Topic Modeling mediante la Asignación Latente de Dirichlet permite efectivamente identificar los temas que son tratados en un gran conjunto de textos. En efecto, se detectó que la principal prioridad de los políticos peruanos ha sido compartir las labores de gestión que conlleva su cargo. Por otro lado, si bien el tema de corrupción judicial se originó recién en el tercer trimestre, su recurrencia ha sido suficiente para poder ser uno de los temas de mayor difusión en el año. El mismo efecto, pero en menor medida, sucede con los eventos acontecidos durante los meses de enero y febrero. Cabe resaltar que los políticos han mostrado poco interés en las elecciones municipales y regionales llevadas a cabo en el mes de octubre de 2018, ya que el análisis no logra localizar este tema, a diferencia de otros eventos puntuales tales como la cuestión de confianza, el mundial de fútbol Rusia 2018 y la visita del papa Francisco, los cuales sí tuvieron eco en las cuentas de Twitter de los políticos peruanos. Al comparar las agendas por grupos políticos, se nota que estos presentan agendas con temas disímiles: mientras que el Poder Ejecutivo otorga importancia a las actividades de inclusión social, Fuerza Popular y las bancadas de izquierda dan espacio a sus actividades partidarias. Por otra parte, el tema de corrupción judicial ha sido mencionado por el Poder Ejecutivo, Fuerza Popular y la Célula Parlamentaria Aprista en menor medida que otros grupos políticos. Para futuros trabajos se recomienda adicionar análisis de sentimientos textuales, ya que no solo se estudiarían los contenidos sino el grado de concordancia o discordancia de los autores con lo que escriben. Además, es posible considerar el uso de bigramas y trigramas, así como de técnicas que permitan la limpieza de términos textuales con errores ortográficos

5. Literatura citada

Alvarado, J.; Carrillo, A.; Forero, J.; Caicedo, L.; Urueña, J. 2016. Análisis de sentimiento político en twitter para las elecciones de la alcaldía de Bogotá 2015. En: XXVI Simposio Internacional de Estadística, Colombia, 8-12 ago, 2016.

Arun, R.; Suresh, V.; Madhavan, C.; Murty, M. 2010. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. En: Zaki, M.J. *et al*. (Eds.). Advances in Knowledge Discovery and Data Mining. Springer, Alemania. 391-402 p.

Blei, D.; Ng, A.; Jordan, M. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research (3): 993-1022.

Cao, J.; Xia, T.; Li, J.; Zhang, Y.; Tang, S. 2009. A density-based method for adaptive LDA model selection. Neurocomputing (72): 1775 – 1781.

Cardenas, R.; Bello, K.; Coronado, A.; Villota, E. 2015. Labor market demand analysis for engineering majors in Peru using Shallow Parsing and Topic Modeling. Machine Learning Summer School. Japón.

Darling, W. 2011. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. Reporte técnico. Disponible en http://u.cs.biu.ac.il/~89-680/darling-lda.pdf

Deveaud, R.; Sanjuan, E.; Bellot, P. 2014. Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval. Revue des Sciences et Technologies de l'Information - Série Document Numérique (17): 61-84.

Presentan moción de vacancia contra PPK por "incapacidad moral". 2018. Correo, Lima, Perú; 8 marzo. Disponible en https://goo.gl/MX9n1s

PPK: presentan moción de vacancia presidencial. 2017. El Comercio, Lima, Perú; 15 diciembre. Disponible en https://goo.gl/cwvvGx

El diálogo entre Kenji, Bocángel y Mamani. 2018ª. El Comercio, Lima, Perú; 21 marzo. Disponible en https://goo.gl/fxuzGG

Osmán Morote, el cabecilla de Sendero que deja el penal de Ancón. 2018b. El Comercio, Lima, Perú; 18 abril. Disponible en https://goo.gl/7tXzLS

Audios complican situación de jueces y miembros del CNM. 2018c. El Comercio, Lima, Perú; 20 julio. Disponible en https://goo.gl/nujjmh

Pugna entre Ejecutivo y FP se reaviva: Vizcarra y Fujimori frente a frente. 2018d. El Comercio, Lima, Perú; 28 agosto. Disponible en https://goo.gl/BMGTb1

Martín Vizcarra: "Los cuatro proyectos de reforma tienen que aprobarse". 2018e. El Comercio, Lima, Perú; 19 setiembre. Disponible en https://goo.gl/24UrJj

Odebrecht: Jorge Barata revela aportes a PPK, Fuerza Popular, Nacionalistas y Apra. 2018. Gestión, Lima, Perú; 28 febrero. Disponible en https://goo.gl/k9SuHW

Agricultores de papa se van otra vez a la huelga. 2018a. La República, Lima, Perú; 31 enero. Disponible en https://goo.gl/e7PCki

Congreso aceptó renuncia de PPK. 2018b. La República, Lima, Perú; 23 marzo. Disponible en https://goo.gl/SNihbK

Eyvi Ágreda: Murió la joven que fue quemada en Miraflores. 2018c. La República, Lima, Perú; 1° junio. Disponible en https://goo.gl/rQfjBB

José Luis Cavassa habría trabajado para partido PPK en campaña de 2016. 2018d. La República, Lima, Perú; 7 setiembre. Disponible en https://goo.gl/mZ1iRM

A 51 se eleva la cifra de muertos tras accidente en Pasamayo. 2018a. Perú21, Lima, Perú; 3 enero. Disponible en https://goo.gl/2xSgQG

Así se desarrolló la 'Marcha por la Vida' en Lima. 2018b. Perú21, Lima, Perú; 5 mayo. Disponible en https://goo.gl/zgUjh8

Datum. 2018c. Credibilidad de políticos desciende a niveles alarmantes, advierten. Perú21, Lima, Perú; 14 ago. Disponible en https://goo.gl/ytao8z

Fang, A.; Ounis, I.; Habel, P.; Macdonald, C.; Limsopatham, N. 2015. Topic-centric Classification of Twitter User's Political Orientation. In: 6th Symposium on Future Directions in Information Access, Grecia.

Farías, M. 2017. Twitter como vía para mediar los conflictos sociales: análisis del caso #Conga, Perú. Tesis de licenciatura. Universidad de Piura, Piura. Perú. 89 p.

Fowks, J. 2017. Protesta masiva en Lima contra Kuczynski por el indulto a Fujimori. El País, Lima, Perú; 29 dic. Disponible en https://goo.gl/SbhqQt

Greene, D.; Cross, J. 2017. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. Political Analysis (25): 77-94.

Griffiths, T.; Steyvers, M. 2004. Finding scientific topics. National Academy of Sciences of the United States of America (101): 5228-5235.

Grimmer, J. 2009. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. Political Analysis (18): 1-35.

Grün, B.; Hornik, K. 2011. Topicmodels: An R package for fitting topic models. Journal of Statistical Software (40): 1-30.

Heinrich, G. 2008. Parameter estimation for text analysis. Reporte técnico. Disponible en http://www.arbylon.net/publications/text-est.pdf

Hidalgo, M. 2017. La vacancia de PPK se resuelve el próximo jueves 21. El Comercio, Lima, Perú; 16 dic. Disponible en https://goo.gl/GJKnnP

Linares, R.; Herrera, J.; Cuadros, A.; Alfaro, L. 2015. Prediction of tourist traffic to Peru by using sentiment analysis in Twitter social network. In: Latin American Computing Conference, 19-23 oct, 2015. Lima, Perú.

Mateo, J. 2016. Análisis de contenidos en Social Media: Clasificación de mensajes e identificación de influyentes en el Banco Central Europeo (BCE). Trabajo de máster. Universidad Complutense de Madrid, Madrid, España. 71 p.

Montesinos, L. 2014. Análisis de sentimientos y predicción de eventos en Twitter. Tesis de pregrado. Universidad de Chile, Santiago de Chile. Chile. 60 p.

Pla, F.; Hurtado, L. 2014. Political Tendency Identification in Twitter using Sentiment Analysis Techniques. In: 25th International Conference on Computational Linguistics, 23-29 agosto, 2014. Irlanda.

Presidencia del Consejo de Ministros. 2018. Decreto Supremo que convoca a Elecciones Regionales y Municipales 2018; 4 ene. Disponible en https://goo.gl/xKjcrE

Ramos, I. 2018. Vizcarra convoca a referéndum para combatir la corrupción en Perú. Diario Financiero, Lima, Perú; 10 oct. Disponible en https://goo.gl/Bu1JQe

SBS [Superintendencia de Banca y Seguros]. 2018. SBS Informa. Boletín Semanal N° 021. Disponible en https://goo.gl/AGvHLi

Sigueñas, M. 2016. Técnicas de Minería de Textos para el Análisis de Discursos y Documentos. Disponible en https://goo.gl/i79hg8

Vilcachagua, P. 22 de enero de 2018. Papa Francisco en el Perú: Lo que nos dejó la visita del Sumo Pontífice. Perú21. Disponible en https://goo.gl/xX1dan

Vílchez, C.; Alhuay, J. 2016. Use of text mining for understanding Peruvian students and faculties' perceptions on bibliometrics training. In: 3rd Annual International Symposium on Information Management and Big Data, 1-3 set, 2016. Lima, Perú.

Vollenweider, C. 2018. 2018: El año de la crisis peruana. Disponible en http://www.celag.org/2018-el-ano-crisis-peruana/

Yano, T.; Cohen, W.; Smith, N. 2009. Predicting response to political blog posts with topic models. Proceedings of Human Language Technologies. In: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 1° jun., 2009. Estados Unidos.