UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Posgrados

Análisis del Discurso Político del Presidente Rafael Correa en los Enlaces Ciudadanos

María Gabriela Juncosa Calahorrano

Carlos Jiménez, Ph.D.
Director de Trabajo de Titulación

Trabajo de titulación de posgrado presentado como requisito
para la obtención del título de Maestría en Matemáticas Aplicadas

Quito, 18 de diciembre de 2017

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE POSGRADOS

HOJA DE APROBACIÓN DE TRABAJO DE TITULACIÓN

Análisis del Discurso Político de Rafael Correa en los Enlaces Ciudadanos

María Gabriela Juncosa Calahorrano

Firmas

Carlos Jiménez, Ph.D.

Director del Trabajo de Titulación _____

Carlos Jiménez, Ph.D.

Director de la Maestría de Matemáticas Aplicadas _____

César Zambrano, Ph.D.

Decano del Colegio de Ciencias e Ingenierías,

Politécnico _____

Hugo Burgos, Ph.D.

Decano del Colegio de Posgrados _____

Quito, 18 de diciembre de 2017

© Derechos de Autor

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante: _____

Nombre: María Gabriela Juncosa Calahorrano

Código de estudiante: 00126510

C. I.: 1719566638

Lugar, Fecha Quito, 18 de diciembre de 2017

AGRADECIMIENTOS

Agradezco a Carlos Jiménez y Ricardo López por los valiosos comentarios que enriquecieron este trabajo. También, agradezco a Felipe Vaca y Wilson Moreno por sus sugerencias y asistencia en todo el proceso.

RESUMEN

Los enlaces ciudadanos del presidente Rafael Correa se han convertido en un referente de opinión pública y discurso político. Este trabajo busca analizar los enlaces ciudadanos utilizando estrategias de minería de texto. Para ello, se implementan dos metodologías: (1) análisis de sentimientos y, (2) modelamiento de tópicos. El análisis de sentimientos arroja que la mayoría de enlaces utilizan, en promedio, palabras que no son ni activas, ni pasivas y, tampoco son ni agradables, ni desagradables. Por otro lado, los enlaces utilizan palabras, en promedio, más fáciles de imaginar. Finalmente, el modelamiento de tópicos ajustó 20 tópicos y cada enlace pertenece a tres tópicos.

Palabras clave: modelamiento de tópicos, análisis de sentimientos, Rafael Correa, sabatinas, discurso político

ABSTRACT

Rafael Correa's weekly speeches (*enlaces ciudadanos*) have become a reference for public opinion and political discourse. This works aims to analyze Correa's weekly speeches using data mining strategies. Two methodologies have been used for such purpose: (1) sentiment analysis and, (2) topic modeling. Sentiment analysis shows that Correa uses in his weekly speeches words which, on average, are neither active, nor passive and, neither pleasant, nor unpleasant. On the other hand, Correa uses words which, on average, are easier to imagine. Finally, topic modeling adjusted 20 topics and speeches belong to, at most, three topics.

Key words: Topic modeling, sentiment analysis, Rafael Correa, sabatinas, political discourse

Índice

Índice de figuras

Índice de cuadros

1. Introducción

La importancia política de los Enlaces Ciudadanos del presidente Rafael Correa los han convertido en un referente de opinión pública y discurso político.

Los Enlaces Ciudadanos, al principio conocidos como Diálogo con el Presidente, fueron pensados como un espacio para que el presidente transmita a la ciudadanía la ideología detrás de su gestión. Desde el inicio, el blanco de los Enlaces fue la prensa que, en los primeros enlaces, tenía un segmento propio para hacer preguntas directas al Presidente. Si bien todos los gobiernos han tenido una relación conflictiva con la prensa, el gobierno de Rafael Correa es quizá el primero en la historia del país que interactuó directamente con la prensa, no como un medio de resistencia sino como el actor de resistencia en sí mismo.

Como es de esperarse, parte importante de los Enlaces Ciudadanos fueron destinados a criticar a la prensa. Esto nos lleva a pensar que pueden existir otros temas recurrentes en el discurso de Correa. Con esto en mente, este trabajo busca identificar los temas subyacente del discurso del presidente Rafael Correa con el objetivo de (1) caracterizar el discurso de Correa e, (2) analizar el tono o sentimiento de su discurso.

Para caracterizar el discurso de Correa, este trabajo utiliza el **Modelamiento de Tópicos**. Un modelo de tópicos es un algoritmo no supervisado que descubre estructuras latentes subyacentes en el corpus para asignar palabras a tópicos. Para ello, se asume que las palabras en un documento se generan a partir de un perfil de probabilidades que forma cierto tópico y así se puede saber si un documento que habla de cierto tópico utilizará ciertas palabras con más o menos frecuencia. Supongamos que se cuenta con un documento cuyo tópico es "estadística"; entonces, existe un 20 % de probabilidad que una palabra en este documento sea "varianza", 10 % que esta palabra sea "inferencia" y 5 % que sea "aproximación". Por otro lado, si escogemos como tópico "antropología" existe un 30 % de probabilidad que una palabra en un documento con énfasis en este tópico sea "cultura", 10 % que una palabra sea "identidad" y 5 % que sea "diversidad".

Por otro lado, para estudiar el tono o sentimiento del discurso se utilizó el Análisis de

Sentimientos que es un estrategia de **codificación basada en diccionarios**. Una estrategia de Análisis de Sentimientos utiliza un diccionario de sentimientos construido previamente en el cual cada palabra tiene un puntaje que cuantifica el grado de expresión de un sentimiento en particular. Después, se debe empatar las palabras de dicho diccionario con las palabras en el texto que se desea evaluar. Finalmente, se calcula un indicador de sentimiento, este indicador puede ser un promedio o un suma simple de puntajes, dependiendo del contexto.

Para este trabajo se utiliza dos diccionarios. El primero, registra en una escala del 1 al 3 el grado de imaginabilidad, agrado y actividad de un conjunto de palabras. Los autores de este diccionario pidieron a un grupo de voluntarios que califiquen a un conjunto de palabras en los 3 sentimientos. El diccionario contiene el promedio de estas calificaciones. Para obtener un puntaje de imaginabilidad, agrado y actividad de las sabatinas se ha calculado un promedio de los puntajes asignados a las palabras.

El segundo diccionario evalúa la positividad y negatividad de las palabras en el diccionario. En este caso, se ha hecho una sumatoria simple para representar este sentimiento. Si la sumatoria es mayor a cero, se puede decir que la sabatina en cuestión contiene más palabras positivas; por otro lado, si la sumatoria es negativa, el texto tiene más palabras con connotación negativa.

Después de hacer este análisis, hemos identificado que el discurso del presidente Correa no es ni activo, ni pasivo y tampoco es agradable o desagradable. Sin embargo, el lenguaje que utiliza se asocia con palabras que son más fáciles de imaginar; es decir, su discurso no es abstracto.

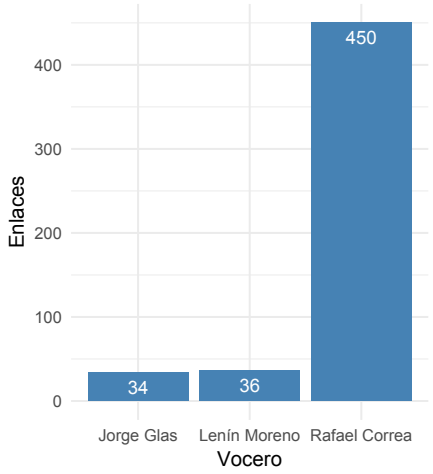
Este trabajo se organiza de la siguiente manera: primero, se describe brevemente qué es una Enlace Ciudadano y luego se presenta un resumen de las estrategias que se puede utilizar para analizar el discurso político. Segundo, se describe los datos disponibles para este trabajo y la preparación de los mismos. Tercero, se describe la metodología de trabajo. Cuarto, se presenta los resultados y; finalmente, se concluye el trabajo con algunas recomendaciones.

2. Los Enlaces Ciudadanos y el Análisis del Discurso Político

2.1. Los Enlaces Ciudadanos

El Enlace Ciudadano es el programa semanal transmitido por 362 radios FM, 76 radios AM, 99 canales de televisión local, 2 canales de televisión nacional y YouTube [1], en el que el presidente Rafael Correa rinde cuentas sobre la gestión de la semana, temas coyunturales y gestión gubernamental. Desde enero 2007, los Enlaces Ciudadanos se han transmitido todos los sábados a partir de las 10h00 y tienen una duración promedio de 2 horas, 57 minutos, 56 segundos [2]. Hasta abril del 2017, se transmitieron 520 enlaces que en total representan más de 1500 horas de discurso. De los 520 enlaces, el presidente Correa dirigió 450, el ex-vicepresidente del primer periodo, Lenín Moreno, 36 y el vicepresidente del segundo periodo y el actual vicepresidente electo, Jorge Glas, 34 enlaces [2].

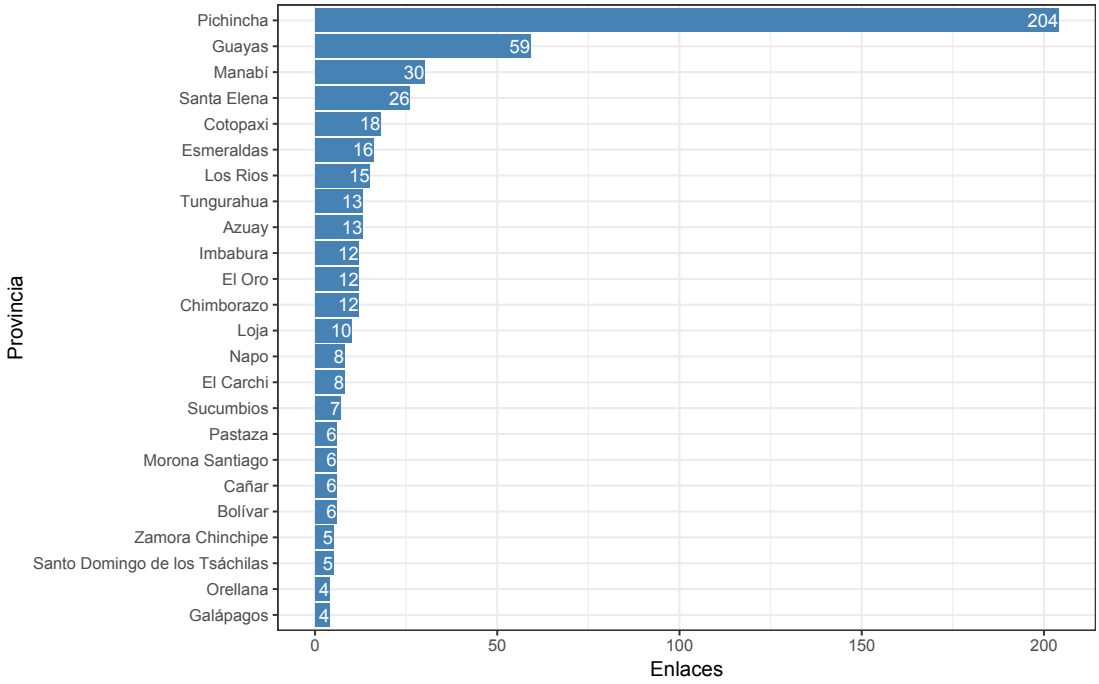
Figura 1: Voceros de los Enlaces Ciudadanos, enero 2007-abril 2017



Los Enlaces Ciudadanos se han trasmitido desde todas las provincias del Ecuador, al menos 1 vez. La mayoría de enlaces se transmitieron desde la provincia de Pichincha (204), seguida de Guayas (59), Manabí (30) y Santa Elena (26). Las provincias que tuvieron menor presencia de enlaces fueron las Islas Galápagos y Orellana, cada una con 4 enlaces registrados,

y las provincias de Santo Domingo de los Tsáchilas y Zamora Chinchipe, cada una con 5 enlaces.

Figura 2: Enlaces Ciudadanos por Provincia, enero 2007-abril 2017



Los Enlaces Ciudadanos, al principio conocidos como Diálogo con el Presidente, fueron pensados como un espacio para que el presidente explique a la ciudadanía la ideología detrás de su gestión. Los primeros enlaces se organizaron bajo el formato pregunta-respuesta. El Presidente iniciaba con una descripción de las actividades de la semana, aprovechando la oportunidad para transmitir sus pensamientos e impresiones personales. Después, daba paso a preguntas de la ciudadanía y de periodistas invitados a participar en estos espacios. Este formato creó múltiples confrontaciones entre el Presidente, en su papel de vocero principal, y los medios invitados. El principal conflicto fue el incidente con Emilio Palacios, columnista de diario El Universo. Después de este, se eliminó el conversatorio con los medios y la vocería de los enlaces se concentró más y más en la figura del Presidente, con esporádicas intervenciones, principalmente de carácter informativo por parte de los servidores públicos presentes.

Desde el inicio, el blanco de los Enlaces fue la prensa que, en los primeros enlaces, tenía

un segmento propia para hacer preguntas directas al Presidente. El Presidente siempre ha cuestionado abiertamente a la prensa ecuatoriana porque, según su criterio, se ha alejado de su función de informar y en lugar de ello, *manipula* la información para favorecer intereses económicos y políticos de sus dueños o aliados. Si bien todos los gobiernos han tenido una relación conflictiva con la prensa, el gobierno de Rafael Correa es quizá el primero en la historia del país que interactúa directamente con la prensa. De esta manera, la prensa pasa de un medio de resistencia, a ser la resistencia en sí misma. En el pasado, la prensa ha funcionado como una resistencia al margen del escenario político, hoy está en el centro. Antes de Correa, la prensa ecuatoriana no se había enfrentado a una crítica tan persistente, ni sostenida.

El formato de los enlaces, así como su duración, ha evolucionado con el tiempo. Al inicio, el Diálogo con el Presidente tenía tres segmentos: (1) informe de las actividades de la semana, (2) diálogo con los periodistas, y (3) espacio destinado a ciencia y tecnología [3]. A partir de mayo 2007, después del incidente con Emilio Palacio, se eliminó el segmento con los periodistas para ser retomado posteriormente fuera de los enlaces en un conversatorio semanal con los medios. Posteriormente, el segmento Ciencia y Tecnología fue reemplazado por el segmento *La Libertad de Expresión* ya es de Todos y se añadió el segmento La *Cantinflada de la Semana* [4]. En el segmento *La Libertad de Expresión ya es de Todos*, el Presidente escoge notas de los medios que, según su criterio, distorsionan los hechos. Siendo el vocero principal, el Presidente aprovecha estos espacios para transmitir su versión de los hechos. Por otro lado, en el segmento *La Cantinflada de la Semana*, el Presidente selecciona segmentos de programas radiales, televisivos o notas de prensa escrita que el considera absurdos, para ponerles en evidencias del público [3].

Tanto la regresión lineal (ver figura 3) como cuadrática evidencian que la duración de los enlaces ha incrementado a medida que ha pasado el tiempo. Sin embargo, la regresión evidencia que la tasa de crecimiento es decreciente; es decir, no se prevé que la duración de los enlaces incremente indefinidamente, sino que se espera una estabilización del crecimiento que

según el gráfico sucede entre las 3 y 4 horas. Para este propósito se ajustaron las siguientes ecuaciones:

$$duracion\_horas_i = \beta_0 + \beta_1 * no\_enlace_i + e \tag{1}$$

Que dio como resultado:

$$duracion\_horas_i = 1{,}84 + 0{,}004 * no\_enlace + e, \quad R^2_{adj} = 0{,}6462$$

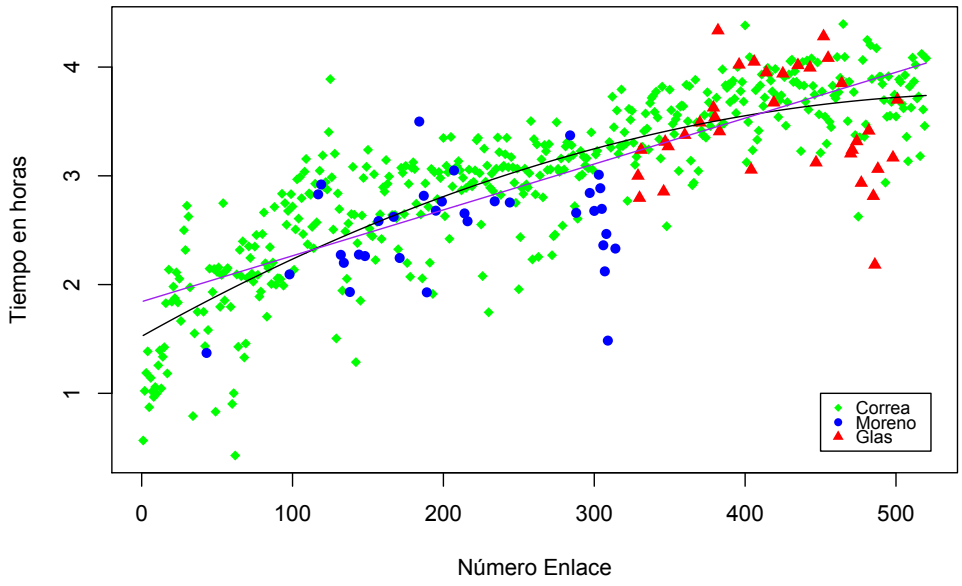
y,

$$duracion\_horas_i = \beta_0 + \beta_1 * no\_enlace_i + \beta_2 * no\_enlace_i^2 + e \tag{2}$$

$$duracion\_horas_i = 1{,}524 + 7{,}778e-03 * no\_enlace - 6{,}767e-06 * no\_enlace^2 + e$$

$$R^2_{adj} = 0{,}6758$$

Figura 3: Evolución de la duración de los enlaces, diferenciado por vocero



2.2. Estrategias de Análisis del Discurso Político

Para analizar un texto de ciencias políticas se puede utilizar 5 estrategias distintas: lectura a profundidad, codificación humana, diccionarios automatizados, aprendizaje supervisado y

modelamiento de tópicos [5]. Cada estrategia tiene sus costos y beneficios, para analizarlos resulta útil enfocarse en dos aspectos: (1) qué tanto conocimiento específico requieren los métodos y (2) el tiempo que una persona requiere para completar el análisis de un cuerpo de texto. Estos costos se deben evaluar en las tres etapas del análisis: (1) la fase del pre-análisis en dónde se conceptualiza y operativiza el problema, (2) la fase de análisis donde se categoriza los textos de interés y, (3) la fase del pos-análisis en la que los resultados de la fase de análisis se interpretan y evalúan [5]. A continuación, se describe brevemente las características de estos 5 métodos, ordenando la presentación de menor a mayor automatización; es decir, se presenta desde el método con menor automatización (clasificación por lectura a profundidad) hasta el de mayor automatización (modelamiento de tópicos).

El método más común para extraer significado de un texto es la **lectura a profundidad**. En esta estrategia, las categorías relevantes dependen del lector. Cualquier persona puede crear el número de categorías que considere necesario, en cuento tenga acceso a mayor información del contenido de un texto. Como resultado, este método se basa en menos supuestos que otros métodos más automatizados. El número de temas no se fija a priori, no se asume ningún tipo de relación entre categorías, los textos se pueden analizar holísticamente y la categorización se la ejecuta caso por caso; es decir, no existe un proceso algoritmizado para especificar las categorías. Un método con pocos supuestos como la **lectura a profundidad** permite mucha flexibilidad en el análisis; sin embargo, implica costos significativos. La lectura de texto requiere un nivel de conocimiento alto sobre el o los temas a tratar y el lenguaje en el que están escritos, además, requiere de una alta dedicación de tiempo. Finalmente, el procesamiento de información requiere excelente criterio, dominio del tema, experiencia y mucha reflexión por parte de quien hace el trabajo [5].

Una alternativa al método de la lectura a profundidad es la **codificación de textos**. La codificación es el método estándar en la investigación cualitativa y análisis de contenido de las ciencias sociales. Para la codificación manual es necesario conocer y fijar las categorías de antemano. Las personas que codifican leen las unidades de texto y asignan uno de los códigos

finitos que han sido preestablecidos. El método asume conocimiento de cualquier relación entre las categorías de interés. La clasificación no se basa en ningún requisito preestablecido y el proceso exacto de mapeo de textos es desconocido. Se puede validar la clasificación comparando el trabajo de dos personas independientes; sin embargo, no se puede conocer cómo estos individuos llegaron a tales conclusiones. La codificación manual es útil cuando se cuenta con abundantes recursos humanos y las categorías interés están bien definidas, son mutuamente excluyentes y exhaustivas; sin embargo, el proceso de mapeo es altamente complejo y desconocido para todos excepto para quien codifica. La ventaja de este método por sobre la clasificación por lectura es que requiere menos conocimiento del tema del texto, aunque se requiere algo de conocimiento para comprender el lenguaje y contextualizar correctamente el contenido. A pesar de ello, sigue siendo menos costoso que la lectura a profundidad, aunque los costos iniciales son mayores ya que crear un esquema de categorización útil requiere conocimiento profundo del tema a tratarse y una importante dedicación de tiempo [5].

La **codificación basada en diccionarios** es el primer paso hacia la automatización del análisis de contenido. Este método requiere que el analista desarrolle una lista o diccionario de palabras y frases que pueden indicar afiliación a una categoría particular. Después, un algoritmo calcula las frecuencias de palabras en cada texto y determina el grado de pertenencia a una de las categorías preestablecidas. Al igual que con la codificación humana, el método del diccionario tiene costos iniciales altos, debido a la necesidad de definir las categorías, identificar la lista de palabras que conforman una categoría y definir la estrategia de mapeo. Sin embargo, una vez que se cuenta con un buen diccionario, los costos de análisis son más bajos que la codificación humana [5].

El **aprendizaje supervisado** presenta una alternativa tentadora a la codificación basada en diccionarios, utilizando métodos estadísticos para automatizar el proceso de codificación manual. Los algoritmos supervisados requieren de codificación manual para un subconjunto de los textos—conocido como el grupo de entrenamiento—y el resto servirán como el conjunto de prueba. Los algoritmos de aprendizaje de máquinas se utilizan para inferir el

mapeo entre las características de los textos y las categorías del conjunto de entrenamiento. El mapeo resultante de este análisis se valida con el conjunto de prueba, a través de medidas de exactitud de predicción. Para la aplicación de métodos de aprendizaje supervisado, las categorías se asumen conocidas y fijas. Esto implica que se debe identificar un conjunto de características relevantes a priori; sin embargo, el algoritmo se encarga de discriminar cuáles de estas características son relevantes y definir la estrategia de mapeo hacia las categorías de interés. El aprendizaje supervisado tienen costos iniciales altos, ya que requieren de intervención humana para la codificación inicial de los documentos que pertenecen al conjunto de entrenamiento. Sin embargo, dado que el proceso de asignación está automatizado, se puede asignar grandes cantidades de texto a categorías en poco tiempo. La ventaja y, por lo tanto, la reducción de costos yace en la posibilidad de procesar cantidades de texto mayores [5].

Finalmente, el **modelamiento de tópicos** es una estrategia de clasificación que busca automatizar el proceso de categorización. El **modelamiento de tópicos** es capaz de rescatar la flexibilidad del método de lectura a profundidad, ya que no asume ni fija las categorías, ni establece relaciones entre ellas de antemano. Este supuesto es fundamental en el método de codificación manual, codificación basada en diccionario y aprendizaje supervisado. En este método, las categorías de interés son objeto de inferencia. El método asume que las palabras son una característica relevante del contenido de los temas de un texto; además, asume que el mapeo de palabras a temas toma una forma paramétrica particular. El **modelamiento de tópicos** busca identificar, en lugar de asumir, las categorías temáticas, los parámetros que describen el mapeo de palabras a temas y la categoría temática de un texto dado. Los costos de este método tienen una estructura distinta al resto de métodos antes descritos. Mientras que los métodos anteriores concentran sus costos en las etapas del pre-análisis (codificación manual, codificación basada en diccionarios, aprendizaje supervisado) y análisis (lectura a profundidad, codificación manual), el **modelamiento de tópicos** requiere poco tiempo y conocimiento en estas etapas; sin embargo, requiere mayor tiempo y esfuerzo, pero no mayor conocimiento, en la etapa del pos-análisis, debido a que el usuario debe interpretar los temas

que crea el algoritmo de tal forma que sean relevantes para el contenido de los textos por analizar [5].

En este trabajo se utiliza dos de estas cinco estrategias. Primero, utilizamos una estrategia de **codificación basada en diccionarios** conocida como Análisis de Sentimientos. Una estrategia de Análisis de Sentimiento asigna un puntaje a cada palabra en el léxico para caracterizar el sentimiento que se busca identificar. Después, se debe rastrea las palabras del léxico que están contenidas en el texto de interés y se desarrolla un indicador de sentimiento. Los indicadores de sentimiento suelen ser promedios o sumas de puntajes, dependiendo del contexto. En nuestro caso, contamos con dos bases de datos. La primera cuenta con medidas sobre el agrado, la imaginabilidad y la actividad o activación de una palabra, utilizando una escala del 1 al 3 para denotar el grado de presencia de estos sentimientos [7]. Cada palabra en el léxico cuenta con un promedio de puntajes otorgado en un trabajo con voluntarios. Para este trabajo, hemos calculado el puntaje promedio de cada uno de estos sentimientos por Enlace Ciudadano. La segunda base de datos describe la positividad y negatividad de las palabras en el léxico [6]. Con ello, es posible calcular la sumatoria de los puntajes otorgados. Si esta es mayor que cero, el texto tiene más palabras con connotación positiva; por otro lado, si la sumatoria es negativa, el texto tiene más palabras con connotación negativa.

En segundo lugar, utilizamos el **modelamiento de temas** utilizando los dos metodologías disponibles: Latent Dirichlet Allocation (LDA) y Correlated Topic Model (CTM). Se han corrido múltiples versiones de los modelos. Primero, LDA tiene tres opciones de métodos para estimar los parámetros: (1) Variational Expectation Maximization algorithm (VEM), (2) VEM con parámtero α fijo y (3) Muestreo de Gibbs. Para el modelo CTM solo existe una opción de método de ajuste (VEM). Dado que el algoritmo requiere el número de tópicos a priori, primero se realiza una simulación con 2, 3, 4, 5, 10, 20, 30, 40, 50, 75, 100 temas y se escoge aquel modelo con el menor indicador de *perplexity*.

textitPerplexity es una medidad de qué tan buena es una predicción. Supongamos que deseamos predecir un secuencia de números del uno al seis. Si lo hacemos utilizando un dado

de seis lados, existe una probabilidad de uno entre seis de predecir el número correctamente. El indicador *perplexity* para este dado, entonces, es seis. Para conjuntos de datos más grandes, se corre un modelo que describe el comportamiento de los datos y *perplexity* nos da una medida de qué tan bueno es el desempeño de los datos en comparación con un dado de x-lados. Por ejemplo, supongamos que ponemos a una persona a predecir la siguiente palabra de un texto que ha sido cortado. Las personas somos buenas predictoras de palabras, es decir, nuestro indicador de *perplexity* es 247. Esto significa que nuestras predicciones son tan buenas como las de un dado de 247 lados [8]. Formalmente, *perplexity* se define [9]:

$$Perplexity\,(w) = exp\left\{\frac{\log\left(p\left(w\right)\right)}{\sum_{d=1}^{D}\sum_{j=1}^{V} n^{(jd)}}\right\} \tag{3}$$

A continuación, presentamos las características de los datos utilizados para este análisis, así como las estrategias de pre-procesamiento necesario para la implementación de las metodologías.

3. Descripción de los Datos

3.1. Datos

Para analizar las características sobresalientes del discurso del presidente Correa, era necesario contar con transcripciones de los enlaces ciudadanos. En la etapa temprana del trabajo, se exploró la posibilidad de hacerlo mediante programas informáticos de transcripción automática, sin tener éxito.

Surgió la posibilidad de contar con recursos para contratar a un equipo que transciba los enlaces ciudadanos. La cantidad de enlaces ciudadanos que se analizaron para este trabajo estuvo restringida por los recursos disponibles para la transcripción. Se contrató a un equipo de 15 personas que en tres semanas transcribieron un total de 40 enlaces. Los audios para las transcripciones se tomaron de la página web www.enlaceciudadano.gob.ec, excluyendo aquellos enlaces que no están disponibles en el archivo del sitio web. Adicionalmente, 2 personas del equipo trabajaron 1 semana en verificar la calidad de las transcripciones. Esta verificación implicó una revisión breve de la ortografía, revisión de la coherencia de las transcripciones y una comparación aleatoria de la transcripción con el audio original. En total, el proceso de recolección de datos tomó cinco semanas, con una semana adicional para que el equipo de transcripción corrija las fallas identificadas. Si bien los textos no son perfectos en un 100 %, podemos garantizar que las transcripciones están completas, en su mayoría, son fieles al audio original y los errores tipográficos y ortográficos no sobrepasan el 97 % del texto.

Una vez determinado el número de enlaces ciudadanos a ser transcritos, la selección puede tomar varios caminos. Se puede, por ejemplo, elegir un periodo de interés y concentrar el análisis con un objetivo claro en mente o se puede escoger un muestra aleatoria de todo el periodo. Dado que el objetivo de este trabajo es investigar las características del discurso de Correa en general, se decidió tomar una muestra que sea representativa de todo el periodo, procurando contar con un enlace por, al menos, cada trimestre desde enero 2007 hasta abril

2017.

3.2. Preprocesamiento de lo Datos

Como insumo de trabajo, requerimos lo que se conoce como la matriz documento-término que es un formato común para aplicaciones de aprendizaje de máquinas. En el transcurso de este trabajo utilizaremos terminología común en enfoques de minería de texto, como por ejemplo[12]:

- String: los datos de texto usualmente se leen en memoria en este formato, en el lenguaje de programación que utilizamos para este trabajo—R—los textos se leen en vectores de caracteres.

- Token: es una unidad de texto significativa y de interés para el análisis. Usualmente se refiere a una palabra; sin embargo, en algunas aplicaciones pueden ser n-gramas o frases. La tokenización es el proceso de dividir el texto en tokens.

- Corpus: estos objetos contienen strings sin procesar e incluyen metadata adicional y detalles de caracterización.

- Matriz documento-término: es una matriz dispersa que describe una colección de documentos, un corpus, con una fila para cada documento y una columna para cada término. El cuerpo de la matriz registra la ocurrencia de una palabra o el índice tf-idf.

La estructura genérica de una base de datos organiza las variables que se quieren medir en las columnas y ubica a los individuos a los que se observará en las filas. En minería de texto queremos transformar un documento a un vector de frecuencias de términos. Para ello, primero se debe definir con qué *token*[9] deseamos trabajar. Un *token* es una unidad de texto de importancia para el análisis y la *tokenización* se refiere al proceso de dividir el texto en *tokens*. Podemos trabajar con unigramas, por ejemplo, donde cada *token* está conformado

por una palabra. Para otras aplicaciones puede ser relevante trabajar con bigramas, es decir, *tokens* de dos palabras. En general en minería de texto, un *token* puede ser palabras individuales, n-gramas, oraciones o párrafos.

Después de identificar el *token* con el que se desea trabajar, se debe preparar el texto para el análisis eliminando las mayúsculas, removiendo puntuación, números y palabras vacías y omitiendo las palabras que tiene menos letras que un mínimo establecidos previamente. Hay quienes sugieren que parar Modelamiento de Tópicos es necesario que las palabras se transforme a raíces [9][10]. Así por ejemplo, *hacer*, *hacía* y *hacemos* en lugar de ser tres términos separados se unirían en una sola raíz *hac*. Sin embargo, hay quienes considerar que transformar las palabras a raíces no siempre añade valor. Los algoritmos que se utilizan para este propósito en algunos casos combinan términos que deberían considerarse como distintos en un contexto. No transformar las palabras a raíces no inserta errores preocupantes ya que esperamos que las variaciones de una palabra terminen en un mismo tópico[11].

Para el objeto de este curso es suficiente trabajar con unigramas, si bien conocemos por experiencia que se puede enriquecer el análisis con n-gramas que sabemos existen, se ha decidido empezar con el caso más simple que no imponga ningún supuesto. Para futuros trabajos, se puede experimentar con n-grama o incluso frase.

Para analizar los datos de mejor manera es necesario eliminar la puntuación del texto, convertir todo el texto a minúsculas, eliminar los espacios que separan los párrafos y eliminar las palabras "vacías". Las palabras "vacías" son palabras que tiene una función gramatical pero, no añaden valor en términos de contenido (un, uno, una, le, la los). En este trabajo se utilizó una base de datos de palabras vacías y el comando propio de R para este fin. Además del diccionario de palabras vacías, se incorporaron una lista adicional de palabras que arrojaban frecuencias altas pero no aportaban significado.

A continuación describimos en detalle las metodologías que se utilizaron en este trabajo.

4. Metodología

4.1. Análisis de Sentimientos

Una vez que hemos explorado la estructura general del corpus analizando la frecuencia del uso de palabras, podemos incursionar en el análisis de sentimientos. Cuando los seres humanos leemos un texto o escuchamos un discurso utilizamos nuestro conocimiento sobre la intención emocional de las palabras para inferir si una porción de discurso tiene una intención positiva o negativa.

Este enfoque entra dentro de los que conocemos como **métodos de codificación basada en diccionarios**, en este caso utilizaremos este método para encontrar la percepción general de un enlace sumando y promediando los puntajes de sentimientos para cada palabra que aparece en el enlace y los diccionarios. Para ello, requerimos encontrar diccionarios que sirvan para este propósito. En este trabajo, utilizamos dos diccionarios que se basan en unigramas. A estos unigramas se les ha asignando una categoría (positivo-negativo ó agrado-activación-imaginabilidad), según el criterio de voluntarios que cooperaron con la valoración.

El primer diccionario que utilizamos es el Diccionario de Afectos en Español [7], un léxico de 2880 palabras más comunes en español que recibieron un puntaje manual por 662 voluntarios. Cada voluntario recibió alrededor de 20 palabras y puntuó en tres dimensiones afectivas: agrado, activación e imaginabilidad, utilizando una escala discreta con tres niveles. Los posibles valores para cada una de las tres dimensiones se presentan en el Cuadro 1.

Cuadro 1: Valores posibles para las tres dimensiones afectivas

	Agradable	**Activación**	**Imaginabilidad**
1	Desagradable	Pasivo	Difícil de imaginar
2	Ni agradable, ni desagradable	Ni activo, ni pasivo	Ni difícil, ni fácil de imaginar
3	Agradable	Activo	Fácil de imaginar

La base de datos final contiene el promedio de las valoraciones de los voluntarios por cada dimensión afectiva. Es decir, cada palabra en léxico tiene un puntaje de agrado, activación e imaginabilidad. En general, las palabras de léxico presentan un 2.23 en promedio de agrado, 2.33 en activación y 2.55 en imaginabilidad, como se presenta en el cuadro a continuación. Gravado[7] demuestra que las correlaciones entre los puntajes de las dimensiones son débiles, esto implica que el agrado, la activación y la imaginabilidad son dimensiones afectivas independientes.

Cuadro 2: Estadísticas descriptivas de las valoraciones del léxico

	Media	**SD**	**Skewness**	**Kurtosis**
Agradable	2.23	0.47	-0.47	-0.06
Activación	2.33	0.48	-0.28	-0.84
Imaginabilidad	2.55	0.42	-0.90	0.18

Cuadro 3: Correlaciones entre las dimensiones afectivas del léxico

	Agradable	**Activación**	**Imaginabilidad**
Agradable	1.00	0.14	0.10
Activación		1.00	0.11
Imaginabilidad			1.00

Finalmente, se presenta las cinco palabras que recibieron los puntajes más altos y más bajos en cada dimensión afectiva.

Cuadro 4: Palabras del léxico con valoración más altas y más bajas

Agradable		**Activación**		**Imaginabilidad**	
+	-	+	-	+	-
jugar	asesinato	idea	yacer	sucio	consistir
beso	caro	publicar	espiritual	silencio	constar
sonrisa	ahogar	violento	quieto	dar	morfología
compañía	herida	sexual	esperar	pez	piedad
reir	cigarro	talento	cadáver	pensar	tendencia

El segundo diccionario que utilizamos es el léxico de sentimientos en español desarrollado por Perez, Barnea y Mohalcea[6]. Este diccionario atribuye una cualidad de negativo o positivo partiendo de diccionarios desarrollados para este mismo propósito en otros lenguajes, inglés en este caso. Para crear el diccionario, las autoras utilizan herramientas manuales (codificación) y automáticas disponibles para interpretar el sentimiento otorgado al léxico en inglés a su equivalente en español. La base de datos disponible para este trabajo no cuenta con todas las entradas clasificadas en español. Como aproximación, para aquellas palabras que no han sido clasificadas en español, tomamos la clasificación en inglés.

Al interpretar estos resultados es importante tener en cuenta que no todas las palabras aparecen en los diccionarios y que el puntaje de sentimiento no considera el contexto; es decir, no analiza calificadores que aparecen antes o después que podrían alterar el puntaje de sentimiento. En particular, la falta de contexto hace que sea imposible evaluar el sarcasmo. El sarcasmo es un recurso muy utilizado por Correa en su discurso, como tal, la aproximación del sentimiento de sus textos que ofrecemos tendrá la limitación de no puede identificar este recurso.

4.2. Modelamiento de Tópicos

El modelamiento de tópicos es el algoritmo de agrupamiento probabilístico más popular. La idea central es crear un modelo generativo probabilístico para el corpus[10] que descubra los temas subyacente del texto. Para ello, asumimos que las palabras en un documento se generan a partir de un tema que asigna probabilidades a las palabras que se mencionarán en dicho tema con más y menor frecuencia.

Los modelos de temas se construyeron sobre y ampliaron el alcance de métodos clásicos del procesamiento de lenguaje porque permiten la afiliación múltiple; es decir, no asumen que los documentos pertenecen a un único tema, sino pertenecen a varios temas y la distribución de temas varía entre documentos[10]. Los modelos de temas ajustan modelos probabilísticos a la frecuencia de ocurrencias de palabras en un documento. Este modelo ajustado se puede utilizar para estimar la similitud entre documentos y entre un conjunto de palabras claves utilizando una estrato adicional de variables latentes o temas[10].

Estos modelos se basan en un supuesto de intercambiabilidad de las palabras de un documento; es decir, estos modelos asumen que la información contenida en el orden de ocurrencia de las palabras es poco significativa y como tal, solo se utiliza las frecuencias de las palabras. Esto lleva a una representación simplificada del texto utilizada comúnmente en algoritmos de procesamiento de lenguajes naturales conocida como modelo de lista de palabras[13].

El modelo de lista de palabras representa un documento como una colección de sus palabras, sin tener en cuenta la gramática ni el orden de las palabras; pero, mantiene la repetición. El modelo de lista de palabras se utiliza comúnmente en métodos de clasificación de documentos ya que la frecuencias de las palabras es una buena características para entrenar la clasificación[13].

Los modelos de temas se construyeron sobre y ampliaron el alcance de métodos clásicos del procesamiento de lenguaje porque permiten la afiliación múltiple; es decir, no asumen que los documentos pertenecen a un único tema, sino pertenecen a varios temas y la distribución

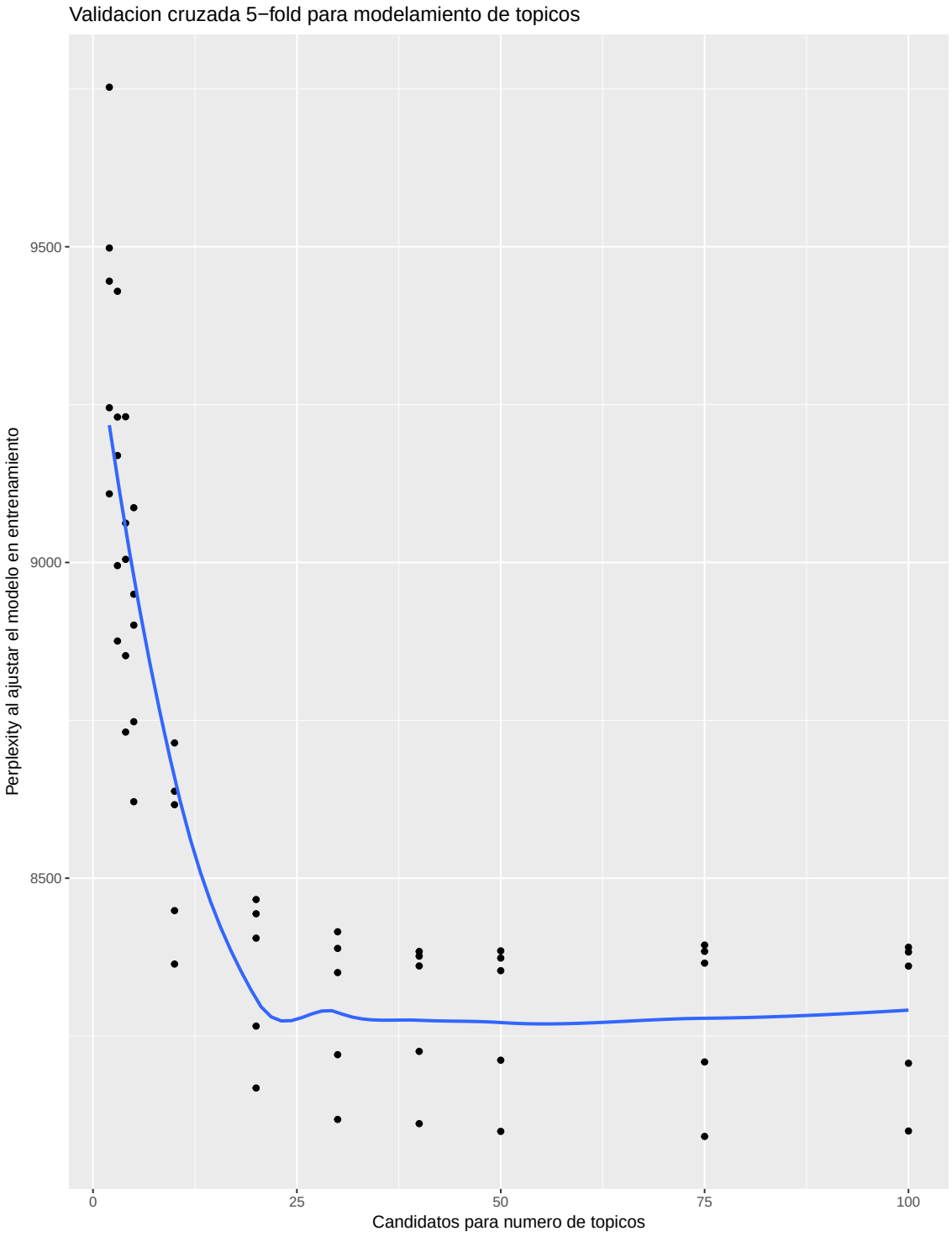
de temas varía entre documentos.

Los modelos de temas ajustan modelos probabilísticos a la frecuencia de ocurrencias de palabras en un documento. Este modelo ajustado se puede utilizar para estimar la similitud entre documentos y entre un conjunto de palabras claves utilizando una estrato adicional de variables latentes o temas.

Tanto LDA como CTM requieren fijar el número de tópicos para iniciar el algoritmo. Para determinar el número óptimo de tópicos, se corrió una validación cruzada 5-fold y se calculó el indicador *perplexity* para cada uno de estos modelos. Esto dio como resultado que el número de temas que minimiza *perpelxity* es 20. Con este número de temas se iniciaron todos los algoritmos.

Adicionalmente, en la limpieza de palabras se conservó solo aquellos término cuyo indicador *sparcity* fue mayor a 0.25. *Sparcity* es un indicador que mide la proporción de textos en los que aparece una palabra. La inspección visual de las frecuencias de palabras en el vocabulario evidenció que existe una gran cantidad de palabras que aparecen una sola vez. La mayoría de estas palabras son errores de tipeo, por ello, se decide eliminarlas del vocabulario para evitar una corrección manual.

Figura 4: Validación cruzada 5-fold para modelamiento de tópicos



4.2.1. Latent Dirichlet Allocation[12]

Esta sección resume el trabajo presentado por David Blei, Andrew Ng y Michael Jordan en su ensayo *Latent Dirichlet Allocation* (2003)[12].

El modelo LDA Latent Dirichlet Allocation es un modelo probabilístico generativo del un corpus. El modelo se basa en la idea que el documento es la representación de una combinación aleatorio de temas latentes, dónde cada tema se caracteriza por un distribución de palabras. LDA asume que cada documento del corpus se generó siguiendo el siguiente proceso:

1. Escoja la cantidad de palabras N de un texto, dónde $N \sim Poisson\left(\xi\right)$
2. Escoja el perfil de uso de palabras θ más frecuentes para su texto, dónde $\theta \sim Dir\left(\alpha\right)$
3. Para cada una de las N palabras w_n:
   a) Escoja un tema z_n, dónde $z_n \sim Multinomial\left(\theta\right)$
   b) Escoja una palabra w_n de la probabilidad multinomial condicionado en z_n:

$$p\left(w_n|z_n,\beta\right)$$

Una variable aleatoria Dirichlet de k-dimesiones, θ, puede tomar cualquier valor que esté contenido en un $(k-1)$-símplex si $\theta_i \geq 0$ y $\sum_{i=1}^{k}\theta_i = 1$. La densidad de probabilidad dentro de este símplex está descrita por la siguiente función:

$$p\left(\theta|\alpha\right) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma\left(\alpha_i\right)}\theta_1^{\alpha_1-1}\ldots\theta_k^{\alpha_k-1} \tag{4}$$

dónde el parámetro $\boldsymbol{\alpha}$ es un vector cuyos componentes $\alpha_i > 0$ con $i = 1,\ldots,k$ y $\Gamma\left(x\right)$ es la función Gamma. Dirichlet es una función conveniente dentro de un símplex porque están en la familia exponencial, tiene estadísticos suficientes de dimensión finita y es conjugada de la distribución multinomial. Estas propiedades facilitan la estimación de parámetros e inferencia necesarios para ejecutar los algoritmos de LDA.

Dados los parámetros α y β, la distribución conjunta de la mezcla de temas, θ, el conjunto de N temas $\boldsymbol{z}$ y el conjunto de N palabras $\boldsymbol{w}$ está dada por:

$$p\left(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta\right) = p\left(\theta|\alpha\right) \prod_{n=1}^{N} p\left(z_n|\theta\right) p\left(w_n|z_n, \beta\right) \tag{5}$$

dónde $p\left(z_n|\theta\right)$ toma el valor de θ_i para el único i tal que $z_n^i = 1$. Si integramos sobre θ y sumamos sobre z, obtenemos la distribución marginal de un documento:

$$p\left(w|\alpha, \beta\right) = \int p\left(\theta|\alpha\right) \left(\prod_{n=1}^{N} \sum_{z_n} p\left(z_n|\theta\right) p\left(w_n|z_n, \beta\right) \right) d\theta \tag{6}$$

Finalmente, si calculamos el producto de las probabilidades marginales de los documentos individuales obtenemos la probabilidad del corpus:

$$p\left(D|\alpha, \beta\right) = \prod_{d=1}^{M} \int p\left(\theta_d|\alpha\right) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p\left(z_{dn}|\theta_d\right) p\left(w_{dn}|z_{dn}, \beta\right) \right) d\theta_d \tag{7}$$

La representación de LDA tiene tres niveles. Los parámetros α y β son parámetros que pertenecen al nivel del corpus. Las variables θ_d son variables que pertenecen al nivel de los documentos, muestreadas una vez por cada documento. Finalmente, tenemos las variables z_{dn} y w_{dn} son variables del nivel de las palabras y se muestrean una vez por cada palabra en cada documento. A diferencia de otros métodos de agrupamiento, LDA no restringe la pertenencia de un documento a un solo tema. Dado que LDA involucra tres niveles, dónde uno de ellos implica tomar muestras de manera repetitivas de los temas dentro de un documento. Entonces, en este modelo los documentos pueden asociarse a múltiples temas. En resumen, LDA plantea que cada palabra de los documentos observados y no observados se generan al escoger aleatoriamente un tema, que a su vez se extrae de una distribución cuyo parámetros se escoge al azar. Este parámetro se selecciona aleatoriamente de una distribución uniforme sobre el símplex de temas.

4.2.2. LDA: Inferencia[12]

El problema de inferencia que se debe resolver para utilizar LDA es el cálculo de la distribución a posteriori de las variables desconocidas, dado un documento:

Esta distribución nos se puede calcular en general. Si bien la distribución a posteriori no se puede resolver para obtener una inferencia exacta, se puede utilizar el algoritmo de inferencia variable basado en la convexidad para aproximarnos a una solución. Este algoritmo utiliza la desigualdad de Jensen para obtener una cota inferior ajustable del algoritmo de la verosimilitud (Jordan et al., 1999). Consideremos una familia de cotas inferiores, indexadas en un conjunto de parámetros variables que se escogen mediante un proceso de optimización cuyo objetivo es encontrar la cota inferior más apretado.

Una manera simple de encontrar un familia manipulable de cotas inferiores es removiendo los nodos y la aristas del modelo gráfico antes presentado. Si consideramos este gráfico, podemos notar que hay una conexión problemática entre θ y β que se origina por las aristas que conectan a θ, $\boldsymbol{z}$ y $\boldsymbol{w}$ entre sí. Si eliminamos estas aristas y también los nodos $\boldsymbol{w}$, el resultado es una representación gráfica más simple con parámetros de variación libre y podemos obtener una familia de distribuciones de variables latentes. Esta familia está caracteriza por la siguiente distribución de variación.

$$q\left(\theta, \boldsymbol{z}|\gamma, \phi\right) = q\left(\theta|\gamma\right) \prod_{n=1}^{N} q\left(z_n|\phi_n\right) \tag{8}$$

dónde el parámetro de Dirichlet γ y los parámetros multinomiales $(\phi_1, \ldots, \phi_N)$ son los parámetros de variación libres.

Ahora que tenemos una familia simplificada de distribuciones de probabilidad , el siguiente paso es definir el problema de optimización para determinar los valores de los parámetros de variación γ y ϕ. El problema de optimización para encontrar la cota inferior más estrecha es:

$$(\gamma*, \phi*) = \underset{(\gamma,\phi)}{argmin} D\left(q\left(\theta, \boldsymbol{z}|\gamma, \phi\right) \parallel p\left(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta\right)\right) \tag{9}$$

La optimización de los valores de los parámetros de variación se encuentran minimizando la divergencia de Kullback-Leibler entre la distribución de variación y la distribución a posteriori real $p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)$. Esta minimización se puede obtener utilizando el método del punto fijo iterativo. Calculando las derivadas de la divergencia KL e igualándolas a cero, se obtiene el siguiente par de ecuaciones actualizadas:

$$\phi_{ni} \propto \beta_{iw_n} exp\left\{E_q\left[\log \theta_i|\gamma\right]\right\} \tag{10}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{11}$$

Se puede demostrar que la esperanza de la actualización multinomial se puede calcular de la siguiente manera:

$$E_q\left[\log\left(\theta_i\right)|\gamma\right] = \Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) \tag{12}$$

dónde Ψ es la primera derivada de la función $\log\Gamma$ que se calcula mediante aproximaciones de Taylor (Abramowitz y Stegun, 1970).

Es importante notar que la distribución de variación es en realidad una distribución condicional que varía como una función de $\boldsymbol{w}$. Esto ocurre porque el problema de optimización en la ecuación 6 se conduce para $\boldsymbol{w}$ fija y como tal, da como resultado los parámetros de optimización (γ^*, ϕ^*) que son una función de $\boldsymbol{w}$. La función de variación resultante está dada por $q\left(\theta, \boldsymbol{z}|\gamma^*\left(\boldsymbol{w}\right), \phi^*\left(\boldsymbol{w}\right)\right)$, en dónde se expresa la dependencia en $\boldsymbol{w}$ de manera explícita. Entonces, la distribución de variación se puede ver como una aproximación de la distribución a posteriori $q\left(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta\right)$.

En el lenguaje del texto, los parámetros de optimización $\gamma^*\left(w\right), \phi^*\left(w\right)$ son específicos a los documentos. En particular, los parámetros de Dirichlet $\gamma^*\left(w\right)$ dan una representación de un documento en el símplex de temas.

Finalmente, se resume el procedimiento de inferencia variacional, con valores de inicialización adecuados para γ y ϕ_n. Queda claro del pseudocódigo que cada iteración de la

inferencia variacional de LDA requiere de $O((N+1)k)$ operaciones. Se puede demostrar empíricamente que el número de iteraciones que se requieren para un solo documento está en el orden del número de palabras en el documento. Esto da como resultado un número de operaciones en el orden de N^2k. El pseudocódigo para este procedimiento es:

1. Inicializar $\phi_{ni}^0 := 1/k$ para toda i y n
2. Inicializar $\gamma_i := \alpha_i + N/k$ para toda i
3. Repetir
4. for $n = 1$ hasta N
5. for $i = 1$ hasta k
6. $\phi_{ni}^{t+1} := \beta_{iw_n} exp(\Psi(\gamma_i^t))$
7. normalizar ϕ_{ni}^{t+1} para que sume 1
8. $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi_n^{t+1}$
9. Hasta convergencia

4.2.3. LDA: Estimación de Parámetros[12]

Dado el un corpus de documentos, $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \boldsymbol{w}_M\}$, nos interesa encontrar los parámetros α y β que maximicen el logaritmo de la verosimilitud de los datos:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\boldsymbol{w}_d|\alpha, \beta)$$

Como sabemos, $p(\boldsymbol{w}_d|\alpha, \beta)$ no se puede calcular; sin embargo, la inferencia variacional nos da una cota inferior para el logaritmo de la verosimilitud y esta cota se puede maximizar con respecto a α y β. Entonces, podemos encontrar estimaciones empíricas para el modelo LDA alternando entre el procedimiento variacional EM (Expectation-Maximization) que maximiza la cota inferior con respecto a los parámetros variacionales γ y ϕ, y después, para

los valores fijo de los parámetro variacionales, se maximiza la cota inferior con respecto a los parámetros del modelo α y β. Para ejecutar el algoritmo variacional EM se debe llevar a cabo el siguiente procedimiento:

1. (Esperanza) Para cada documento, encuentre los valores de optimización para los parámetros variacionales.

2. (Maximización) Maximice la cota inferior del logaritmo de la verosimilitud con respecto a los parámetros del modelo α y β. Esto corresponde a encontrar los estimadores de máxima verosimilitud estadísticos suficientes para cada documento bajo la distribución a posteriori apropiada que se calcula en el paso de la Esperanza.

Estos dos pasos se repiten hasta que la cota inferior del logaritmo de la verosimilitud converja.

4.2.4. Algoritmo CTM[15]

Esta sección resume el trabajo presentado por David Blei y John Lafferty en su ensayo *Correlated Topic Models* (2015)[15].

El modelo LDA asume que las palabras de cada documento se originan de una mezcla de temas, cada uno de los cuáles es una distribución sobre el vocabulario. Una limitación de LDA es que no permite modelar la correlación entre temas. Esta limitación se origina del uso de la distribución Dirichlet para modelar la variación entre las proporciones de los temas. Para evitar este inconveniente, Blei y Lafferty (2005) desarrollaron un modelo que utiliza una distribución log-normal para describir la variación entre las proporciones de los temas.La distribución log-normal es una distribución en el símplex que permite un patrón general de variabilidad entre los componentes, transformando una variable aleatoria normal multivariante. Considerando una parametrización natural de una distribución multinomial K-dimensional:

$$p\left(z|\eta\right) = exp\left\{\eta^T z - a\left(\eta\right)\right\} \tag{13}$$

La variable aleatoria Z puede tomar K valores distintos; está representada por un vector

K-dimensional con un componente igual a uno para indicar un valor dentro de la lista . La función generadora acumulativaes:

$$a(\eta) = \log\left(\sum_{i=1}^{K} exp\{\eta_i\}\right) \tag{14}$$

El mapeo entre la parametrización promedio (símplex) y la parametrización natural es:

$$\eta_i = \log \theta_i/\theta_K \tag{15}$$

La distribución log-normal asume que η tiene una distribución normal y que luego se mapea al símplex con la función inversa de la ecuación 12, es decir, $f(\eta_i) = exp\eta_i/\sum_j exp\eta_j$. La distribución log-normal puede representar las correlaciones entre componentes de una variable aleatoria simplicial a través de la matriz de covarianza de la distribución normal. La distribución log-normal se estudió originalmente en la observación de datos constitutivos tales como las proporciones de minerales en muestras geológicas. Blei y Lafferty (2005) extienden este uso a un modelo jerárquico para describir la composición latente de temas asociados a cada documentos.

Sean $\{\mu, \Sigma\}$ la media y matriz de covarianza, y sea $\beta_{1:K}$ K temas multinomiales sobre el vocabulario fijo. El modelo de temas correlacionados asume que un documento con N palabras se genera de acuerdo al siguiente procesos:

1. Seleccione $\eta|\{\mu, \Sigma\} \sim (\mu, \Sigma)$
2. Para $n \in \{1, \ldots, N\}$:
    a) Seleccione la asignación de temas $Z_n|\eta$ de $Mult(f(\eta))$.
    b) Seleccione las palabras $W_n|\{z_n, \beta_{1:K}\}$ de $Mult(\beta_{z_n})$.

Cabe señalar que este procesos de generación de documentos es idéntico al proceso de generación de LDA, la única diferencias que las proporciones de los temas se seleccionad de

una log-normal en lugar de una distribución de Dirichlet. Esta característica le permita a CTM ser más expresivo que LDA, ya que el supuesto de independencia que implica es uso de Dirichlet no es realista cuando analizamos documentos de un colección

4.2.5. CTM: Inferencia a posteriori y estimación de parámetros[15]

La inferencia a posteriori es el reto central en CTM. La distribución a posteriori de las variables latentes condicionadas en un documento, $p(\eta, z_{1:N}|w_{1:N})$, no se puede calcular ya que una vez que se condiciona sobre observaciones, las asignaciones de temas $z_{1:N}$ y el logaritmo de las proporciones η son dependientes. Blei y Lafferty (2007) utilizan métodos de campo promedio variacional para obtener una aproximación de esta distribución a posteriori.

En modelos gráficos que utilizan pares y mezclas de familias de exponenciales conjugadas, el algoritmo de inferencia variacional se puede obtener a partir de los principios generales. Sin embargo, en el modelo CTM la distribución log-normal no es conjugada de la multinomial. Como consecuencia, Blei y Lafferty (2007) deben encontrar un algoritmo de inferencia variacional que toma en cuenta esta particularidad de CTM. Empezamos por usar la desigualdad de Jensen para acotar el logaritmo de la probabilidad de un documento:

$$\log p(w_{1:N}|\mu, \Sigma, \beta) \geq$$

$$E_q[\log p(\eta|\mu, \Sigma)] + \sum_{n=1}^{N} E_q[\log p(\eta|\mu, \Sigma)] + E_q[\log p(w_n|z_n, \beta)] + H(q) \tag{16}$$

Aquí, la esperanza se calcula con respecto a la distribución variacional de las variables latentes, y $H(q)$ representan la entropía de la distribución. Utilizamos la distribución factorizada:

$$q\left(\eta_{1:K}, z_{1:N}|\lambda_{1:K}, v^2_{1:K}, \phi_{1:N}\right) = \prod_{i=1}^{K} q\left(\eta_i|\lambda_i, v_i^2\right) \prod_{n=1}^{N} q\left(z_n|\phi_n\right) \tag{17}$$

La distribución variacional de las variables discretas $z_{1:N}$ se especifican por los parámetros multinomiales de dimensión K $\phi_{1:N}$. La distribución variacional de las variables continuas

$\eta_{1:K}$ son K Gaussianas univariantes independientes $\{\lambda_i, v_i\}$. Dado que los parámetros variacionales se ajustan utilizando un solo documento observable $w_{1:N}$, introducir un matriz de varianza-covarianza no diagonal no representa un ventaja. El hecho que el logaritmo de la probabilidad no sea conjugada presenta una dificultad al calcular el valor esperado del logaritmo de la probabilidad de una asignación a un tema:

$$E_q\left[\log p\left(z_n|\eta\right)\right] = E_q\left[n^T z_n\right] - E_q\left[\log\left(\sum_{i=1}^{K} exp\left\{\eta_i\right\}\right)\right] \tag{18}$$

Para mantener la cota superior del logaritmo de la probabilidad, se acota por abajo el negativo de la log-normal utilizando una expansión de Taylor:

$$E_q\left[\log\left(\sum_{i=1}^{K} exp\left\{\eta_i\right\}\right)\right] \leq \zeta^{-1}\left(\sum_{i=1}^{K} E_q\left[exp\left\{\eta_i\right\}\right] - 1 + \log\left(\zeta\right)\right) \tag{19}$$

aquí hemos introducido un nuevo parámetro de variación ζ. La esperanza es el promedio de una distribución log normal con media y varianza que vienen de los parámetros variacionales. Dado un modelo y un documentos,el algoritmo de inferencia variacional optimiza la ecuación 4 con respecto a los parámetros variacionales. Dada una colección de documentos o corpus, se estima los parámetros en el modelo de temas correlaciones al intentar maximizar la verosimilitud de un corpus de documentos como un función de los temas $\beta_{1:k}$ y la distribución Gaussiana multivariante (μ, Σ). Para ello, se utiliza el enfoque de maximización de la esperanza variacional (VEM-variational expectation maximization), en el que se maximiza la cota del logaritmo de la probabilidad de un colección de documentos sumando la ecuación 4 sobre todos los documentos.

El paso de la esperanza o paso-E maximiza la cota con respecto a los parámetros variacionales, esto se logra aplicando la inferencia variacional para cada uno de los documentos. En el paso de maximización o paso-M, maximizamos la cota con respecto a los parámetros del modelo. Esto es la estimación de máxima verosimilitud de los temas y la Gaussiana multivariante utilizando estadísticos suficientes, dónde la esperanza se calcula con respecto a la

distribución variacional calculada en el paso-E. El paso-E y el paso-M se repiten hasta que la cota de la verosimilitud converge.

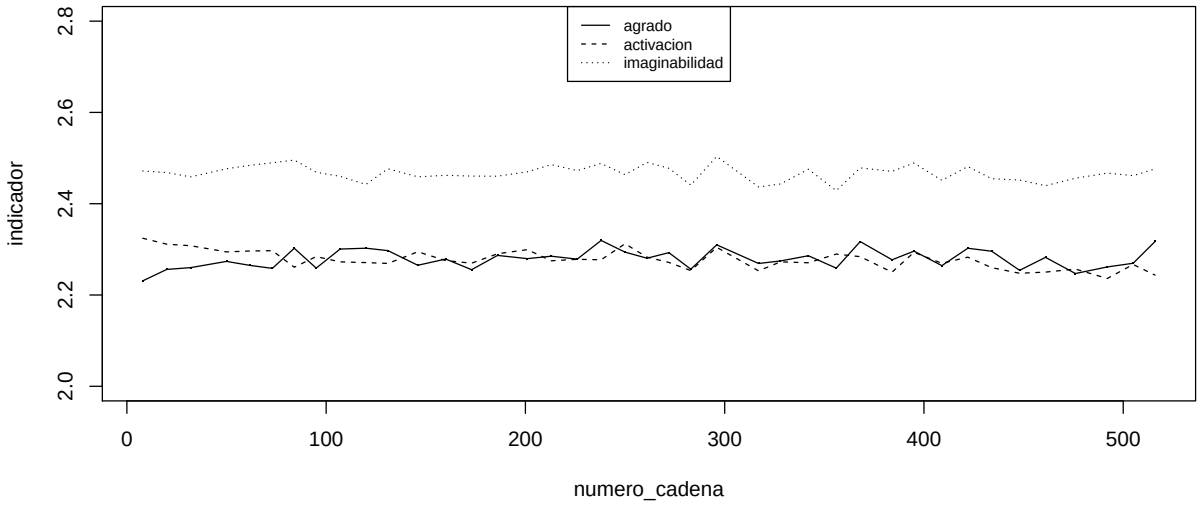
5. Resultados

5.1. Análisis de Sentimientos

El Análisis de Sentimientos se utiliza para determinar la intención del emisor de un mensaje, ya sea verbal o escrito. Para este trabajo, hemos utilizado dos diccionarios. El primero, mide el agrado, imaginabilidad y activación de un texto. El segundo, mide la positividad o negatividad del discurso.

La figura 4 presenta los resultados del análisis de sentimientos de los 40 Enlaces Ciudadanos transcritos para este trabajo. Como se puede apreciar en el gráfico, en general los enlaces ciudadanos se acercan más a puntajes neutros de agrado y activación, es decir, los enlaces ciudadanos no son ni agradables, ni desagradables y tampoco son ni activos, ni pasivos. Por otro lado, la imaginabilidad de los enlaces ciudadanos se aleja más del puntaje neutral; es decir, el lenguaje que se utiliza en los enlaces ciudadanos se acerca más a ser fácil de imaginar, que neutral (ni fácil, ni difícil de imaginar). De entre las 3 características que mide este diccionario, tal vez la más importante es la imaginabilidad. Es crucial que un discurso dirijo a multitudes sea fácil de imaginar, caso contrario corre el riesgo de ser irrelevante para el ciudadano promedio.

Figura 5: Agrado, activación e imaginabilidad de los Enlaces Ciudadanos

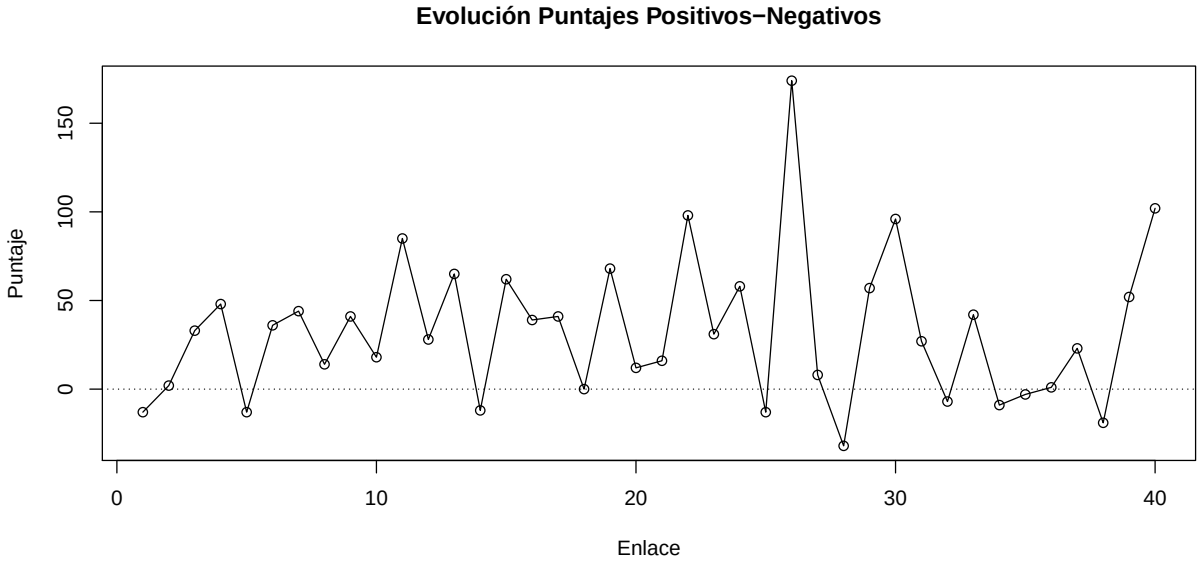


Por otro lado, en cuanto a la positividad o negatividad el lenguaje utilizado para los discursos, tenemos que 9 discursos tienen más atributos negativos que positivos, siendo el enlace número 250, de 17 de diciembre de en que tiene más atributos negativos de todo el corpus. Según El Ciudadano [F], los principales temas que se trataron en este enlace fueron:

- Gobierno ratificó a los migrantes en su día
- Jefe de Estado augura éxitos a los equipos que jugarán la final del campeonato nacional de fútbol
- Correa invita a visitar la Amazonía en las fiestas de Navidad y Fin de Año
- Gobierno prioriza entrega de recursos a área social en lugar de pagar deuda externa
- Correa rechaza acusaciones del asambleista César Montúfar contra su abogado
- La propuesta del Gobierno Nacional con el registro de marcas, frases y slogans no responde a intereses comerciales

- Gobierno da ultimátum a la minería ilegal en las localidades de Esmeraldas, Zamora Chinchipe, Morona Santiago y Napo
- La resolución del Tribunal Contencioso Electoral ante demanda de Vistazo es inconstitucional. Vistazo fue acusada de haber incumplido el silencio electoral durante la Consulta Popular del 7 de mayo 2011
- El Presidente rechazó la información *tergiversada* sobre la inauguración del tramo del tren Salinas-Ibarra
- El asesor presidencial, Francisco Latorre aclaró que acudió a la Fiscalía por motivos personales, no por orden presidencial
- Correa dice que los medios llaman protesta social a la violencia social

Figura 6: Positividad y Negatividad de los Enlaces Ciudadanos



Por otro lado, el enlace ciudadano 342, de 5 de octubre 2013, tiene más atributos positivos que negativos y es el enlace con puntaje positivo más alto de todos en el corpus. Según el

sitio web del Enlace Ciudadano [F], los tema principales que se trataron en aquella ocasión fueron:

- Moradores de Intag repaldan poryecto minero Llurimagua
- Nuevos delitos se incorporarán en el Código Integral Penal
- Una nueva forma de golpismo existe en América Latina
- Playas de Cuyabeno abre el camino hacia la nueva Amazonía
- Obras en Lago Agrio transformarán la vida en la Amazonía
- Gobierno invita a opositores y periodistas a corroborar contaminación causada por Chevron
- Nuevos delitos se incorporarán en el Código Integral Penal

Si bien es útil conocer el sentimiento general del texto, esta estrategia resultaría más útil si es posible identificar secciones específicas del texto. Dado que en este momento no tenemos una estrategia para hacerlo, el análisis debe limitarse al texto en general o secciones arbitrarias, sin posibilidad de identificar secciones del discurso que hablan de un tema en particular.

5.2. Modelamiento de Tópicos

Se corrieron cuatro modelos para las tres opciones de LDA y uno para CTM, con 20 tópicos por descubrir y *sparcity* de 97.5 %.

Después, se inspeccionó visualmente los cuatro modelos, se etiquetó los tópicos que era posible etiquetar y finalmente, se decidió cual de estos cuatro modelos representaba de mejor manera el discurso de Correa. Para inspeccionar las palabras claves de los tópicos y las etiquetas, diríjase a los anexos.

El modelamiento de tópicos correlaciones no arrojó resultados satisfactorios. El modelo ajustó los 20 tópicos con igual probabilidad para todas las sabatinas; es decir, cada enlace tiene un 0.5 de probabilidad de pertenecer a cualquiera de los temas. Estos no es útil porque lo que estamos buscando son categorías que discriminen entre los enlaces. Nuestra sospecha se confirma al inspeccionar los términos que conforman los temas, y podemos constatar que son tópicos muy similares.

LDA con VEM dos temas para cada enlace, en casi todos los casos. El primer tema es el predominante en la sabatina y la prababilidad de que un enlace pertenezca a ese tópico oscila entre 0.88 y 0.99. La probabilidad restante suele concentrarse en un segundo tema secundario y la probabilidad de pertenencia oscila entre 0.8 y 0.15. La probabilidad restante usualmente se reparte en el resto de tópicos con probabilidades muy pequeñas, demasiado cercanas a 0 como para considerarlas relevantes. Este patrón se repite para LDA con VEM y α fijo. Se concluye que LDA con VEM centra los excesivamente los tópicos en lo que se habló en la sabatina, y aunque útil, se requiere de temas más globales y menos concentrados en los detalles de una sabatina. Esto será particularmente problemática cuando se incluya más enlaces para hacer predicciones, si los enlaces son todos distintos, no se podrá confiar en la exactitud de la predicción.

Finalmente, LDA con muestreo de Gibbs arroja los resultados más prometedores. En primer lugar creo un tema que se ha etiquetado como *MISCELANEO* al cual pertenecen todos los enlaces ciudadanos con un probabilidad que oscila entre 0.35 y 0.51. Al inspeccionar los términos en este tópico, se puede ver con claridad que incluye palabras características y recurrentes en el discurso de Correa; es decir, se puede pensar en este tópico como el discurso común en todas los enlaces. Por ejemplo, se encuentra la palabra *patria, social, revolución, prensa, ciudadana, salud, derecho, medios, justicia, mala, familia, desarrollo, cambio, Dios, América, proyecto, noche, mentira, cuidado*, entre otras.

En segundo lugar, todos los enlaces cuentan con un segundo tema predominante y la probabilidad de pertenencia a dicho tema oscila entre 0.28 y 0.53. Cabe recalcar que, contrario

al caso de LDA con VEM, varios enlaces ciudadanos comparten temas. Finalmente, algunos enlaces cuentan con un tercer tema cuya probabilidad de asignación no supera el 0.15.

Por inspección, se determina que LDA con muestreo de Gibbs es el modelo que presenta mejor desempeño. Como tal, se propone las siguientes etiquetas para lo tópicos.

1. Amazonía - Chevron: EC32 23 % y EC342 42 %
2. Impuestos - subsidios: EC1 42 %, EC2 11 %, EC3 15 %, EC4 44 %, EC5 41 %, EC6 8 % y EC9 10 %
3. Lasso - obras públicas: EC505 35 % y EC516 45 %
4. Fuerzas Armadas: EC160 34 % y EC461 37 %
5. Educación Superior - pobreza: EC296 5 %, EC368 5 %, EC434 53 %, EC461 5 % y EC492 5 % y EC9 10 %
6. Policía - 30 S: EC201 33 %, EC213 30 % y EC238 31 %
7. Elecciones - economía: EC95 32 % y EC173 34 %
8. Terremoto Esmeraldas y Manabí: EC476 57 %
9. Política Exterior - derechos humanos: EC368 30 % y EC409 45 %
10. Universidades - invierno- agricultura: EC73 42 %, EC84 42 % y EC146 31 %
11. Maestros - libertad de expresión: EC283 37 % y EC317 38 %
12. Minería: EC20 36 %, EC186 33 % y EC250 29 %
13. Moneda nacional - obras públicas: EC107 37 % y EC120 44 %
14. Galápagos - Yasuní ITT: EC131 35 % y EC492 15 %
15. Snowden - espionaje - prensa: EC296 6 % y EC328 51 %

16. MPD - Día de los trabajadores: EC368 7 %, EC395 9 % y EC422 43 %
17. Macroeconomía: EC434 56 % y EC461 5 %
18. Utilidades: EC356 39 %, EC384 30 % y EC395 7 %
19. Presan - escándalo departamento Bélgica: EC146 6 %, EC226 31 %, EC261 29 %, EC272 40 % y EC296 29 %.
20. MISCELANEO: TODAS

6. Conclusiones y Recomendaciones

Después de hacer este análisis, hemos identificado que el presidente Correa utilizan palabras, en promedio, que no son ni activas, ni pasivas y tampoco son agradables, ni desagradables. Sin embargo, el lenguaje que utiliza se asocia con palabras que son, en promedio, más fáciles de imaginar; es decir, su discurso no es abstracto.

El análisis de sentimiento sería más informativo si utilizara segmentos particulares para el análisis. La interrogante es cómo identificar estos segmentos de una manera sistemática, pero con significado. Dividir el texto en partes iguales es una alternativa, pero no buena. La sugerencia es segmentar el texto manualmente y, posteriormente, conducir el análisis de sentimientos nuevamente.

Además, una limitación de este método es que no permite evaluar el sarcasmo, es decir, no es posible cambiar el puntaje de una palabra de acuerdo al contexto en el que aparece. Finalmente, se debe extender este análisis a n-gramas mayores que uno, para investigar el sentimiento de frases representativas del discurso.

Al interpretar estos resultados es importante tener en cuenta que no todas las palabras aparecen en los diccionarios y que el puntaje de sentimiento no considera el contexto; es decir, no analiza calificadores que aparecen antes o después que podrían alterar el puntaje de sentimiento. En particular, la falta de contexto hace que sea imposible evaluar el sarcasmo. El sarcasmo es un recurso muy utilizado por Correa en su discurso, como tal, la aproximación del sentimiento de sus textos que ofrecemos tendrá la limitación de no puede identificar este recurso.

El modelamiento de tópicos también presenta algunas limitaciones. En primer lugar, requiere de supuestos detallados sobre el proceso generador de documentos (modelo de bolsa de palabras) y cuidadosa especificación de los hiperparámetros. Gallagher et al presentaron un enfoque alternativo a modelamiento de tópicos que no asume un modelo generativo. En lugar de ello, produce tópicos utilizando un enfoque teórico de información. Este marco teórico se puede generalizar a un enfoque jerárquico y extensiones semi supervisadas, sin

supuestos adicionales[16].

Asímismo, vale la pena aplicar proceso de Dirichlet jerárquicos, la versión no paramétrica de LDA, ya que estos enfoques son capaces de calcular el número óptimo de tópicos. Al no tener esta restricción sobre los tópicos, puede tratar una cantidad infinita de tópicos. La limitación de estos modelos es que requieren del supuesto sobre el proceso generador de datos (modelo de bolsa de palabras)[17].

El modelamiento de tópicos con unigramas tiene limitaciones; sin embargo, no se puede generalizar y correr el mismo modelo con cualquier n-grama. Se sugiere identificar los n-gramas más relevantes, por ejemplo, *revolución ciudadana* y correr el modelo con una mezcla de n-gramas.

Finalmente, es recomendable conducir un estudio empírico de **lectura a profundidad** de los 40 enlaces ciudadanos que se utilizaron para este estudio. De esta manera, se podrá comparar los resultados del modelamiento de tópicos y mejorar las etiquetas de los tópicos.

7. Referencias

Referencias

[1] Comunicación personal con Secretaría Nacional de Comunicación, 19 de Mayo de 2017.

[2] Elaboración o cálculo propio.

[3] Plúa, M. L. (2014). *Enlace Ciudadano: Dispositivo de la Revolución Ciudadana* (Doctoral dissertation, Tesis para obtener título de maestría en Comunicación con mención en opinión pública). Facultad Latinoamericana de Ciencias Sociales, Quito, Ecuador).

[4] Zeas, S. (2014). El antes y después de Rafael Correa. Consultado diciembre, 13, 2017.

[5] Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1), 209-228.

[6] Perez-Rosas, V., Banea, C., Mihalcea, R. (2012, May). Learning sentiment lexicons in spanish. In *LREC* (Vol. 12, p. 73).

[7] Rıos, M. G. D. A., Gravano, A. (2013). Spanish DAL: A Spanish dictionary of affect in language. *WASSA 2013*, 21.

[8] Aaron Schumacher: Perplexity, what it is and what yours is, `http://planspace.org/2013/09/23/perplexity-what-it-is-and-what-yours-is/`. Consultado por última vez el 17 de diciembre de 2017.

[9] Hornik, K., Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13), 1-30.

[10] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *arXiv preprint arXiv:1707.02919*.

[11] Ramage, D., Rosen, E. (2011). Stanford topic modeling toolbox.

[12] Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

[13] Elkan, C. (2010). Text mining and topic models. *Lecture notes*.

[14] Greene, D., O'Callaghan, D., Cunningham, P. (2014, September). How many topics? stability analysis for topic models. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 498-513). Springer, Berlin, Heidelberg.

[15] Lafferty, J. D., Blei, D. M. (2006). Correlated topic models. *In Advances in neural information processing systems* (pp. 147-154).

[16] Gallagher, R. J., Reing, K., Kale, D., Steeg, G. V. (2016). Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge. *arXiv preprint arXiv:1611.10277*.

[17] Perotte, A. J., Wood, F., Elhadad, N., Bartlett, N. (2011). Hierarchically supervised latent Dirichlet allocation. *In Advances in Neural Information Processing Systems* (pp. 2609-2617).