ISSN 1870-4069

Análisis y seguimiento de tópicos en las conferencias matutinas del presidente de México

Luis Armando Arias-Romero[1], Gabriela Ramírez-de-la-Rosa[1], Esaú Villatoro-Tello[1,2]

[1] Universidad Autónoma Metropolitana, Unidad Cuajimalpa, México

[2] Idiap Research Institute, Martigny, Switzerland

ariasluis.ar@gmail.com, {gramirez, evillatoro}@cua.uam.mx

Resumen. El lenguaje es un recurso importante para los políticos. El análisis del lenguaje de políticos como el presidente de un país es una tarea importante para diferentes disciplinas como la lingüística, sociología, y comunicación. En este artículo presentamos un método que incorpora la detección automática de tópicos en 534 conferencias matutinas del presidente de México, Andrés Manuel López Obrador, utilizando LDA. Posteriormente, a través de un sistema web, al que denominamos DICTA, una persona puede visualizar de forma rápida los temas tratados en dichas conferencias. En la evaluación experimental, se obtuvo el valor más alto de coherencia al utilizar 18 tópicos.

Palabras clave: Detección y seguimiento de tópicos, LDA, discurso político, procesamiento del lenguaje natural.

Topic Analysis and Tracking from Mexico's President Daily Press Briefing

Abstract. Language is a very useful tool to politicians. Several areas such as linguistics, sociology and communication consider important the study of political discourse analysis. In this paper we present a method for topic detection using LDA in 534 daily press briefing of the Mexico's president: Andrés Manuel López Obrador. Subsequently, through a web system, which we call DICTA, a person can quickly view the topics discussed in these briefing. Through experimental evaluation we found the highest coherence value when using 18 topics.

Keywords: Topic detection and tracking, LDA, political discourse, natural language processing.

1. Introducción

El lenguaje humano es complejo y diverso. A través del lenguaje podemos expresar pensamientos, sentimientos, emociones, etc. [15]. En la política, el lenguaje puede usarse para convencer de una idea, cambiar la forma de pensar de una comunidad de personas, inspirar, pero también dividir. El análisis del lenguaje en el contexto político se ha estudiado durante muchos años por diversos campos como la lingüística, sociología, comunicación entre otros [16].

Dependiendo de la disciplina, el estudio del lenguaje político es diverso. Por ejemplo, se estudia el vocabulario utilizado (aspectos léxicos), la relación del léxico dentro de una oración (relación lexico-gramatical) [13], el estilo del discurso, entre otras características. El análisis del lenguaje en estos campos ayuda a entender fenómenos sociales, económicos y/o políticos [3, 13].

Usualmente, el análisis de este tipo de fenómenos no depende del análisis de un sólo discurso. Así, cuando se requiere estudiar textos políticos que abarcan más de un documento (o discurso), los especialistas se enfrentan con el volumen de textos a analizar [7].

El procedimiento manual involucra la lectura de todos los textos para realizar un estudio a fondo sobre temas que aborda y su influencia o correlación en otras áreas como la afectación (positiva o negativamente) en la economía, polarización [9], o consecuencias sociales [8, 6].

Adicionalmente, la importancia o influencia del emisor del lenguaje está relacionado con el impacto de dicho discurso; particularmente en la política. En México, se podría argumentar que el actor de más influencia política es el presidente de la República, actualmente Andrés Manuel López Obrador.

Desde el inicio de su gestión, en diciembre de 2018, López Obrador ha establecido una rutina de conferencias diarias donde presenta una variedad de asuntos que quiere comunicar al país. En estas conferencias, que se transmiten en vivo por YouTube y que se reportan en la mayoría de los noticieros de cobertura nacional, el presidente anuncia los programas sociales de su gobierno, da instrucciones a sus colaboradoras y envía mensajes políticos. Con frecuencia, en las conferencias participan funcionarios de su gobierno, según el tema que quiera abordar o para atender algún problema específico [10].

La regularidad de estas conferencias genera una gran cantidad de información. Información que deberá ser organizada y analizada tan pronto como se tiene disponible. En este contexto, con ayuda del procesamiento automático de textos, se pretende generar un sistema web que permita a los analistas políticos, pero también a la población general, explorar los tópicos o temas que son discutidos a lo largo de las conferencias de prensa matutinas del actual presidente de México.

Para llevar a cabo la detección y el seguimiento de los tópicos se utilizan las transcripciones de las conferencias[3]. Este sistema no intenta sustituir la labor crítica y especializada de las personas expertas. El objetivo del sistema es apoyar en la organización temática de lo que el presidente de México comunica en las conferencias matutinas diarias.

[3] La colección de transcripciones se pueden consultar en: https://lopezobrador.org.mx/transcripciones/

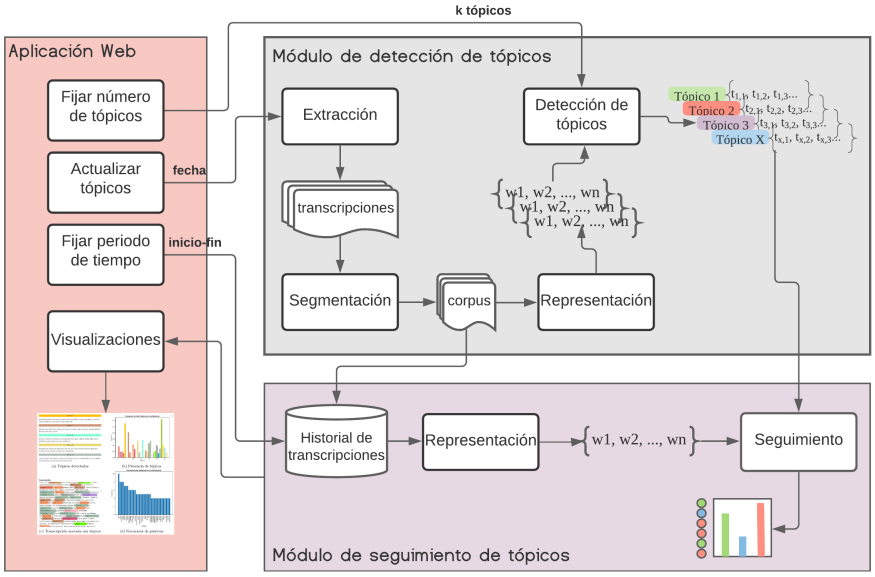


Fig. 1. Esquema general del sistema propuesto que contiene tres módulos principales: detección de tópicos, seguimiento de tópicos y la aplicación web.

En este artículo se propone realizar un sistema web que i) incorpore un método de detección automático de tópicos que demuestre generar tópicos de alta cohesión y variados, ii) incorpore un método de seguimiento de los tópicos encontrados dentro de un conjunto nuevo de transcripciones, y iii) permita explorar, mediante visualizaciones propuestas, los tópicos en una o más conferencias seleccionadas.

Por lo tanto, las aportaciones de este proyecto son dos. Primero, la evaluación de un modelo basado en LDA para la detección de tópicos en transcripciones en español. Segundo, el desarrollo de un sistema web que incorpore dicho modelo y que permite explorar los tópicos o temas que aborda el presidente de México en una o más conferencias.

2. Trabajo relacionado

La Detección y Seguimiento de Tópicos (TDT, por sus siglas en inglés) es un área de estudio dentro del procesamiento del lenguaje natural (PLN). El TDT está conformado por tres tareas [1]: i) Segmentación de una fuente de información en historias. ii) Detección de tópicos no conocidos por el sistema. Y iii) Seguimiento de tópicos conocidos por el sistema.

Existen herramientas que analizan textos en inglés para la detección de tópicos para su posterior visualización. A continuación se describen algunas que tienen características similares a las propuestas:

– **Terminology Extraction**[4]. Esta herramienta identifica términos clave dentro de un documento de texto. En resumen, se compara la frecuencia de palabras en un documento dado, con la frecuencia de uso de las palabras dentro de un lenguaje determinado. Para encontrar las palabras relevante, se utiliza la distribución de Poisson, el método de estimación por máxima verosimilitud y la frecuencia inversa de documentos. Además, utiliza un etiquetador probabilístico para identificar los términos que serán extraídos.
– **Term Extraction**[5]. Similar a la herramienta anterior, extrae la terminología de un texto de entrada. Pero Term Extraction permite la configuración de ciertos parámetros, como el número de términos a buscar dentro del texto, y el número de palabras que pueden conformar un término (un término se construye con un conjunto de palabras).
– **jsLDA**[6]. Esta herramienta implementa, en el lenguaje de programación JavaScript, el modelo de detección de tópicos LDA. Permite realizar la búsqueda de temas dentro de un conjunto documentos. Entre las configuraciones que un usuarios puede realizar están: el número de tópicos a ser encontrados y el número de iteraciones del modelo sobre los documentos. Dentro de las visualizaciones disponibles, la herramienta permite ver las correlaciones que hay entre los tópicos encontrados. Además, por cada tópico es posible visualizar una serie de tiempo de su presencia en los documentos. Adicionalmente, jsLDA muestra algunas estadísticas del vocabulario encontrado, como la frecuencia del término y la especificidad de cada término con respecto a los tópicos.

De manera general, todas las herramientas descritas son capaces de analizar textos en inglés. Aunque la mayoría de ellas no tiene visualizaciones intuitivas para el público general. La herramienta capaz de aceptar diferentes parámetros de configuración, jsLDA, produce visualizaciones orientada a describir el comportamiento del algoritmo LDA, por lo que las gráficas que genera están dirigidas a personas que conocen el funcionamiento de LDA.

3. Sistema propuesto

La Figura 1 muestra el diagrama general del sistema propuesto que contiene tres módulos: i) Módulo de detección de tópicos; ii) Módulo de seguimiento de tópicos; y iii) Aplicación Web Dicta. La descripción de cada módulo se detalla a continuación.

3.1. Módulo de detección de tópicos

La fuente primaria del sistema propuesto, como ya se ha mencionado, son las transcripciones que se publican diariamente en el sitio web del presidente de México actual[7]. Por lo tanto, el primer paso del sistema es la obtención de las transcripciones.

[4] http://labs.translated.net/terminology-extraction/
[5] http://termextract.fivefilters.org/
[6] https://mimno.infosci.cornell.edu/jsLDA/
[7] https://lopezobrador.org.mx/transcripciones/

PRESIDENTE ANDRÉS MANUEL LÓPEZ OBRADOR: Buenos días.

Bueno, vamos a informar y se trata de una muy buena noticia. Como se dijo desde el principio del gobierno, se heredó un déficit de especialistas médicos o médicos especialistas. Fue un saldo -otro más- negativo de la política neoliberal…

ALEJANDRO SVARCH PÉREZ, TITULAR DE LA COORDINACIÓN NACIONAL MÉDICA DEL INSTITUTO DE SALUD PARA EL BIENESTAR (INSABI): Con su permiso, señor presidente, señor secretario. Muy buenos días.

Como ustedes saben, lo hemos platicado en este espacio, nuestro país lamentablemente tiene un déficit estructural de médicos especialistas. Esto ha sido particularmente sensible en momentos como la pandemia que vivimos…

JORGE ALCOCER VARELA, SECRETARIO DE SALUD: Muchas gracias, señor presidente.

Sí, esta es una estrategia, pero no la única. Tenemos desde luego ya de años atrás varias escuelas de medicina y que fueron impulsadas por la doctora Sosa, que conduce este plan, también de becas de carreras de medicina y de otras especialidades a lo largo de todo el país, que superan las 100, 'Benito Juárez' es su nombre y desde luego ahí hay 12 escuelas de medicina…

INTERLOCUTOR: Disculpe, secretario, ¿tendrá alguna fecha de cuándo va a iniciar en funciones esta Universidad de la Salud y con cuántos estudiantes iniciará?...

PRESIDENTE ANDRÉS MANUEL LÓPEZ OBRADOR: Nos acaban de informar que ya está operando, funcionando…

(a) No segmentada

Buenos días. Bueno, vamos a informar y se trata de una muy buena noticia. Como se dijo desde el principio del gobierno, se heredó un déficit de especialistas médicos o médicos especialistas(...) A ver, si puedes explicar sobre las nuevas universidades. Sí, nada más que es importante que se dé a conocer que se han iniciado ya, incluso están en operación, muchas escuelas de medicina en todo el país, tanto del sistema de universidades 'Benito Juárez' como iniciativas que tomaron gobiernos locales(...) Sigue habiendo clases, pero virtuales Vamos a presentarles toda la cobertura, porque son 140 universidades públicas nuevas que están en proceso, es algo extraordinario. Acabo de inaugurar dos, una en Agua Prieta y otra en Tlatizapán, Morelos; en Agua Prieta, Sonora, y en Tlaltizapán, Morelos. Son 140 del sistema de universidades 'Benito Juárez' y están en los municipios, en las regiones más apartadas(...)

(b)Segmentada

Fig. 2. Fragmento de la transcripción del 5 de noviembre de 2019, antes y después de la segmentación usando como actor político de interés a Presidente Andrés Manuel López Obrador.

La salida general de este módulo es un conjunto de vectores, uno por cada tópico detectado. Cada vector-tópico contiene un conjunto de términos asociados a ese tópico. Este módulo se compone por cuatro procesos: Extracción, Segmentación, Representación y Detección.

Extracción. Este proceso es el encargado de realizar la recolección periódica de las transcripciones. Para la implementación de este proceso se usó la biblioteca de Python Scrapy[8]. Al final de la extracción, las transcripciones obtenidas son almacenadas en formato CSV (Comma-Separated Values).

Segmentación. Durante las conferencias de prensa diarias de López Obrador a menudo participan funcionarios de su gobierno. Estas participaciones también son transcritas. Por lo tanto, el objetivo de este proceso es identificar el contenido textual relacionado con el actor político objetivo y el resto de las participaciones es eliminada. En el caso de este proyecto, el actor político objetivo es el presidente de México.

[8] https://scrapy.org/

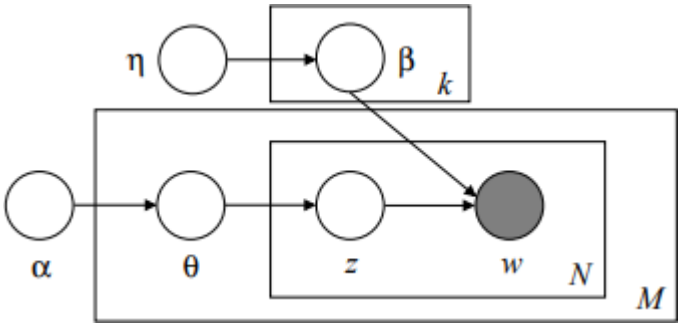


Fig. 3. Representación gráfica del modelo LDA.

Las transcripciones que contienen sólo la información del político de interés formarán parte del corpus de documentos que se usarán para la detección de los tópicos a través LDA. Como puede verse en la Figura 1 el corpus se almacena en una base de datos que contiene el historial de las transcripciones que la aplicación web usará posteriormente.

En la Figura 2a, se puede observar un fragmento de una transcripción sin segmentar (entrada de este proceso) y posteriormente, en la Figura 2b aparece sólo el texto correspondiente al actor de interés (i.e., López Obrador). Este fragmento corresponde a la transcripción de la conferencia de prensa del 5 de noviembre de 2019.

Representación. Una vez que se extraen y segmentan las transcripciones, se preparan los textos para la representación. Dado que el interés del sistema propuesto es la detección de temas o tópicos tratados dentro de las conferencias matutinas, se eliminan todas las palabras que no sean sustantivos, con el fin de conservar únicamente palabras de contenido. El conjunto completo de pasos en el pre-procesamiento se lista a continuación (los tres primeros procesos fueron realizados con expresiones regulares):

- Se transforma todo el texto a minúsculas.
- Se eliminan los números contenidos dentro del texto.
- Se elimina cualquier signo de puntuación.
- De cada una de las transcripciones se extraen los sustantivos[9].

Cabe destacar que las transcripciones se realizan con ayuda de una técnica estenográfica, por lo que en múltiples ocasiones existen errores ortográficos contenidos dentro de la transcripción y por ende dentro del conjunto de documento ya preprocesado.

Detección. Una vez pre-procesado y representado el conjunto de documentos a analizar, se utiliza el algoritmo de detección de tópicos para aprender los tópicos relevantes de ese conjunto. Específicamente, el modelo empleado para la detección de tópicos es Latent Dirichlet Allocation (LDA) [5].

[9] Para obtener los sustantivos se utilizó: https://spacy.io/models/es

Evolución de los tópicos.

Distribución de tópicos en el tiempo.

Participantes en la conferencia del dia 2020-07-21 :

	Nombre
0	SECRETARIO MARCELO EBRARD CASAUBÓN
1	HUGO LÓPEZ GATELL RAMÍREZ
2	PRESIDENTE ANDRÉS MANUEL LÓPEZ OBRADOR
3	JORGE ALCOCER VARELA SECRETARIO DE SALUD

Participantes en las conferencias.

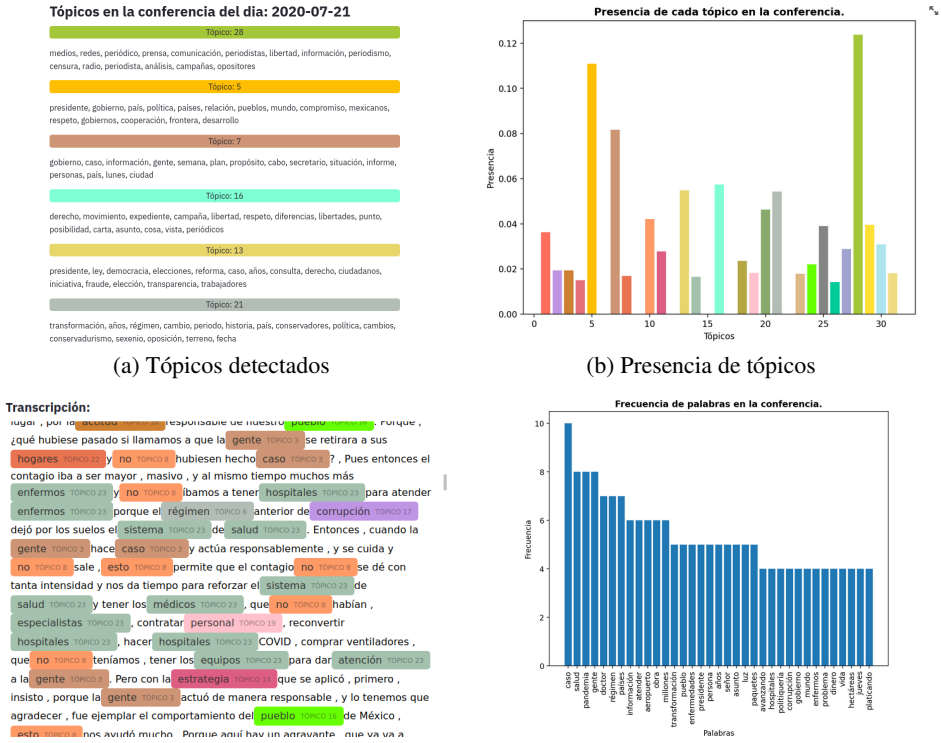
Fig. 4. Visualizaciones generales en DICTA.

En la sección 4 se presentan los experimentos para determinar el número adecuado de tópicos para este conjunto de documentos. La salida de este proceso son los tópicos detectados por el modelo.

Cada tópico esta asociado con un conjunto de términos que describen o que pertenecen al tópico en cuestión. Junto a cada uno de los términos se encuentra un valor que indica la pertenencia que cada término tiene con respecto al tópico.

La idea general detrás del modelo LDA, consiste en que los documentos están representados como una distribución aleatoria sobre tópicos latentes, donde cada tópico se caracteriza por una distribución de términos.

Esto es, se asume que los tópicos existen antes que los documentos y que estos documentos se construyen a partir de tales tópicos [5, 4]. En la Figura 3 se puede ver la representaciósn gráfica del modelo LDA.



(a) Tópicos detectados

(b) Presencia de tópicos

(c) Transcripción anotada con tópicos

(d) Frecuencia de palabras

Fig. 5. Visualizaciones en DICTA que contiene información de una conferencia específica.

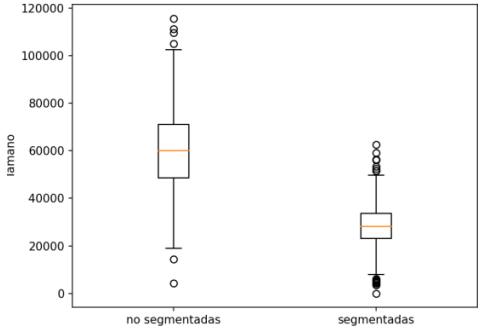
Cada nodo es una variable aleatoria; α y η son distribuciones Dirichlet; β es una distribución de palabras, una para cada tópico; θ es una distribución de tópicos, una por cada documento; N, M y k denotan replicación; N denota la colección de palabras dentro de cada documento; M es el conjunto de documentos en la colección; y k el número de tópicos. Finalmente, w denota una palabra en un documento y z un tópico dentro de un conjunto de tópicos [5].

3.2. Módulo de seguimiento de tópicos

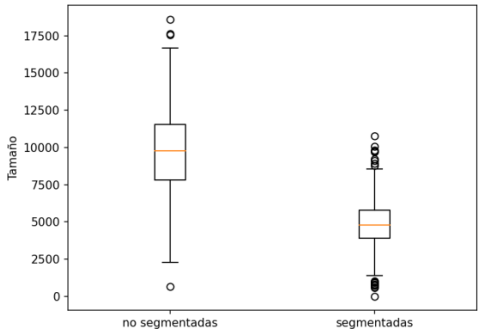
Este módulo hace el seguimiento de los tópicos (previamente detectados) en un conjunto dado de transcripciones de las conferencias matutinas. Para su funcionamiento, este módulo necesita dos parámetros.

El primer parámetro es el periodo de tiempo de las transcripciones a analizar del historial de transcripciones (ver Figura 1). Estas transcripciones deberán ser pre-procesadas y representadas usando los mismos procesos descritos en la etapa de Detección de tópicos (ver Sección 3.1).

El segundo parámetro, es el conjunto de vectores correspondientes a los tópicos detectados que corresponde a la salida del Módulo de detección de tópico explicado en la sección anterior.



(a) Tamaño (en caracteres) de las transcripciones no segmentadas y segmentadas.



(b) Tamaño del vocabulario de las transcripciones no segmentadas y segmentadas.

Fig. 6. Estadísticas del conjunto de datos usados en la evaluación experimental.

La salida de este módulo es un nuevo vector que contiene los tópicos encontrados dentro de la transcripción y el valor correspondiente a la pertenencia de cada tópico dentro del documento analizado.

Luego, este vector es enviado al sistema web para generar la visualización correspondiente. En la Figura 1 se representa esta salida como una distribución de los tópicos en los documentos analizados y las palabras (círculos) de cada tópico encontradas.

3.3. Dicta: Aplicación web de visualización de tópicos

El sistema web denominado DICTA[10], integra los módulos descritos anteriormente y permite al usuario visualizar los tópicos de un conjunto de conferencias matutinas dado un rango de fechas. Como puede verse en la Figura 1, la comunicación con el módulo de detección de tópicos se realiza a través de la selección del número de tópicos simultáneamente con la opción de actualizar los tópicos.

La comunicación con el módulo de seguimiento de tópicos se lleva a cabo a través de fijar el periodo de tiempo de análisis y posteriormente, con la visualización generada. Las funcionalidades más relevantes de DICTA son:

[10] DICTA y la base de datos utilizada están disponibles en https://github.com/lyr-uam/dicta

Tabla 1. Primeras 20 palabras asociadas a los cuatro tópicos generados con Top2Vec. Entre paréntesis, el tema asignado por los autores a cada tópico.

Tópico ID	Términos asociados
0 (política exterior)	migratorio Centroamérica ebrad marcelo donald humanos aranceles fenómenos confrontación migración cooperación relaciones exteriores soberanía guerra trump naciones paz derechos violencia
1 (educación / clases)	educativo gratuita maestros educativa educación medicina apartadas calidad normales servicios universidades medicinas mejorar medicamentos superior vendían salud escuelas niveles unam
2 (salud)	epidemia endemia coronavirus camas enfermos intensiva terapia recomendaciones hugo ventiladores enfermeras salvar proyecciones hospitalización especialistas medico crisis cuidarnos normalidad científicos
3 (energía)	estancias pemex infantiles cancelar energética contratos transparencia lopez simulación obrador comisión signifique expediente electricidad debate gasoducto organizaciones llamada organismo barriles

– Fijar el número de tópicos. Este proceso permite al usuario indicar cuántos tópicos desea explorar. Entre mayor el número de tópicos, más fina es la organización de temas a visualizar. Si no se fija un número, el sistema usará el obtenido en la etapa de evaluación (que se describe en la Sección 4).

– Actualizar tópicos. Esta opción, permite lanzar el módulo de detección de tópicos para que se puedan acceder a las transcripciones más recientes. Adicionalmente, se considera como entrada el número de tópicos fijado por la opción anterior.

– Fijar periodo de tiempo. Este proceso valida que las fechas ingresadas contengan una transcripción dentro de la base de datos (o historial de transcripciones). Usualmente en los días festivos en México o fines de semana no existen conferencias matutinas. Después de esta validación se cargan a memoria las conferencias que pertenezcan al rango de tiempo especificado.

Una parte importante del sistema propuesto es la generación de visualizaciones dirigidas al público general. Por lo tanto, el sistema genera seis diferentes gráficos divididos en información general de las conferencias (Figura 4) e información específica de una conferencia a analizar (Figura 5). A continuación se describen brevemente cada visualización.

– Distribución de tópicos en el tiempo. Esta gráfica (Figura 3.1) muestra la presencia, en porcentajes de cada tópico en la conferencias de prensa de la base de datos en un momento dado. En el eje de las x se grafican las conferencias de la más antigua a la más nueva; y en el eje de las y se grafica el porcentaje relativo al 100 % de la transcripción dada, de la presencia de cada tópico. Por ejemplo, el tópico 5 aparece en todas las conferencias entre el 10 y el 20 % del total de cada transcripción.
– Participantes en las conferencias de prensa. Este gráfico (Figura 3.1) muestra los participantes de una conferencia dada. En este proyecto nos enfocamos en el análisis de un único actor político: el presidente de México, sin embargo, existe información que en un futuro puede ser relevante.

Tabla 2. Primeras 30 palabras asociadas a los cuatro tópicos generados con LDA (Gensim). Entre paréntesis, el tema asignado por los autores a cada tópico.

Tópico ID	Términos asociados
0 (intro. a la conferencia)	vamos entonces va si mexico mil ahora pueblo gobierno gente ver aquí van país bien así como voy caso bueno hacer decir tiempo luego mismo importante haciendo mañanera ser puede
1 (sin asignación)	entonces vamos si va gobierno corrupción ahora ver pues caso ser bueno como mismo gente méxico ahí aquí van presidente así voy hacer poder puede bien tiempo pueblo luego país
2 (salud / economía)	mil va salud millones vamos si entonces empresas médicos hospitales pesos ahora presupuesto medicamentos van como año avión créditos dos petroleo caso gente ahí trabajadores dinero seguro bueno hacer deuda
3 (errores de transcripción)	delpresupuesto alas demanera vamos estesemana periodneoliberal quese muybien poderlegislativo atransparenar lacorrupcion enel ainformar delpoder camade elpresupuesto antidemocratica peregrinos ingrese laseguridad enlos experimentados siestamos tenemospetroleo sanchezcordero supuestmenteesta combustibleque tenerproteccion superdelegado tenganprincipios

– Tópicos detectados. Gráfica (Figura 3.1) que lista los términos de los tópicos con mayor presencia en la conferencia analizada. Cada tópico se lista con un color particular y el tópico con mayor porcentaje de presencia en la conferencia se muestra primero (note que el número del tópico es sólo un identificador, no corresponde a la importancia o mayor presencia).
– Presencia de tópicos. En esta gráfica de barras (Figura 3.1) se muestra la presencia de cada tópico en la conferencia analizada. El tópico se representa por un identificador numérico y un color asignado (cabe hacer notar que el color del tópico es consistente en todas las gráficas mostradas en esta sección).
– Transcripción anotada. Esta visualización (Figura 3.1) muestra el documento transcrito de la conferencia analizada indicando con un mismo color todos los términos pertenecientes al mismo tópico. Se muestra el tópico al que pertenece dicho término de modo que en conjunción con la gráfica de la Figura 3.1 se pueda hacer un análisis más detallado de cada conferencia.
– Frecuencia de palabras. Finalmente, en esta gráfica de barras (Figura 3.1) se muestra la frecuencia de los términos más mencionados en cada transcripción analizada. Esta gráfica no está relacionada con los tópicos directamente, sino que es un conteo independiente del número de palabras en una conferencia.

4. Evaluación del modelo de detección de tópicos

Como complemento a la aportación principal presentada en este artículo, realizamos dos tipos de validación del modelo de detección de tópicos. Por un lado se compararon de forma cualitativa dos algoritmos de detección de tópicos para elegir el modelo a

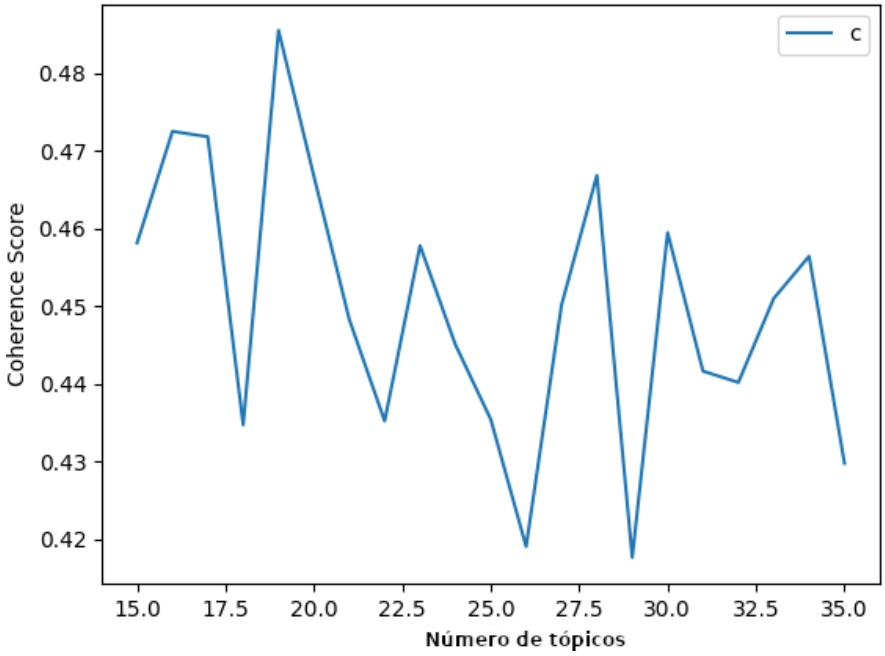


Fig. 7. Valor de coherencia en los tópicos obtenidos usando LDA.

incorporar en DICTA. Por otro lado, una vez elegido LDA, se usó una métrica de cohesión para determinar el número de tópicos recomendado para la tarea.

4.1. Conjunto de datos

El conjunto de datos está compuesto por 534 transcripciones segmentadas (como se describe en la Sección 3.1). Estas conferencias se dieron entre el 3 de diciembre del 2018 y el 8 de febrero del 2021. Por un lado, en la Figura 3.2 se puede observar que el tamaño promedio de caracteres de las transcripciones antes y después de la segmentación varía considerablemente, esto nos indica que la participación del actor de interés (el presidente de México) ocupa aproximadamente la mitad de las conferencias diarias, a pesar de que el número de participantes es más de 1.

Particularmente, el tamaño promedio de las transcripciones no segmentadas es de 60364 caracteres ($\sigma = 17106$) y el tamaño promedio de las transcripciones segmentadas es de 28199 caracteres ($\sigma = 915$). Por otro lado, en la Figura 3.2 se pueden observar el tamaño promedio del vocabulario usado en las transcripciones, antes y después de la segmentación.

El vocabulario es definido como palabras únicas. Las transcripciones no segmentadas tienen un tamaño de vocabulario promedio de 9728 ($\sigma = 2746$) palabras. Las transcripciones segmentadas tiene en promedio un tamaño de vocabulario de 4834 ($\sigma = 1572$).

4.2. Selección del modelo de detección de tópicos

En esta sección se evaluaron dos modelos de detección de tópicos: Top2Vec [2] y LDA [5] (bajo la implementación de Gemsim[11]). Top2Vec encuentra el número de tópicos de forma automática dentro del conjunto de documentos, mientras que a LDA se debe especificar el parámetro k (Figura 3). Por lo tanto, primero realizamos la prueba con Top2Vec para después utilizar el número de tópicos encontrado como parámetro de k para la prueba con LDA.

En la Tabla 1 se pueden ver los 4 tópicos generados por el modelo Top2Vec. De aquí se observa que los tópicos son identificables en áreas generales del gobierno de un país: política exterior, educación, salud, y energía. Como segundo experimento se utilizó LDA para generar 4 tópicos (este número de tópicos fue guiado por el número identificado de forma automática por Top2Vec). En la Tabla 2 se pueden observar los tópicos generados por LDA.

Es notorio que el conjunto de tópicos generados por LDA no son de la misma calidad que los generados por Top2Vec. Sin embargo, debido a la gran variedad de temas que se abordan en las conferencias matutinas del presidente de México actual, es deseable tener la posibilidad de generar tópicos más finos. A partir de la observación del tópico 2 de la Tabla 2 es viable inferir que a un mayor número de tópicos LDA podría generar temas de mejor calidad, dividiendo el tópico 2 en dos o más temas particulares.

4.3. Validación del número de tópicos

En este apartado se describe la validación del número de tópicos para las transcripciones de las conferencias del presidente de México, López Obrador. La calidad de los tópicos detectados por modelos como LDA, es medida por su grado de coherencia. Se puede decir que un tópico es coherente si la gran parte de los términos que describen a ese tópico en particular están relacionados [14].

Dado que LDA requiere que se especifique el valor de k (ver Figura 3), es importante determinar un valor que de k que responda a las necesidades de la tarea. Un valor pequeño de tópicos resultará en tópicos demasiado generales; mientras que un valor muy grande de k podría resultar en tópicos que no se pueden interpretar o que podrían combinarse.

Así, para medir la coherencia de los tópicos se puede utilizar un conjunto de métricas llamadas Coherence Measures. Estas métricas evalúan los tópicos a través de un promedio de la similitud entre pares de términos, que son tomados de los términos principales de un tópico [12]. En [11] se desarrolló una nueva métrica denominada C_v y que tiene una alta correlación con lo que los humanos consideran buenos tópicos.

Esta métrica esta basada en la combinación de las siguientes tres Indirect Cosine Measure, NPMI(Nomalized Pointwise Mutual Information), y Boolean Sliding Window (la definición formal de la métrica se puede encontrar en [11]). El valor de C_v está normalizado y va de 0 a 1; entre mayor el valor, mayor calidad de los tópicos.

[11] https://radimrehurek.com/gensim/

Para este experimento se generaron modelos incrementando el número de tópico (de 15 a 35). En la Figura 7 se puede observar el valor de cohesión C_v de los 20 modelos construidos. Los parámetros de LDA utilizados para este experimento fueron: $\alpha = 0.1$ y $\eta = 0.9$. De aquí, el mejor valor de coherencia obtenido es con un número 18 de tópicos, seguido de 28.

5. Conclusiones

En este artículo se presentó un sistema web que apoya en el análisis del discurso político del presidente de México, Andrés Manuel López Obrador. El sistema comprende tres módulos principales: el módulo de detección de tópicos, el módulo de seguimiento de tópicos y la aplicación web.

La aplicación propuesta permite a un usuario visualizar los temas o tópicos que están presentes en el dichas conferencias. Entre las funcionalidades del sistema se encuentran: fijar un periodo de tiempo para el análisis, actualizar el modelo de detección de tópicos (fijando el número de tópicos a detectar).

El módulo de detección de tópicos utiliza LDA con la implementación de Gensim para la generación automática de tópicos. Se evaluó la cohesión de los tópicos generados por LDA, obteniendo un valor máximo de Coherence score de 0.49 con 18 tópicos. Actualmente el sistema analiza a un actor político de interés. Como trabajo futuro se buscará que el usuario pueda determinar el conjunto de actores políticos (participantes en las conferencias) que desee analizar.

Agradecimientos. Los autores y autora agradecen a la Universidad Autónoma Metropolitana Unidad Cuajimalpa por el apoyo otorgado durante la realización de este proyecto. El tercer autor además fue apoyado parcialmente por Idiap Research Institue y el SNI-CONACyT México.

Referencias

1. Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study final report (1998)
2. Angelov, D.: Top2vec: Distributed representations of topics (2020) doi: 10.48550/arXiv.2008.09470
3. Bhatia, A.: Critical discourse analysis of political press conferences. Discourse & Society, vol. 17, no. 2, pp. 173–203 (2006) doi: 10.1177/0957926506058057
4. Blei, D. M.: Probabilistic topic models. Communications of the Association for Computing Machinery, vol. 55, pp. 77–84 (2012) doi: 10.1145/2133806.2133826
5. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation. Journal of Machine Learning Research, vol. 3, pp. 993–1022 (2003)
6. Francisco-Ortega, D.: Coronavirus outbreak in Mexico: A critical discourse analysis of AMLO's speech. Open Journal for Studies in Linguistics, vol. 3, no. 2, pp. 93–100 (2020)
7. Grimmer, J., Stewart, B. M.: Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, vol. 21, no. 3, pp. 267–297 (2013) doi: 10.1093/pan/mps028

8. Marini, A. M.: El mesías tropical: Aproximación a fenómenos populistas actuales a través del discurso de López Obrador, no. 139, pp. 153–170 (2018)
9. Navarro, F., Tromben, C.: Estamos en guerra contra un enemigo poderoso, implacable: Los discursos de Sebastián Piñera y la revuelta popular en chile. Literatura y lingüística, pp. 295–324 (2019) doi: 10.29344/0717621x.40.2083
10. Nájar, A.: Así son las mañaneras, la novedosa estrategia para gobernar de AMLO en México. BBC News Mundo, (2019)
11. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. pp. 399–408 (2015) doi: 10.1145/2684822.2685324
12. Rosner, F., Hinneburg, A., Röder, M., Nettling, M., Both, A.: Evaluating topic coherence measures (2014) doi: 10.48550/arXiv.1403.6397
13. Sarfo, E., Krampa, E. A.: Language at war: A critical discourse analysis of speeches of Bush and Obama on terrorism. International Journal of Social Sciences and Education, vol. 3, no. 2 (2012)
14. Syed, S., Spruit, M.: Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics. pp. 165–174 (2017) doi: 10.1109/DSAA.2017.61
15. Tausczik, Y. R., Pennebaker, J. W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology, vol. 29, no. 1, pp. 24–54 (2010) doi: 10.1177/0261927X09351676
16. Torío, M. Á. R.: Características del lenguaje político: La designación. Philologia Hispalensis, no. 10, pp. 7–22 (1995) doi: 10.12795/ph.1995.v10.i01.01