UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

SISTEMA PARA CUANTIFICAR LA COBERTURA DE LOS MEDIOS DE COMUNICACIÓN EN CHILE

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELÉCTRICO

ALVARO SEBASTIÁN TOLEDO MONTERO

PROFESOR GUÍA:
ANDRÉS ABELIUK KIMELMAN
PROFESOR CO-GUÍA:
CRISTIAN CANDIA VALLEJOS

COMISIÓN:
AIDAN HOGAN

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: ALVARO SEBASTIÁN TOLEDO MONTERO
FECHA: 2022
PROF. GUÍA: ANDRÉS ABELIUK

SISTEMA PARA CUANTIFICAR LA COBERTURA DE LOS MEDIOS DE COMUNICACIÓN EN CHILE

En un mundo conectado las personas tienen diversas formas para acceder a noticias de interés. Las redes sociales han dado paso al surgimiento de medios alternativos por los cuales informarse y medios tradicionales se han adaptado a estas redes. Los medios noticieros en redes sociales tienden a compartir una variedad de noticias de distintos tópicos que dependen del enfoque temático de la cuenta o el contexto local en la que se encuentra. Debido al exceso de información en el mundo moderno, el usuario se ve dificultado para filtrar noticias con tópicos relevantes para éste. Por lo tanto, sería de utilidad tener una herramienta que permita saber los tópicos que abarcan los noticieros de forma rápida, fácil de entender y permita realizar comparaciones entre los distintos medios.

En este trabajo se propone una forma de visualizar la distribución o cobertura de tópicos presentes en las noticias chilenas. Se realiza un seguimiento a las noticias publicadas por 371 medios tradicionales y alternativos en *Twitter*. Para poder determinar los tópicos de los titulares utilizados se implementó *Distributed Dictionary Representations* (DDR). A través de diccionarios basados en la taxonomía descrita por el *International Press Telecommunications Council* (IPTC) se clasifica cada noticia en los distintos tópicos existentes. Finalmente, se realizó un análisis comparativo entre cuentas para identificar aquellas que tengan ventaja a la hora de publicar noticias de un tópico particular utilizando como medida el *Revealed Comparative Advantage* (RCA). Este último también se utilizaría para el cálculo de medidas de proximidad entre cuentas, y así lograr formar grupos de cuentas que sean similares entre sí, y formar conexiones entre estas.

Para el proceso de evaluación del método utilizado se realizó una serie de encuestas, donde los encuestados debían asignar un tópico (usados en el proceso de clasificación) a una muestra aleatoria de noticias. De esta forma se pueden comparar los tópicos que asigna el modelo planteado y los que serían asignados por personas. Los resultados de estas encuestas muestran una clasificación aceptable (pero mejorable) de los tópicos de nivel más generales. Se logra la visualización para el porcentaje de tópicos presentes en las cuentas a través del tiempo. A través de los gráficos obtenidos se puede identificar la influencia de eventos importantes en los porcentajes de publicación en ciertas cuentas. Por otro lado, es posible identificar cuentas que no se ven afectadas en gran medida por los hechos ocurridos a nivel país o mundo. Se logra comprobar que RCA funciona como medida para tener un ranking de las cuentas en cuanto a las publicaciones que realizan respecto a un tópico como Política. Finalmente se logran formar grupos de cuentas que se asemejan entre sí, de la misma forma en que existen cuentas que no pueden ser encasilladas con otras. Es posible una variación en el ranking y formación de grupos cuando se considera una ventana de tiempo específica (por ejemplo, solamente las publicaciones realizadas en la última semana) o medida que se actualiza la base de datos con publicaciones más recientes.

Gracias a quienes confiaron en mí
Cuando yo no lo hice

Saludos

Agradecimientos

El primer agradecimiento va hacia mi profesor guía, Andrés Abeliuk, por todo el apoyo que me ha otorgado, por las dudas resueltas, por las alternativas y soluciones propuestas, por darse el tiempo de preguntar como iba el trabajo y por juntarse a resolver dudas y fijar los pasos a seguir. Agradezco infinitamente que me haya aceptado como memorista, sobre todo porque me encontraba complicado para encontrar un profesor.

A mis amigos, por haberme dado energía y ganas de continuar, por haberme permitido recobrar un poco más de seguridad en mi mismo y permitirme disfrutar de instancias sociales. Por todo el apoyo emocional que han sido durante este proceso.

A mis perros, Tritón y Beethoven (por siempre en mis recuerdos), por siempre sacar sonrisas y entregar cariño incondicionalmente. A mi familia, por siempre preguntar como voy, mi estado emocional y la paciencia durante todo el proceso.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

La actual masificación en la producción de información y noticias dificulta el proceso que debe realizar el usuario para obtener aquellos temas relevantes para éste. Además, se agrega la frecuencia con la que los medios de comunicación constantemente cambian de tópicos para abarcar nuevos, en una especie de olvido selectivo [1] generando un mayor excedente en información nueva. Este exceso de información también implica un olvido de los tópicos recientemente abarcados debido al surgimiento de nuevas noticias, en acto de olvido por anulación [2]. Según la definición de la Real Academia Española los medios de comunicación se utilizan como instrumento para transmitir públicamente todo tipo de información a la población [3]. Entre estos, se pueden distinguir la televisión, periódicos y revistas, internet, y radio.

Para el contexto chileno, en el estudio realizado por *Reuters Institute de la Universidad de Oxford* [4] se observa que la modalidad online es la más utilizada a la hora de buscar información. En este mismo estudio se observa que las redes sociales tienen un rol mucho más relevante a la hora de comunicar noticias. Con un 73 % de la población utilizando las redes sociales se destrona a la televisión como el medio tradicional para la difusión de noticias con tan solo un 66 %. Uno de los mayores factores a considerar para la explicación de este fenómeno puede deberse a que el 83 % de la población cuenta con un Smartphone. En este mismo estudio se puede observar que los medios noticieros tradicionales como TVN, Chilevisión, entre otros, han adaptado sus noticieros a una modalidad online (incluyendo redes sociales). Sin embargo, en el contexto de las redes sociales son los grandes noticieros los que siguen logrando captar más la atención de las personas. Es tanto así que páginas como Emol, BíoBíoChile, 24horas, entre otros, siguen siendo aquellas con mayor alcance, con un porcentaje de 27 % a 24 %.

A partir de las manifestaciones originadas el 18 de octubre de 2019 surgió un descontento general de la población hacia los medios tradicionales de comunicación chilenos [5]. Del estudio realizado por Luna, Toro y Valenzuela [5], se tiene que el seguimiento de medios alternativos a través de las redes sociales se sostiene a través del tiempo. El decaimiento en la confianza a los medios tradicionales es consistente al revisar los resultados publicados en *Reuters Institute Digital News Report 2020* [4]. Se tiene una caída del 15 % en la confianza hacia las noticias, quedando con un total del 30 % y posicionando a Chile en el lugar 28 en el ranking de confianza de los 40 países que se incluyeron en el estudio.

De las redes sociales, es Facebook la más utilizada para la obtención de información en Chile con un 63 %. Le siguen WhatsApp con un 40 %, Instagram con un 28 % [4] y Twitter con 22 %. El nivel de impacto que llegan a tener las redes sociales como medio informativo

puede ser bastante grande en general. Como ejemplo, basta destacar que se realizó un total de 3.5 millones de publicaciones en Twitter relacionadas al proceso constituyente entre el 2019 y 2020 [6]. Este nivel de cobertura es con respecto a un tópico en particular, y no realizado exclusivamente por medios noticieros tradicionales o alternativos.

Otro aspecto interesante de las noticias es la atención que se le otorgan a estas dependiendo del impacto que tienen en la población o en los eventos a nivel nacional/global. Un evento ocurrido en el presente puede ser cubierto por una gran cantidad de medios. Sin embargo, a la siguiente semana se puede ver opacado por otros tópicos. En ese aspecto, la atención colectiva que otorgan los medios a distintos temas pueden influenciar la velocidad en la que se olvidan eventos pasados. Así mismo, tener se facilita el registro de esta información y la fecha del suceso para una especie de memoria cultural [7].

1.1. Identificación y Formulación del Problema

A partir de lo mencionado resulta interesante preguntar cómo los diferentes medios (tradicionales y alternativos) cubren los diferentes tópicos que ocurren a nivel país. La asignación de tópicos a noticias es una tarea que se ha realizado [8], sin embargo, el trabajo realizado para textos en español es más escaso. Además, al realizar este tipo de análisis a medios chilenos permite extraer información con respecto a eventos importantes ocurridos a través del tiempo. El enfoque de los noticieros varía a medida que nuevos sucesos ocurren, como puede ser la pandemia causada por el COVID-19 o las elecciones presidenciales.

Por lo tanto surgen las preguntas: ¿Cómo cubren los medios chilenos los distintos tópicos a medida que surgen nuevos eventos? ¿Es posible comparar los distintos noticieros utilizando como medida el porcentaje de cobertura que le dan a distintos temas? ¿Existe una relación entre la atención que le dan los medios a un tema particular y el alcance que tienen? ¿Cúales tópicos tienden a ser más hablados por los medios noticieros?

1.2. Objetivos del Trabajo de Título

Para poder responder a la problemática planteada y desarrollar un sistema que sea capaz de cuantificar la cobertura de los noticieros. Por lo tanto, el objetivo principal del Trabajo de Título es:

Diseñar un sistema capaz de cuantificar y comparar la cobertura que tienen los medios de comunicación chilenos con respecto a diferentes tópicos a través del tiempo, implementando una forma de visualización de la información obtenida.

Para poder cumplir el objetivo principal del Trabajo de Título es necesario combinar distintas etapas cada una con su propio objetivo. Se utilizarán las publicaciones de las cuentas de Twitter de distintos medios tradicionales y alternativos. Dado que se busca hacer un análisis con respecto a diferentes tópicos no se realizará un filtrado de publicaciones. Además, se busca realizar un análisis lo más cercano a tiempo real, obteniendo las últimas publicaciones realizadas por las cuentas seguidas. Se debe solucionar la obtención de datos, determinar las cuentas que se utilizarán para el estudio seguimiento y el periodo de extracción de los datos. Se deben probar distintos algoritmos de clasificación, y definir aquel que se utilizará para el desarrollo final. Debido a esto los objetivos secundarios y específicos a cada etapa son los siguientes:

- Extraer información de Twitter de diferentes medios noticieros para crear una base de datos que será utilizada en el Trabajo de Título
- Estudiar, probar y seleccionar los diferentes métodos de clasificación por tópicos a utilizar. Realizar los ajustes necesarios para la optimización de estos
- Definir medidas de validación de los algoritmos utilizados
- Establecer los diversos tópicos que se asignarán a las noticias estudiadas
- Implementar una forma de visualización para los resultados obtenidos

Capítulo 2

Marco Teórico y Estado del Arte

Para cumplir el objetivo principal se requiere el estudio previo de los temas relacionados a cada etapa, y la revisión bibliográfica de trabajo realizado en materias similares. En particular se requiere la revisión de los distintos métodos de aprendizaje de máquinas y la categorización de los diferentes tópicos con los que se desea clasificar.

2.1. Machine Learning

La Inteligencia Artificial o *Artificial Intelligence* (AI) es un campo dedicado a la producción de sistemas capaces de aprender y realizar tareas humanas [24]. En AI existe una gran cantidad de técnicas que se pueden aplicar para realizar distintos tipos de tareas. Entre ellas se pueden encontrar las Redes Neuronales, Deep Learning, Lógica Difusa, Algoritmos Genéticos, Procesamiento de Lenguaje Natural, Robótica, Visión Artificial, entre otros [25]. El Aprendizaje de Máquinas o *Machine Learning* (ML) es una de las técnicas que presenta esta área de la computación. Dadas las características del problema que se desea resolver (clasificación de noticias) es que se indagará más sobre ML.

Machine learning se caracteriza por buscar que el sistema obtenga información a partir de los datos que le son entregados. El objetivo principal es que el programa encuentre patrones y datos que le permitan realizar decisiones sin tener que depender de humanos [26]. ML puede ser aplicado en una gran cantidad de escenarios donde uno de estos es la clasificación o agrupación de datos [27]. Existen diferentes técnicas ML y pueden ser clasificadas entre 4 a 6 clases distintas, sin embargo, son el aprendizaje supervisado y no supervisado los 2 mayores exponentes [28]. Estas técnicas son:

- Aprendizaje supervisado: Utilizada datos etiquetados. El sistema aprende de estos para etiquetar datos nuevos.
- Aprendizaje no supervisado: No hay datos etiquetados. El sistema extrae características de los datos para agruparlos.
- Aprendizaje semi-supervisado: Utiliza ambos tipos de datos.
- Aprendizaje reforzado: Aprende a través de la realimentación que se le entrega al sistema.

2.1.1. Aprendizaje supervisado

Este método de *Machine Learning* se caracteriza por varios aspectos, uno de ellos es la utilización de datos etiquetados. Estas etiquetas corresponden a las clases o lo esperable

de los datos donde el objetivo principal es lograr que el sistema aprenda del patrón de estos datos y sea capaz de predecir con datos nuevos. El aprendizaje supervisado se puede subdividir en distintas técnicas dependiendo de la herramienta matemática que utilizan. En la Figura 2.1 se pueden ver los diferentes algoritmos de aprendizaje supervisado y la rama a la que pertenecen. Uno de los requerimientos esenciales de esta forma de aprendizaje son los datos etiquetados lo que puede consumir tiempo si se trabajan con una base de datos nueva. Un aspecto positivo dentro de esta técnica de ML es la facilidad para validar los modelos obtenidos. Cómo se tienen los datos etiquetados se puede calcular la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Con estos valores se puede tener las medidas de Precisión y Recall, F1-Score, el área bajo la curva, entre otros, que permiten evaluar la calidad del modelo obtenido [32].

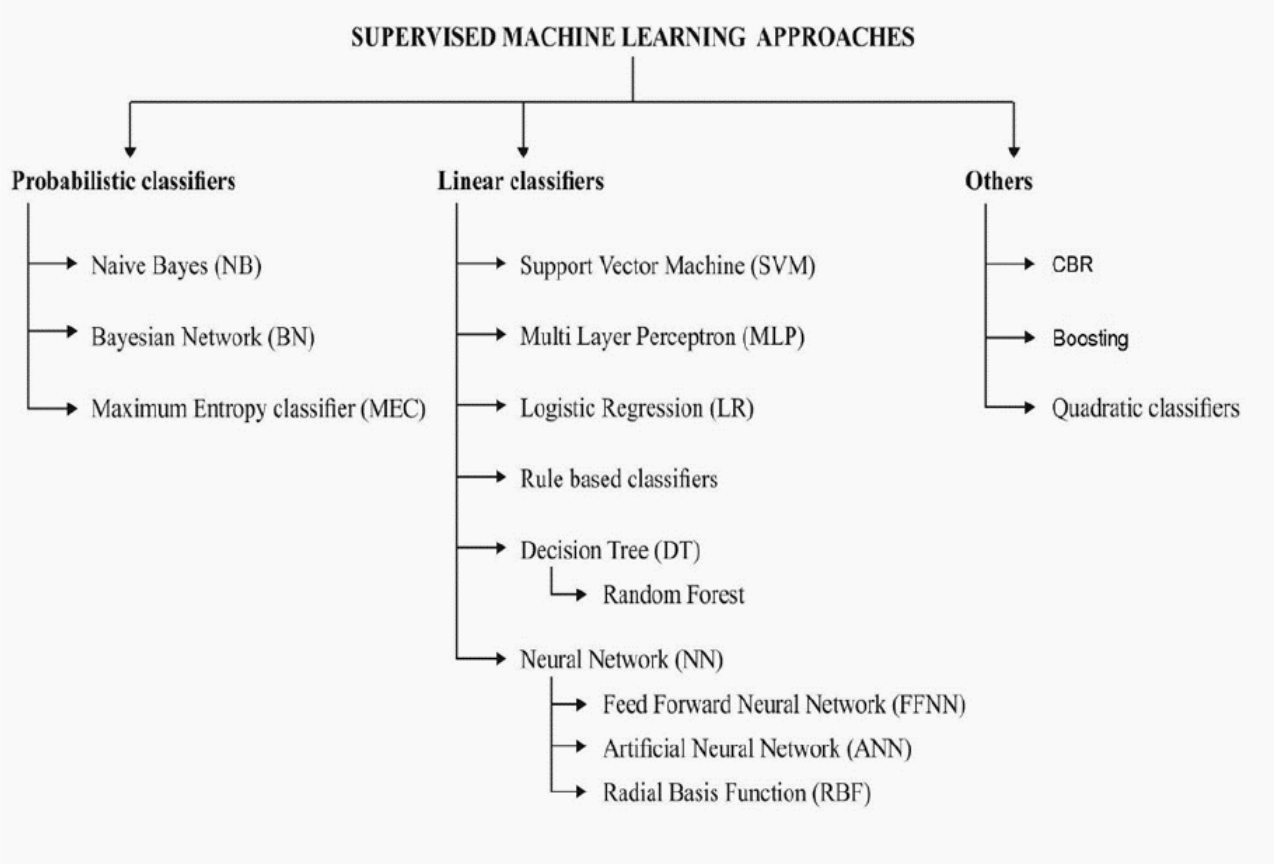

Figura 2.1: Distintas técnicas de aprendizaje supervisado y su categorización [26].

Por una parte los clasificadores probabilísticos realizan su predicción mediante la probabilidad de un dato de pertenecer a una clase. Por ejemplo, el clasificador Naives Bayes implementa el Teorema de Bayes que se ve en la Ecuación 2.1 para calcular la probabilidad que una característica pertenezca a una clase/etiqueta [29]. Así mismo las Redes Bayesianas utilizan variables aleatorias y dependencia condicional para la computación de sus resultados [30]. El trabajo de David [43] utiliza un clasificador de Bayes para asignar tópicos y sentimientos a Tweets relacionados con el área científica (los cuales fueron etiquetados manualmente). Sin embargo, dada la baja cantidad de datos usados no se logró una precisión alta teniendo un promedio del 30 %.

Definición 2.1 *Teorema de Bayes*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

Los clasificadores lineales obtienen sus resultados a través de la combinación lineal de las características de los datos para obtener la clasificación de salida [31]. Entre las técnicas existentes de clasificadores lineales se encuentran las *Support Vector Machine* (SVM), *Multi-Layer Perceptron*, *Random Forest*, etc. Son estos tipos de clasificadores que podrían ser utilizados para resolver el problema de clasificación presentado en este trabajo.

En el trabajo de Kanish et al. [33] realizaron un estudio comparando 3 tipos de clasificadores para texto. Comparan el rendimiento de *Regresión Logística*, *KNN* y *Random Forest* utilizando como base de datos noticias de BBC. Estas noticias traen consigo 5 etiquetas distintas correspondientes al tópico que pertenecen. En cuanto a precisión y exactitud los 3 métodos obtuvieron buenos resultados con una precisión y exactitud promedio por sobre el 90 %. Esto implica que en caso de poder adquirir los datos etiquetados utilizar aprendizaje supervisado puede ser una gran forma de resolver problemas de clasificación. Sin embargo, otro aspecto importante a tener en este tipo de ML es que específicamente para texto se requiere una forma de representar el texto a valores numéricos con los cuales el sistema sea capaz de trabajar.

Tener una forma de representar el texto es vital para poder desarrollar un sistema clasificador utilizando ML. Para esto se debe vectorizar el texto que se desea utilizar ya sea utilizando un *word embedding*, frecuencia de término–frecuencia inversa de documento o *TF-IDF*(Term frequency–Inverse document frequency) o simplemente una bolsa de palabras combinado con conteo de éstas. Un *word embedding* es un espacio vectorial de baja dimensión en el cual se pueden representar palabras a través de vectores [34]. Ya que la palabra pasa a ser un vector es posible realizar operaciones matemáticas con estos tales como combinaciones lineales necesarias para el aprendizaje supervisado. Estos *word embedding* provienen del área NLP del área de AI y pueden ser desarrollados prácticamente para varios idiomas (mientras se tengan los datos). Otro aspecto interesante de los *word embeddings* es que pueden pertenecer a temas específicos incluyendo una mayor cantidad de palabras relacionadas a dicho tópico, sin embargo, un buen *word embedding* debería ser capaz de representar de forma genérica y ser aplicable a distintos tipos de problemas [35]. Entre uno de los problemas para los cuales pueden ser aplicados es para la detección y clasificación de sentimientos en textos, similar al problema planteado en este trabajo. Por otra parte TF-IDF hace uso de la ocurrencia de las palabras en el texto utilizado. Siguiendo esta línea, entre más veces se encuentre una palabra en un texto mayor frecuencia tendrá, pero entre menor sea la ocurrencia de una mayor importancia tendrá dentro del documento [36]. Esto en combinación con diccionarios es otra forma de representar el texto/palabras de forma numérica y que facilita a las técnicas de aprendizaje supervisado trabajar con este tipo de datos.

2.1.2. Aprendizaje no supervisado

Como se mencionó anteriormente el aprendizaje no supervisado es otro de los 2 grandes métodos de Machine Learning. Una de sus principales diferencias es que los datos utilizados no poseen etiquetas y la máquina busca categorizar los datos en distintos grupos o *clusters* en base a las características que los datos poseen [37]. Para este tipo de problemas también es importante saber cuáles *features* de los datos son esenciales para realizar el proceso de

clustering, teniendo en cuenta la naturaleza del problema y lo que se espera que solucione. El trabajo de Dy et al. [37] se explica con mayor detalle el proceso para selección de features en base a distintos criterios.

Al igual que para el caso supervisado, el método no supervisado posee una gran cantidad de técnicas para utilizar. En el trabajo de Memoona et al. [38] realizan un breve resumen de varias técnicas de aprendizaje no supervisado, entre las que se encuentran *K-Means* (bastante popular para clustering), *Probabilistic Latent Semantic Analysis*, *KEEL*, entre otros. Desde análisis de texto [39] hasta clustering de imágenes, reconocimiento de patrones [41] o incluso aprendizaje de otros algoritmos de aprendizaje [40], este método se pueden ajustar a varios tipos de problemas. Debido a que estas técnicas no utilizan datos etiquetados se debe buscar una forma de evaluar y validar los modelos entrenados [42]. La forma de evaluación dependerá principalmente del tipo del problema que se esté buscando resolver.

2.1.2.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) es una técnica de aprendizaje no supervisado para agrupar datos en base a sus características. Esta técnica es frecuentemente utilizada cuando se trabaja con texto, por ejemplo, descubrir los tópicos presentes en dichos textos [44], análisis de sentimientos en críticas realizadas por usuarios [45] e incluso para realizar un análisis espacial de imágenes de satélites [46].

Dado que LDA se puede utilizar en textos tiene sentido razonar que pueda ser utilizado para analizar datos extraídos de Twitter específicamente su componente de texto. Por ejemplo, en el trabajo de Sanandre et al. [47] realizan una modelado de tópicos para los Tweets realizados relacionados con al crisis financiera de la Universidad Nacional de Colombia. Utilizaron LDA para poder encontrar las temáticas dentro de los Tweets analizados y agrupar estos entre las 12 distintas temáticas. El modelado de tópicos resulta ser una aproximación interesante para el análisis de datos de Twitter. Esto se debe a que permite crear tópicos a partir de los datos que se pueden utilizar posteriormente para clasificar los Tweets entre estos tópicos y utilizar técnicas de ML supervisado [48].

LDA es un modelo generativo que utiliza probabilidades para realizar sus cálculos sobre el texto. En el trabajo de Blei et al.[49] explican detalladamente la matemática detrás de este modelo, sus aplicaciones, cálculo de parámetros, entre otros. Brevemente, se tiene que un documento puede ser representado como un tópico (aleatorio) donde cada tópico está compuesto por una distribución por sobre palabras. Las palabras son representadas como un vector con valor 1 en la posición correspondiente a la palabra dentro del vocabulario y 0 en todo el resto del vector. La probabilidad de corpus D conformado por M documentos con N_d palabras se muestra en la Ecuación 2.6. Las variables α y β son variables correspondientes al corpus utilizado y pueden calcularse. Las variables z_{dn} y w_{dn} corresponden a los tópicos posibles para asignarse a los documentos y las palabras que conforman estos documentos respectivamente. Finalmente la variable θ_d es una variable aleatoria de Dirichlet de dimensión k.

Definición 2.2 *Probabilidad calculada para un corpus D*

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

En la Figura 2.2 se observa una representación visual del modo de operación de LDA[50].

El Corpus pasa por el primer parámetro de Dirichlet α y comienza la iteración por cada documento iniciando con el parámetro θ. A continuación se comienza a analizar cada documento iniciando por la asignación de tópicos z por palabras y finalmente se observan las palabras w dentro del documento, al mismo tiempo el parámetro β se aplica para la distribución entre los tópicos y sus palabras. El resultado de este proceso entrega tanto los distintos tópicos encontrados (con la cantidad de tópicos definidos previamente), y la distribución de cantidad de tópicos en todos los documentos analizados del corpus.

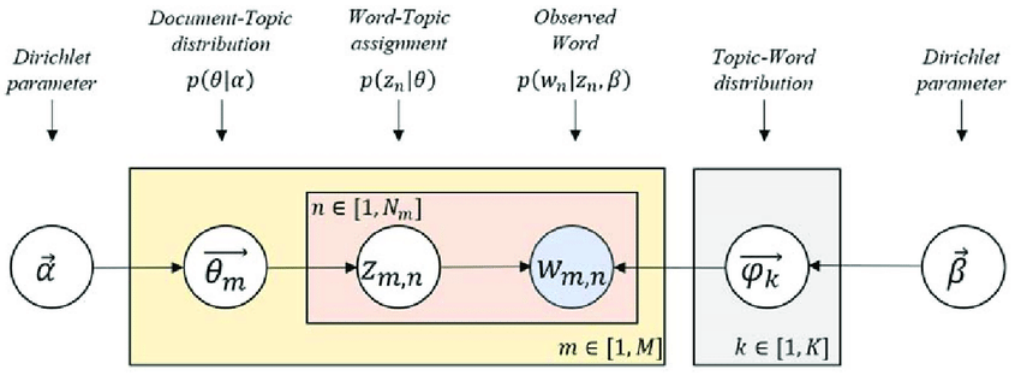

Figura 2.2: Esquema gráfico de LDA[50].

LDA puede obtener resultados bastantes interesantes sobre todo en el área del modelado de tópicos. Sin embargo, trabajar con datos de Twitter puede traer consigo ciertas complicaciones a tener en cuenta. Una de ellas es la cantidad de texto por Tweet dado que el límite es de 140 caracteres. Esto implica que cada documento del corpus está compuesto por un Tweet distinto y se debe considerar lo corto que este puede ser. Por otro lado también se tiene la presencia de datos no estructurados, enlaces, emojis (que pueden contener significado), abreviaciones utilizadas en el lenguaje, sarcasmo, entre otros, que limitan la cantidad de información que puede extraer LDA [51]. Así mismo, la repetición de palabras dentro de distintos Tweets puede implicar una repetición de estas en distintos tópicos. LDA puede obtener buenos resultados generando tópicos coherentes y con palabras representativas para estos, sin embargo, dependerá muchas veces del nivel de pre-procesamiento, ajuste de parámetros adecuado y datos utilizados. Por ejemplo en el trabajo de Guo et al. [52] obtuvieron como resultado que el método LDA entregaba mayor cantidad de falsos positivos a comparación de su método basado en diccionario que entregaba mayor falsos negativos

2.1.2.2. Distributed Dictionary Representations (DDR)

El método *Distributed Dictionary Representations* o DDR fue propuesto por Garten et al. [23] en 2017 y presentan otra forma para operar con los diccionarios. A grandes rasgos DDR busca representar el concepto de una colección de palabras a través del promedio de su representación en un espacio semántico. Este promedio se puede utilizar para comparar con otros documentos y calcular la similitud que estos presentan al concepto obtenido. Puede ser utilizado para el análisis de sentimientos en documentos e incluso la creación de diccionarios a partir de la comparación de palabras que lo conforman con nuevas que se desean agregar.

El método busca encontrar el concepto que es representado por las palabras que conforman un diccionario. Esto implica que las palabras que conformen dichos diccionarios deben ser preferentemente representativas del concepto, y evitar aquellas que puedan ser utilizadas por otros conceptos en diferentes contextos. Esto permite que el investigador que utilice DDR se

enfoque principalmente en la creación de un diccionario consistente con el término que busca interpretar en lugar de considerar cada palabra posible que se relacione a este. También permite utilizar este tipo de técnica en documentos con una menor cantidad de palabras, característica bastante interesante considerando la gran cantidad de investigaciones que se han realizado en el último tiempo utilizando datos de redes sociales que se caracterizan por una menor cantidad de texto. Algunos ejemplos de esto último vendrían siendo los trabajos realizados por Chen et al. [48] o Yang et al. [51], además del trabajo realizado en esta memoria.

Para poder realizar la tarea anteriormente mencionada DDR necesita de una representación distribuida R de texto n-dimensional (como los *words embeddings*) en el cual interpretar las palabras como vectores de dimensión n y poder realizar los cálculos matemáticos correspondientes. Este espacio vectorial puede ser específico para el tipo de problema en el que se está utilizando o uno más general entrenado en un corpus con una gran cantidad de oraciones.

Más detalladamente, el razonamiento detrás de DDR sigue los siguientes pasos:

1. Se crea un diccionario D de m palabras.

   Definición 2.3 *Diccionario D*

   $$D = [w_1, w_2, ..., w_m]$$

2. Se obtienen los vectores de cada palabra del diccionario representados en el espacio vectorial R de dimensión n con vocabulario V.

   Definición 2.4 *Representación de las palabras*

   $$R(w_i) = [d_1, d_2, ..., d_n], i = [1, m] \forall w \in V$$

3. Para obtener el concepto representado, se promedian los vectores de cada palabra que conforman el diccionario. Este resultado se normaliza para reducir el nivel de error a la hora de calcular las distancias.

   Definición 2.5 *Concepto representado*

   $$C_R = \frac{\sum_{\forall w \in D_R} R(w)}{\parallel \sum_{\forall w \in D_R} R(w) \parallel}$$

4. Se realiza el mismo procedimiento para el texto con el que se desea realizar la comparación, obteniendo T_R.

5. Se calcula la distancia entre los obtenidos para C_R y T_R utilizando la similitud coseno. Dado que se normalizaron los vectores previamente, la expresión se reduce simplemente a $cos\theta = C_R * T_R$

   Definición 2.6 *Similitud coseno*

   $$cos\theta = \frac{x * y}{\parallel x \parallel \parallel y \parallel}$$

6. El valor obtenido de la similitud coseno tiene un rango entre -1 y 1, donde 1 indica

que ambos conceptos son ampliamente similares, y un valor de -1 indica que ambos conceptos son completamente distintos.

Una de las ventajas que posee DDR por sobre el conteo de palabras es que añadir términos de alta frecuencia no afecta en mayor parte el concepto que busca representar el diccionario. Sin embargo, añadir una gran cantidad de términos no tan relacionados con lo que se intenta representar terminaría entregando resultados distintos utilizando DDR, dado que la representación del concepto del diccionario cambiaría junto con estos términos.

Los resultados obtenidos en su trabajo indican que DDR es una buena herramienta para ayudar a métodos existentes a la hora de tratar con textos cortos. Diccionarios con aproximadamente 30 palabras tienden a tener un mejor rendimiento, dado que se concentra principalmente a aquellas palabras importantes del concepto. El F1 score para algunos de los experimentos realizados ronda entre 0.75 y 0.8, lo que es bastante aceptable para un método principalmente no supervisado. Sin embargo, si se desea aplicar esta técnica para poder realizar la detección de tópicos es necesario validar correctamente los resultados obtenidos y realizar mejoras al algoritmo si es necesario. Estas mejoras se pueden realizar, por ejemplo, incluyendo otras técnicas a la hora de analizar los tópicos.

2.2. Tópicos

En caso de utilizar un modelo de aprendizaje supervisado o DDR, se requerirá la definición previa de los diccionarios que representarán los diversos tópicos. En el trabajo realizado por De Clercq et al. [9] utilizan la taxonomía propuesta por el *International Press Telecommunications Council* (IPTC) para la caracterización de los tópicos. Uno de los objetivos que busca cumplir el IPTC es la estandarización para optimizar el manejo y transmisión de información [10]. Se fundó en Londres en el año 1965 y hoy en día cuenta con más de 60 compañías, asociaciones y organizaciones de medios noticieros. Entre sus miembros se encuentran organizaciones provenientes de Estados Unidos de América, Francia, China, España, entre otros [11].

Para el año 2010 el IPTC lanzó una taxonomía con más de 1200 términos enfocado principalmente en la clasificación de los textos noticieros. Esta taxonomía se categoriza por niveles que van desde el nivel 1 al nivel 5 donde en el primer nivel se encuentran aquellos términos que abarcan una gran cantidad de sub-términos [12]. Los términos tienen asociados un código único facilitando la asignación de estos a distintos textos, además de poder contar con un enlace a la página de *Wikidata.org* de dicho texto. Wikidata [13]es una base de conocimiento gratuita y de libre acceso, conteniendo items de información que tanto humanos como máquinas pueden editar. Los términos dispuestos por el IPTC se encuentran en 12 idiomas, incluyendo el español e inglés que viene a ser de utilidad a la de utilizar esta estandarización.

La división de tópicos dispuesta por el IPTC cuenta con 17 términos en el nivel más alto (nivel 1). Estos corresponden a la clasificación más general posible para el tópico que representa una noticia. Estos tópicos en conjunto a su definición otorgada por el IPTC son los siguientes [14]:

- Arte, Cultura, Entretenimiento y Medios (ACEM): Todas las formas de arte, entretenimiento cultura y medios.
- Catástrofes y accidentes (Catástrofes): Acontecimiento natural o provocado por humanos que resulta en la pérdida de vida o lesiones a criaturas vivas y/o daño a objetos o propiedades.

- Ciencia y tecnología (CyT): Todos los campos relacionados con la comprensión humana, así como el estudio metódico y la investigación de las ciencias naturales, formales y sociales, como la astronomía, la lingüística o la economía.
- Conflicto, guerra y paz (Conflicto): Actos de protesta o de violencia cuyos motivos pueden ser sociales o políticos, actividades militares, conflictos geopolíticos, esfuerzos de resolución de conflictos.
- Deporte: Actividad competitiva o habilidad que involucre un esfuerzo físico y7o mental y organizaciones involucrados en dichas actividades.
- Economía, negocios y finanzas: Todo asunto relacionado con planificación, producción e intercambio de riquezas.
- Educación: Todo aspecto relacionado la fomentación de conocimiento, tanto formal como informal.
- Estilo de vida y tiempo libre (Estilo vida): Aquellas actividades realizadas por placer, relajación, recreación fuera de empleo remunerado, incluyendo comer y viajar.
- Interés humano, animales, insólito (Interés humano): Artículos sobre individuos, grupos, animales, plantas u otros objetos con enfoque emocional.
- Mano de obra (Trabajo): Aspectos sociales, organizaciones, reglas y condiciones que inciden en el empleo del esfuerzo humano para la generación de riquezas o provisiones de servicios y el sustento económico de los desempleados.
- Medio ambiente: Todos los aspectos sobre protección, daño y condición del ecosistema del planeta Tierra y su entorno.
- Meteorología: Estudio, observación y pronóstico de fenómenos meteorológicos.
- Policía y justicia (Justicia): El establecimiento y/o declaración de reglas de conducta en la sociedad, la aplicación e dichas reglas, infracciones a estas, el castigo de los infractores y organizaciones involucrados en estas actividades.
- Política: Ejercicio de poder, local, regional, nacional e internacional, o la lucha por el poder, y las relaciones entre entidades de gobierno y estados.
- Religión y culto (Religión): Todo los aspectos de la existencia humana que involucran teología, filosofía, ética y espiritualidad.
- Salud: Todos los aspectos relacionados al bienestar físico y mental de seres vivos.
- Sociedad: Preocupaciones, problemas, asuntos e instituciones relevantes en la interacción social humana, los problemas y el bienestar, tales como pobreza, derechos humanos y planificación familiar.

De estos 17 tópicos generales se tienen 127 tópicos en el nivel 2 que detallan distintos aspectos de los tópicos generales. Sin embargo, no todos los tópicos de segundo nivel tienen necesariamente una sub-división. La forma en que cada término se encuentra identificado en la página del IPTC se puede observar en la Figura 2.3. En esta figura se puede apreciar el

nivel que le asigna el IPTC al concepto mediante los cinco recuadros en la sección lateral izquierda donde la cantidad de asteriscos indican el nivel del concepto. Además, se puede ver el código asociado al concepto como, por ejemplo, Política que tiene el código 11000000. Se observa también en Elecciones un enlace que direcciona al concepto en Wikidata, al concepto más amplio y a conceptos relacionados.

Figura 2.3: Visualización de los conceptos en la página del IPTC

Un ejemplo de visualización en forma de árbol se puede apreciar en la Figura 2.4 [15]. En esta imagen se aprecia cómo se subdividen los conceptos y se tienen los 5 niveles posibles de categorización. Para el ejemplo mostrado se tiene que la avalancha es un tipo de deslizamiento de terreno que a su vez es un desastre natural y que a su vez es un desastre general.

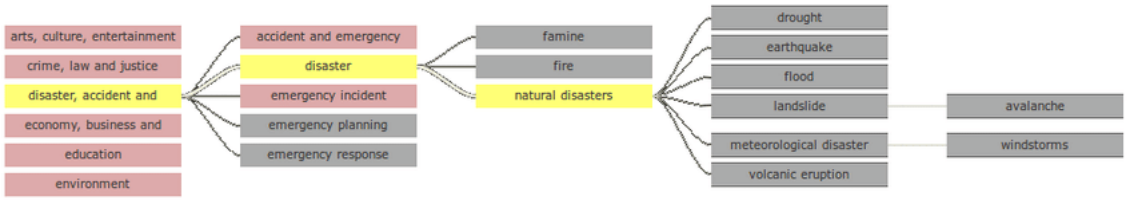

Figura 2.4: Visualización en forma de árbol de los conceptos estandarizados del IPTC

2.3. Trabajos similares

Una basta cantidad de investigaciones se han realizado utilizando datos de Twitte, dada la gran cantidad de información que se puede extraer de un correcto análisis de estos y la gran variedad de problemas en los que se pueden utilizar. Por ejemplo, como se mencionó en la sección de LDA, Sanandres, et al. [47] utilizaron LDA para encontrar tópicos dentro de una base de datos de Twitter y observar si se encontraban tópico relacionados a la crisis financiera que afrontaba la Universidad Nacional de Colombia. Es un problema que en esencia es similar al que se propone en este trabajo, pero con la diferencia que en el trabajo de Sandares simplemente se buscan encontrar tópicos relacionados a la crisis. En cambio, en este trabajo se proponen clasificar noticias en una mayor variedad de tópicos y tener un registro temporal de estos, sumado a todas las comparaciones y agrupaciones que se pueden realizar posteriormente.

Un ejemplo de análisis de noticieros se puede observar en el trabajo realizado por Sáez-Trumper, et al. [8], en el que utilizan la información de 80 noticieros (algunos de Twitter)

y utilizan técnicas de aprendizaje no supervisados para descubrir el tipo de sesgos que estos poseen. Uno de los aspectos interesantes de este trabajo es que se enfocan en las características que poseen las cuentas a partir de las noticias que publican, y definir si poseen un sesgo selectivo, de cobertura o declarativo. Uno de los objetivos del Trabajo de titulo es poder caracterizar los noticieros a partir de la cobertura que se les da a los tópicos, por lo que, el trabajo de Sáez-Trumper es una de las fuentes de inspiración para este objetivo y la razón por la que se busca obtener un gráfico que muestre agrupados las distintas cuentas noticieras dependiendo de los tópicos que publican.

Finalmente, en el trabajo de Lee, et al. [61] donde buscan clasificar los *trends* que van surgiendo en Twitter en 18 tópicos. Este trabajo es bastante similar al propuesto como tesis, con la diferencia que esta enfocado a tweets en inglés. Otra diferencia visible es que la clasificación de tópicos se realiza a todos los tweets que se van publicando, y no se realiza un seguimiento a noticieros. También, en el trabajo de Lee tienen diferentes tópicos a los propuestos del IPTC y utilizan 2 técnicas distintas de aprendizaje de máquinas: 1 supervisada(Clasificador de Naive Bayes Multinomial) y otra no supervisada (Bolsa de palabras), donde se logró una precisión del 65 % y 70 % respectivamente. Esto último también ayuda a tener una idea del nivel de precisión que se puede aspirar a conseguir en este tipo de problemas.

Ya se tiene explicado los distintos algoritmos disponibles a utilizar para la clasificación de las noticias en distintos tópicos. Además, se explicó una de las formas para la definición de distintos tópicos utilizando la estandarización propuesta por el IPTC. Se puede proceder al diseño de las distintas partes que conformarán el sistema, realizar las pruebas correspondientes y mostrar los resultados de dichas pruebas.

Capítulo 3

Datos

3.1. Datos

Para el desarrollo del Trabajo de Título y lograr cumplir el objetivo principal es necesario la utilización de datos. Estos datos corresponden a las noticias que distintos medios de noticias publican a través de su plataforma. Debido a esto una parte bastante importante del trabajo consiste en construir la base de datos que se utilizará para el desarrollo de este trabajo. Se seguirán cuentas de noticieros chilenos que se encuentren en la plataforma de Twitter como se ha mencionado previamente.

La motivación de usar Twitter por sobre otras plataformas reside en que hay una gran parte de trabajos realizados que trabajan en la clasificación o agrupación de tweets, por ejemplo *Theme Based Clustering of Tweets* [16]. Para poder obtener los datos es necesario crear una cuenta de desarrollador de Twitter. Con esta cuenta se tiene los *tokens* y *llaves* necesarias para acceder a la *Application Programming Interface* (API) lo cual resulta vital para obtener los tweets.

Aparte de la cuenta de desarrollador se necesita un *wrapper* que pueda extraer la información solicitada. Estos *wrappers* trabajan directamente con la API de Twitter [17], y varían en los lenguajes de programación en los que están desarrollados. Continuando con lo anterior, el lenguaje a utilizar para el diseño del sistema es *Python*. Una de las ventajas presente al utilizar Python es la amplia cantidad de librerías dedicadas para el procesamiento de texto y lenguaje, aprendizaje de máquinas, manejo de tablas de datos, y en esta caso particular *DDR*. Entre los wrapper para Python que trabajan con la API de Twitter se decidió utilizar Tweepy. Esta herramienta funciona en las versiones de Python que van desde la 3.6 hasta la 3.10 [18]. Cuenta con una vasta documentación para todos sus métodos junto con la referencia correspondiente a la página de desarrollador de Twitter. Esto también facilita el entender los resultados que entrega Tweepy en base a lo que Twitter informa en su plataforma.

Una vez resuelto el problema de cómo obtener los datos, se debe decidir cuáles son las cuentas a las que se les realizará un seguimiento. Se desean seguir medios noticieros chilenos y para esto se debería realizar una exploración por Twitter para descubrir varias de estas cuentas. En el trabajo realizado por Elejalde et al. [19] se realiza un análisis a cuentas de noticieros chilenos en Twitter y como presentan un sesgo en base a la ubicación geográfica de sus seguidores. Dado que trabajan en el contexto chileno, implica que tienen una lista de todos las cuentas que utilizaron para su análisis. Cuentan con una lista de 404 cuentas chilenas para el 27 de noviembre de 2018. Se utilizó esta lista como guía para añadir más cuentas creadas recientemente y descartar aquellas que hayan sido eliminadas de la plataforma entre los años

2018 y 2021. Después de realizar el filtrado de cuentas se obtuvo una lista con 375 cuentas que varían entre tradicionales y alternativo ubicados a lo largo de todo el país.

Se inició el periodo de extracción de datos a partir del 09 de agosto de 2021 y se terminó de extraer información el 08 de noviembre de 2021 con una separación de 1 semana por cada fecha de extracción, para un total de 12 semanas de tweets. Durante este periodo se buscó extraer la cantidad máxima posibles de Tweets desde el día de extracción hacia fechas anteriores. Existen limitaciones a la hora de utilizar la API de Twitter y es por esto que en las cuentas que usualmente publican en mayor cantidad se logró extraer aproximadamente 3000 tweets cada semana. Esto se debe a que en la documentación de la API se detalla que para el método *timeline* se puede hacer 15 peticiones por usuario, en una cantidad de hasta 200 veces por cuenta [20].

Una vez se tienen los datos se realiza un análisis exploratorio de estos buscando identificar palabras con alta frecuencia, cantidad de datos, información de los Tweets, entre otros. De las 375 cuentas de las que se extrajo información se conservaron los datos de 366. Esto se debe a que algunas de estas cuentas habían sido desactivadas, y de otras no se logró obtener ningún Tweet. En la Tabla 3.1 se observa la cantidad total de publicaciones sin repetición obtenidas es aproximadamente 1,6 millones. Además, se tiene el rango existente entre la publicación más antigua del conjunto de datos y la más reciente (que resulta ser la última fecha de extracción).

Tabla 3.1: Información acerca de la base de datos conformada por los Tweets obtenidos

Cantidad de TWeets	1.571.418
Fecha más reciente	08 de noviembre de 2021
Fecha más antigua	04 de mayo de 2009

Para el procesamiento de texto se transformaron todos los caracteres del alfabeto a minúscula. Además se eliminaron todos los símbolos pertenecientes a otros idiomas que no fueran el español, signos dobles, símbolos como el *hashtag(#)*, emoticones y los enlaces presentes en las publicaciones. De esta forma la publicación final solamente consiste en el texto que conforman los Tweets en minúscula y de esta forma se evalúan todas las publicaciones bajo un mismo estándar.

Textos tanto en español o inglés están conformados por una gran variedad de palabras, sin embargo, algunas de éstas son utilizadas con gran frecuencia debido a que son palabras auxiliares. Los pronombres, conectores y preposiciones no otorgan mayor información a la hora de referirse a un tópico, pero le entregan coherencia a las oraciones. Es por esto que si se quiere trabajar con texto una buena práctica es eliminar este tipo de palabras que no otorgan mayor información al tópico en el que está insertado, también conocidas dentro del área del procesamiento de lenguaje como *STOPWORDS*. En el paquete *Natural Language Tool Kit (NLTK)* de Python viene incorporado una librería con stopwords en español junto con un método para filtrar estas palabras en el texto deseado. Además, se pueden extender esta librería agregando manualmente otras palabras que se desearan filtrar dependiendo del contexto del trabajo realizado [22].

Al realizar un análisis exploratorio de los datos más profundo se encontraron particularidades en los Tweets. Una de estas es la presencia de palabras muy frecuentes que no aportan en gran medida a la distinción de los tópicos. Otro tipo de palabra frecuencia corresponde a aquellas que hacen referencia a la página de la que provienen. Estas palabras están nor-

malmente asociadas a un hashtag lo que implica que son palabras concatenadas, pero que solamente hacen mención a la cuenta autora del Tweet. Se realizaron nubes de palabras para observar gráficamente cuáles eran estas palabras frecuentes y si era posible visibilizar tópicos presentes en la cuentas una vez realizado este filtrado.

Las siguientes nubes de palabras se obtuvieron utilizando la librería *WordCloud* para Python, en conjunto con las *stopwords* otorgadas por NLTK. En la Figura 3.1 se observa la nube de palabras obtenida para la cuenta *@24horastvn*. En esta Figura se puede observar que palabras como vivo, sigue, señal detalles, envivo, entre otras, tienen una gran representación a lo largo de las publicaciones. Analizando las palabras más pequeñas se puede obtener la presencia de algún tópico, por ejemplo la presencia de *Variante Delta* o *Copa América*. Cabe destacar palabaras como *vivo* o *24play* no aportan mayor información al tópico del que se habla en la noticia, pero si invitan al usuario a visitar el enlace anexado al Tweet o informan que es un suceso ocurriendo en el momento. Sin embargo, también existen cuentas como *@rsumen* cuya nube de palabras se puede ver en la Figura 3.2. Esta cuenta no presenta ninguna de las palabras mencionadas anteriormente por lo que existen casos donde añadir estas nuevas stopwords no afectan el texto del Tweet.

Figura 3.1: Nube de palabras de la cuenta @24horastvn sin stopwords añadidas

La lista de stopwords agregadas hasta el momento está conformada por:

- señal, vivo, detalles, detalle, comenta, aquí, sigue, link, gt, ver, envivo, así, conversamos, video, |, →, », lo, hoy, ahora, mañana, día, todos, revisa.

Posterior a agregar estas nuevas palabras se procede a generar nuevamente nubes de palabras para observar los cambios en cuentas que utilizan las stopwords añadidas. En la Figura 3.3 se puede observar la nube de palabras de la cuenta @24horastvn después de filtrar las palabras mencionadas anteriormente. Se puede ver que en esta ocasión si se aprecia de manera más sencilla los tópicos como Copa América o Variante Delta/COVID. Sin embargo, las palabras como *24play*, *24horascentral*, entre otras, siguen apareciendo en una medida

Figura 3.2: Nube de palabras de la cuenta @rsumen sin stopwords añadidas

considerable. En cuentas como la de @rsumen se observa que este tipo de palabras de auto-referencia no suelen aparecer por lo que no se considerará para la siguiente etapa.

Figura 3.3: Nube de palabras de la cuenta @24horastvn con nuevas stop-words añadidas

Se procede a eliminar estas palabras de auto-referencia y observar si otras palabras pasan a ser más prominentes. La eliminación de estas palabras no se realiza para el filtrado final del trabajo, solamente se realiza para observar el comportamiento que tiene esta cuenta y poder extender este análisis a otras cuentas similares. La razón para conservar estas palabras es que son un estándar para la cuenta, y se evita tener que actualizar la lista de palabras a borrar

en caso del surgimiento de nuevas. La Figura 3.4 muestra la nube de palabras obtenidas para la cuenta de @24horastvn, donde claramente se ve la influencia de eliminar las palabras de auto-referencia dado de la presencia más evidente de Variante Delta.

Figura 3.4: Nube de palabras de la cuenta @24horastvn con palabras de auto-referencia eliminadas

Finalmente queda un sólo elemento de los Tweets que falta por estudiar y es el hashtag. Anteriormente se mencionó que se eliminó el símbolo #, pero no se eliminó la palabra asociado a ese símbolo. Esto implica que se conservó una palabra extra que se incluye dentro del texto analizar lo que puede resultar en un aumento de la información con respecto al tópico a tratar. El problema nace con los hashtags que son formados por palabras concatenadas. Para palabras como 24horascentral no aportan información para el tópico de la noticia. Sin embargo, con palabras concatenadas como *convencionpuedes* o *avanceantiviral* es visible que la separación de estas palabras influye en una ganancia a la información del tema tratado que en este caso sería la convención constitucional y salud respectivamente. Resolver este tipo de tareas no es trivial dado que otras palabras presentes en el texto como *constitucionalidad* podría separarse en *constitucion* y *alidad*. Por lo tanto, encontrar una forma de separar las palabras concatenadas sin afectar el resto del texto es una tarea que podría mejorar la detección de tópicos presentes en las noticias.

Capítulo 4

Diseño, Visualización y Métricas

4.1. Tópicos

Como se mencionó en el capítulo anterior, si se desea utilizar DDR o algún tipo de aprendizaje supervisado es necesario conocer las etiquetas/diccionarios que se desean utilizar. En ambos casos se utilizará, ya sea, la taxonomía o definiciones otorgadas por el IPTC. En el caso de utilizar aprendizaje supervisado se utilizarían las 17 tópicos de nivel uno para asignar las etiquetas a las noticias. En el caso de utilizar DDR se necesitan definir los diccionarios que se utilizarán para representar los diferentes tópicos. En este segundo caso se optó por utilizar los 127 tópicos de segundo nivel, y posteriormente asociarlos a su término de primer nivel.

Dada la falta de conocimiento en ciertas temática y la posible superposición con otros tópicos se utilizaron solamente 88 de los 127 términos de segundo nivel. La motivación detrás de utilizar los términos de segundo de nivel se debe a 2 motivos. El primero de estos corresponde a la mejor representación de tópicos presentes en los textos, es decir, algo que habla netamente de elecciones será detectado con mayor facilidad. La segunda motivación se complementa con la primera, y es que permite a los diccionarios ser mucho más específicos con el concepto que intentan representar y con esto poder representar más fielmente la noticia. La lista de los tópicos (generales y específicos) se puede encontrar en la sección de Anexos, Capítulo *Diccionarios utilizados*. La lista de los tópicos no utilizados se puede encontrar en la sección de Anexos, Capítulo *Términos no utilizados/fusionados* Tabla A.1. La selección de los tópicos no utilizados fue un proceso manual, y entre los principales motivos para no utilizar tópicos cómo *Diálogo interreligioso*, *Transferencia deportiva* o *Estudiante* se encuentran el desconocimiento con respecto a estas áreas, lo confusos que pueden ser en definición y su superposición con otros tópicos. Por ejemplo, para conformar el tópico de *Diálogo interreligiosos* resulta complejo encontrar palabras que hagan referencia a esta temática y que al mismo tiempo no sobrepongan con los otros tópicos de Religión. Para esto también se utilizó como referencia la definición entregada por el IPTC para cada uno de estos tópicos y en caso de encontrar que dicha definición complicará la construcción de un diccionario para éste tópico, simplemente se saltaba o buscaba si se podían juntar 2 tópicos que tuvieran una gran cantidad de palabras en común, como es el caso *Emigración e Inmigración*.

Los diccionarios se crean en archivos *txt* que es el formato que acepta la librería de DDR en Python. Esta librería exige que el nombre del concepto a representar sea el nombre del archivo de texto, que las palabras que conforman este diccionario estén escritas en la primera línea del archivo separadas por espacio y sin salto entre líneas. En la Tabla 4.1 se puede

observar el diccionario escrito para el tópico de elecciones y que sigue el formato mencionado anteriormente. Es importante notar que la inclusión de las palabras *elección_local* y *elección_nacional* a pesar de que *elección* ya se encuentra en el diccionario. Esto se realiza principalmente para llevar el promedio calculado del diccionario a un concepto de elección política, que es el contexto en cuál se enmarca este diccionario.Cabe destacar que en el trabajo de Garten et al. [23], llegan a la conclusión que la cantidad óptima de palabras que conforma un diccionario que utiliza DDR es de 30 palabras. Por lo tanto, una mejora de diccionarios es una posible alternativa si se busca mejorar el rendimiento. Pero se debe tener en consideración que las palabras que conformen estos diccionarios sean representativas del tópico que representan y evitar repetir varias palabras en distintos de estos diccionarios. El resto de diccionarios se mostrará en la sección de Anexos, cada uno con su macro tópico respectivo.

Tabla 4.1: Ejemplo de diccionario utilizado para la implementación de DDR.

Diccionario	Palabras
y Elecciones	campaña candidato debate elección elección_local elección_nacional revocación renovación elecciones_primarias referendo sistema_electoral votación

4.2. Técnicas utilizadas

Este capítulo se centrará en documentar el desarrollo del algoritmo para determinar los tópicos de las noticias y los detalles para visualización de los resultados. Como se ha mencionado anteriormente el lenguaje de programación a utilizar es Python 3. Existe gran variedad de librería implementando las distintas técnicas de machine learning, como LDA, BERTopic, Support Vector Machines e incluso DDR. Esta última posee librería únicamente para Python 2, por lo que se realizaron los cambios correspondientes para utilizarse en Python 3.

Para realizar la tarea de clasificación se utilizarán principalmente algoritmos de aprendizaje no supervisado. Entre los principales motivos para esta decisión se encuentran la gran cantidad de datos con los que se debe trabajar y la cantidad de tiempo que tomaría encontrar las etiquetas correctas para cada noticia dentro de las 1,5 millones aproximados con las que se está trabajando. De haber contado con una base de datos previa con datos etiquetados se hubiera optado por trabajar con técnicas supervisadas dada la facilidad para su evaluación y mejoras para obtener buenos resultados.

Para implementar LDA simplemente se instaló la librería *gensim* para Python que tiene incorporado el método LDA. Para utilizar LDA y cualquiera de los métodos a continuación se utiliza el texto filtrado de los Tweets de las cuentas que se obtuvieron en el capítulo anterior. Se realizaron 2 experimentos, uno utilizando LDA con la información de las cuentas por separado a modo de observar las temáticas que trataba cada una. El primer experimento consistió en el uso total de los datos de todas las cuentas para obtener los tópicos que se hablan en general en los medios. En el segundo experimento se utilizó los datos de cada cuenta por separado y se fijó que el número máximo de grupos ha formar fueran 17 para observar si el consigue algo similar a los 17 tópicos de primer nivel según IPTC. Entre las ventajas que tiene aplicar este tipo de técnica es poder encontrar temáticas habladas en la actualidad y que se pueden ver representados en los tópicos, como se mencionó en el capítulo 2. Sin embargo, una complicación que traería sería la dificultad para visualizar la cobertura

general de los tópicos que dan los noticieros. Además, posterior a obtener los clusters con los tópicos se agregó a los datos el tópico predicho y las palabras que conforman dicho cluster.

Para implementar DDR se necesita de un word embedding que se utilizará para representar las palabras como vector. El utilizado para la realización de este trabajo es el *Spanish Billion Words Corpus and Embeddings* de Cristian Cardellino [53]. Este word embedding está compuesto por más de un millón de tokens únicos, con una dimensión de 300 y fue entrenado en un corpus compuesto por fuentes en español. Existe una gran variedad de embeddings en español, como por ejemplo *BETO* [54], y esta la opción de ir probándolos para encontrar aquél que mejor se adapte al problema. Se implementa utilizando el método *word2vec* que viene incorporada con la librería de *Gensim* para Python. Por lo tanto, una vez descargo el word embedding simplemente se carga utilizando *word2vec* y se puede utilizar para la representación de palabras en español.

Como ya se tienen tanto los diccionarios para los 17 tópicos, los datos filtrados y la librería DDR implementada en Python 3 se puede proceder a diseñar el resto del programa. Simplemente se implementa el método *make_agg_vec* de esta librería para obtener los vectores representados de dimensión n (n=300) de los Tweets y de los diccionarios. Posteriormente, se utiliza el método *get_loadings* para obtener la similitud de cada Tweet (separados según la cuenta) a los 88 tópicos de segundo nivel, donde este valor varía entre 0 y 1. Con este valor ya se puede comenzar a discernir entre cuáles de los tópicos el titular analizado tiende a enfocarse más. Se procede a seleccionar aquellos que sean más representantes para el Tweet. Se selecciona el tópico con mayor relevancia que esté con un valor de similitud por sobre 0,65 asegurándose que la selección de los documentos sea lo más filtrada posible. También se puede realizar un proceso de normalización de los valores entregados por DDR a la hora de elegir el tópico que le corresponde al titular utilizando como umbral la desviación estándar (entre más alta, más se parece el tópico). Sin embargo, el código utilizado para la implementación de DDR ya tiene considerada la normalización de los datos. Una vez se tienen los tópicos de segundo nivel seleccionados se le agrega una nueva columna a la base de datos, incluyendo el tópico de primer nivel o *Macro tópico* dependiendo obviamente de cuál sea el asignado en el paso anterior. Se juntan todos los Tweets clasificados para poder extraer de forma aleatoria y con mayor imparcialidad los 100 titulares que serán utilizados para la evaluación del modelo. Además, también será utilizado para poder realizar un análisis global gráfico del porcentaje de tópicos cubiertos por los medios.

4.3. Visualización

Como se mencionó anteriormente la etapa de visualización toma en consideración diferente tipo de información a utilizar. La primera de estas consiste en un gráfico de barras que muestra la cantidad total de post realizados para cada tópico (que recibieron una etiqueta). El objetivo principal de este gráfico es mostrar como es la distribución total de tópicos tomando todas las cuentas de la muestra. De esta forma se puede saber cuáles tópicos tienen mayor representación de manera general en medios noticieros chilenos.

Para poder mostrar la variación de la cobertura de tópicos a través del tiempo se utilizó la información recopilada por cada cuenta por separado. Se toma la base de datos de una cuenta con los tópicos agregados y se agrupan las noticias según su fecha (semanal) y tópico. Posteriormente se saca la cantidad total de post realizados por semana y por tópico, y con esta información se calcula el porcentaje de representación que posee cada tópico en dicha semana. Esta información se plasma en un gráfico de tiempo mostrando como varía el porcentaje de

representación del tópico a lo largo de las semanas en la cuenta específica. Este tipo de visualización también permite relacionar el aumento en el porcentaje representado de un tópico con las noticias ocurridas durante un espacio de tiempo específico. Se agrega una leyenda con los 17 tópicos, cada uno con un color designado previamente evitando colores fuertes que puedan molestar al usuario. El gráfico es un archivo *.html* por lo que se abre en una ventana de navegador. Esto fue posible gracias a la librería *bokeh* de Python, que permite montar gráficos en navegador y poder lograr que estos sean interactivos. Gracias a esto se puede implementar la opción de mostrar tanto la fecha (por semana) como porcentaje exacto de un tópico ubicando el ratón sobre la línea de dicho tópico, logrando así una comparación mucho más precisa con otros tópicos dentro de la misma cuenta. Algo que no se ha podido implementar fue una leyenda interactiva, que fuese capaz de ocultar tópicos haciendo click sobre dichos tópicos. Esto puede ser de utilidad si se desea ver el progreso de algún tópico específico sin confundirse con el resto de los tópicos. Esto es simplemente un impedimento de la librería utilizada, y que podría actualizarse en el momento en que implementen un método para ocultar la leyenda [56]. Un ejemplo de gráfico que se busca obtener se puede ver en la Figura 4.1, donde se observa un gráfico de líneas con distintas temáticas y al mismo tiempo muestra información respecto al gráfico al ubicar el ratón sobre éste.

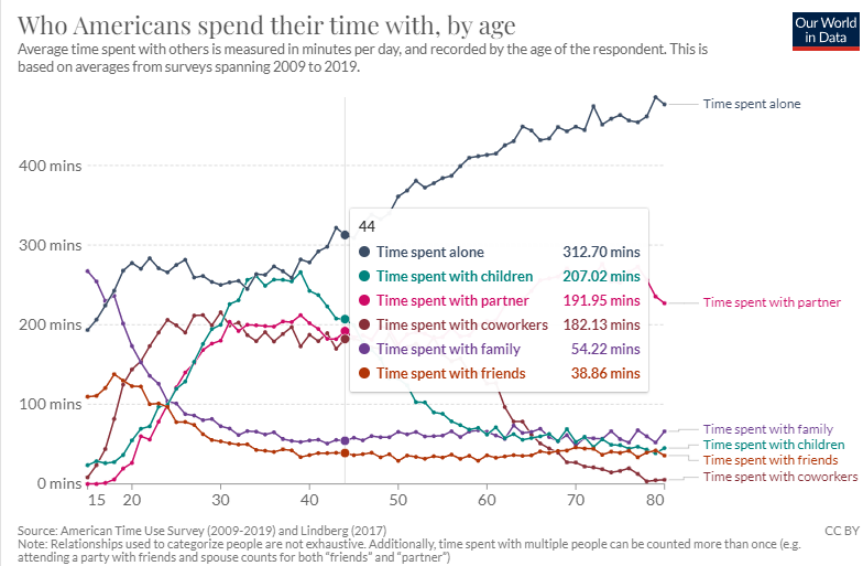

Figura 4.1: Formato de gráfico que se desea obtener que contiene leyenda, distinción por color e información en un punto de tiempo específico. La figura fue extraída de *Our World in Data*, de la sección *Who do we spend time with across our lifetime?*[55]

4.4. Métricas

Una vez se tienen los gráficos para la totalidad de los tópicos y los gráficos de variación temporal de las cuentas, se plantea mostrar visualmente una comparación entre cuentas. Específicamente, se plantea realizar un análisis de ventaja comparativa entre los diferentes noticieros. Para cumplir con esta tarea, se planean utilizar distintas métricas que utilizan los porcentajes de tópicos publicados para realizar estas comparaciones. El *Revealed Comparative Advantage* (RCA) o Índice de Balassa es una medida de comparación propuesta por Béla Balassa [57]. El RCA se rige por la Ecuación 4.1, donde i y k corresponde a los índices para

los países, j y l son los índices para las mercancías y T es el ítem exportado. Esta medida es utilizada principalmente en aspectos económicos como se pudo observar en su definición, pero puede ser extrapolado a otros contextos con una estructura similar. Los valores posibles para el RCA están en el orden de los reales positivos, donde se hace una diferenciación si se encuentran por sobre o bajo 1. Si el valor es menor que 1, implica que el ítem exportado presenta una desventaja frente a otros países con respecto a ese ítem exportado. Por otra parte si el valor es mayor que 1, implica que el país está x veces por sobre el promedio para el ítem en cuestión, donde x corresponde a la distancia del RCA a 1. El RCA se puede visualizar a través de tablas de ranking donde se toman las cuentas con el mayor RCA para cada tópico.

Definición 4.1 *Revealed Comparative Advantage (RCA)[57]*

$$RCA_{ij} = \frac{T_{ij} / \sum_{l \neq j} T_{il}}{\sum_{k \neq j} T_{kj} / \sum_{k \neq l} T_{kl}}$$

Se puede hacer un reemplazo en las definiciones de ciertas variables para hacer calzar la fórmula de RCA en el contexto de tópicos de noticias. En este caso i y k pasan a representar noticieros o cuentas, j y l representan los tópicos y T son las noticias/textos analizadas. Por lo tanto, el cálculo del RCA sigue la secuencia:

1. Del dataset de todos los titulares juntos, obtener la cantidad total de noticias (X_4).
2. Agrupar este dataset por tópicos y obtener la cantidad de noticias de cada uno (X_3).
3. Obtener la cantidad total de noticias para una cuenta c (X_2).
4. Agrupar las noticias por tópico j y obtener la cantidad de estas para cada uno (X_1).
5. Calcular el RCA para cada tópico j de la cuenta i con $RCA_{cp} = \frac{X_1/X_2}{X_3/X_4}$

A continuación se procede a agregar una columna extra al dataset que indica si el RCA del tópico es mayor o menor a 1. En caso de ser mayor a 1, en la columna *PASS* se agrega un 1 y se agrega un 0 en caso contrario. Este sistema binario es de utilidad a la hora de querer filtrar el dataset para conservar solamente aquellos tópicos que efectivamente sean mayor a 1. Con esta información se puede proceder a realizar una análisis de proximidad tanto para tópicos como para medios noticieros. Se utiliza la definición de proximidad dada por el *The Atlas of Economic Complexity: Mapping Paths to Prosperity* [58] donde la relación existente entre 2 productos es en base a la probabilidad en la que ambos sean exportados por el mismo país. Esta definición de proximidad utiliza RCA para sus cálculos, pero solamente aquellos que tengan valor por sobre 1. En la Ecuación 4.2 se muestra la definición matemática de proximidad utilizada, en esta p y p' corresponden a las mercancías, c corresponde al país, la matriz M posee valor 1 o 0 dependiendo si el RCA para el producto p es mayor a 1 o menor a un respectivamente y que se refiere a la ubicuidad. Ubicuidad se define en este mismo libro como la cantidad de países que producen un producto p y se define matemáticamente como la Ecuación 4.3.

Definición 4.2 *Proximidad [58]*

$$\theta_{pp'} = \frac{\sum_c M_{cp} M_{cp'}}{max(k_{p,0}, k_{p',0})}$$

Definición 4.3 *Ubicuidad [58]*

$$k_{p,0} = \sum_{c} M_{cp}$$

Ambas definiciones se pueden utilizar en el contexto del problema presentado en este trabajo al igual como se realizó con RCA. Si se desea analizar la proximidad entre todos los tópicos, se tendrá una matriz cuadrada de tamaño 17 (la cantidad total de tópicos de primer nivel). Para calcular la matriz de proximidad con tópicos se siguen los siguientes pasos:

1. Se agrupan todos los noticieros por tópico cuyo RCA sea mayor a 1, guardando como tablas dicha agrupación. Se crea la matriz donde se guardarán los datos de los cálculos realizados.

2. Para la tabla i (*i=[1,17]*):

    a) Se une con la tabla j (*j=[1,17]*) en su columna *Cuenta*, dejando aquellas que se encuentren en la tabla del tópico i como tópico j.

    b) Se utiliza el tamaño de esta tabla como numerador y se divide por el máximo entre el tamaño de las tablas i y j (la ubicuidad).

    c) Se reduce el valor de j en 1. De esta forma queda una matriz condensada con los valores de proximidad de los tópicos.

Para realizar el mismo procedimiento, pero buscando encontrar la proximidad entre las distintas cuentas, se deben realizar unas ligeras modificaciones. Se utiliza la información del RCA de los noticieros de forma separada. Cada noticiero es una tabla, por lo que, en lugar de ser una matriz cuadrada de tamaño 17 es una matriz cuadrada de tamaño igual a la cantidad de noticieros a los que se les pudo calcular al menos 1 tópico con RCA mayor a 1. Ahora la unión entre tablas se realiza para cada una de las cuentas y el *JOIN* se realiza en la columna de *Macro* (haciendo referencia al macro tópico o de primer nivel). Así se tienen los tópicos en los que ambas cuentan coinciden y son mayor que 1 y se usa el valor del tamaño de esta tabla como numerador en el cálculo de la proximidad. El valor de ubicuidad pasa a ser el tamaño de las tablas de cada cuenta utilizada en el cálculo de proximidad. Teniendo la matriz de proximidad en su forma condensada se pueden utilizar los métodos de la librería *Scipy* para obtener un dendrograma que muestre la proximidad existente entre las cuentas o tópicos de este trabajo y facilite entender esta proximidad. Un ejemplo de dendrograma se puede observar en la Figura 4.2, donde se muestra la similitud entre los distintos macro-invertebrados de un río en La Paz, Bolivia.

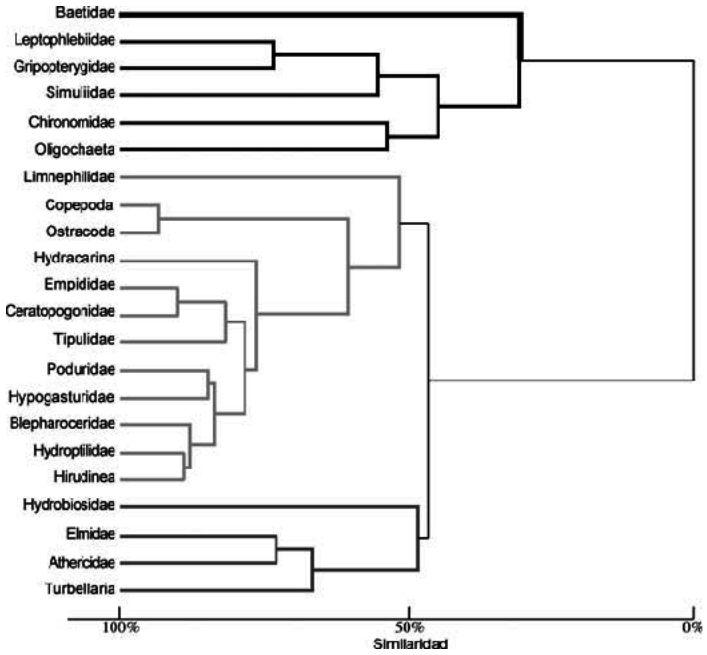

Figura 4.2: Formato de dendrograma que se desea obtener, incluyendo el nombre de los objetos de estudio que en este caso serían macro-invertebrados (en el trabajo serían cuentas noticieras). La figura fue extraída de la Figura 6 del paper: *Estructura de macroinvertebrados acuáticos en un río altoandino de la Cordillera Real, Bolivia: variación anual y longitudinal en relación a factores ambientales* [60]

Capítulo 5

Resultados

En este capítulo se revisarán los resultados obtenidos tanto en el proceso de clasificación de titulares, como el cálculo de RCA y proximidad. Además se mostrarán los resultados del proceso de visualización descrito en el capítulo anterior. Los resultados obtenidos de la implementación de LDA se pueden ver en la sección de Anexos C.

5.1. Distributed Dictionary Representations (DDR)

5.1.1. Validación del método utilizado

Se comenzará mostrando los resultados del proceso de validación del método utilizado. Verificar que el modelo presentado tiene un rendimiento aceptable es vital para asegurarse que los resultados representen los tópicos de las noticias. Como se mencionó en el capítulo 4 del conjunto de titulares clasificados se extraen un total de 100 noticias de forma aleatoria. Con este conjunto de titulares se crea una encuesta en donde el usuario debe asignar un tópico al titular presentado de entre los 17 tópicos de primer nivel y los 88 de segundo nivel (cada sub-tópico se encuentra asociado a su macro tópico). El encuestado asigna un tópico solamente a 15 titulares seleccionados aleatoriamente de los 100 del conjunto. Lo titulares pueden diferir entre encuestados, por lo que titulares pueden recibir muchas respuestas o ninguna. Una vez se tienen los resultados de la encuesta se asignan los tópicos correspondientes de estas de acuerdo a elección mayoritaria. En caso de ocurrir un empate tanto a nivel de sub como macro tópico se le asignará ambas etiquetas. Posterior a esto simplemente resta comparar la etiqueta asignada por DDR con la obtenida a través de la encuesta y calcular tanto la precisión como el valor-F(*recall*). Información demográfica (detalles en la sección de Anexos) también fue preguntada en la encuesta a modo de tener una visual de las características de los encuestados, cómo rango de edad, alineación política, localización, que tienen relevancia en la forma de percibir el mundo.

Dado que no se necesita normalizar los datos se consideró utilizar los valores que entrega directamente DDR, pero evaluando distintos umbrales para la selección de noticias. Los umbrales de distancia coseno son de 0.5, 0.55, 0.6, 0.65, 0.7 y 0.75. Como se mencionó anteriormente, se extraen de forma aleatoria 100 noticias clasificadas para cada caso de umbral utilizado. En la Tabla 5.1 se observa la cantidad total de noticias clasificadas y el porcentaje de noticias clasificadas con respecto al total de noticias. Además, se incluye una sección que muestra los resultados de una evaluación previa realizada a los modelos. Sin embargo, se debe considerar que estos valores no son cercanos a la realidad debido al sesgo

producto del conocimiento de las sub-divisiones de los tópicos y palabras que conforman los diccionarios de cada uno, información que no se le entrega a los encuestados (se guían solamente a través de la palabra que representa el concepto del diccionario).

Tabla 5.1: Información de los datos con los distintos umbrales utilizados

Umbral	Cantidad de noticias	Porcentaje total	Precisión en evaluación previa
0,50	1.456.418	92,68 %	50 %
0,55	1.270.736	80,87 %	58 %
0,60	894.565	56,93 %	65 %
0,65	437.209	27,82 %	72 %
0,70	145.970	9,29 %	79 %
0,75	33.249	2,16 %	95 %

Con esta información, se decidió utilizar los datos de los casos con los umbrales 0,6 y 0,65, dado que abarcan un porcentaje amplio de los datos (57 % y 28 % respectivamente) y pueden entregar unos valores aceptables (por sobre 65 % de precisión en la clasificación)en la encuesta. Se realizó un total de 2 encuestas, la primera de estas tuvo un total de 30 respuestas y la segunda tuvo 20 respuestas. Los resultados obtenidos de precisión para tópico y macro tópico, y valor-F del macro tópico para ambos casos se pueden observar en la Tabla 5.2. Se puede corroborar que el umbral 0,65 consigue una precisión considerable para el macro tópico, al mismo tiempo que un recall alto. Sin embargo, se puede ver que en ambos casos la precisión de los tópicos para ambos casos es menor del 50 %. Esto se puede explicar con la gran cantidad de tópicos que existen para clasificar al mismo tiempo de la dificultad de identificar los alcances que una palabra puede representar un tópico en las noticias. Se debe tener en consideración que con 50 encuestas realizadas se clasificaron manualmente un total de 750 noticias en 88 posibles tópicos, lo que deja un promedio de 8.5 noticias por tópico (en el caso hipotético que se encuentren en cantidades iguales). Además, se debe tener en consideración que cada persona tiene su propio criterio para asignar alguno de los 88 tópicos y no necesariamente conoce la definición asignada por el IPTC. Por lo tanto, bajo estas circunstancias, conseguir por sobre un 65 % de precisión en la asignación de tópico representa un valor aceptable.

Tabla 5.2: Resultados de la encuesta realizada

Umbral	Precisión Tópico	Precisión Macro Tópico	Recall Macro Tópico
0,60	28,92 %	49,69 %	82,61 %
0,65	40,04 %	68,47 %	80,85 %

5.1.1.1. Ejemplos de clasificaciones realizadas

A partir de los resultados mostrados en la Tabla 5.2, se mostrarán los resultados de los datos filtrados con el umbral de 0.65, que incluye la evolución temporal de las cuentas, distribución general de las noticias, cálculo de RCA, entre otros. En primer lugar se mostrarán los valores que entrega DDR al momento de comparar los diccionarios con la noticia. Además, en la tabla se incluirán solamente los primeros 3 tópicos en orden alfabético junto a los valores de cercanía de cada uno al tópico del tweet. Una de los principales motivos para

mostrar únicamente los primeros 3 tópicos se debe al gran número total de éstos (88 en total). Utilizar los primeros 3 tópicos para mostrar un ejemplo de resultado que entrega DDR permite visualizar el rango de clasificación y relacionar los valores obtenidos con los tweets.

Tabla 5.3: Similitud a los primeros 3 tópicos para la cuenta @latribunacl

Texto	Accidente	Alerta meteorológica	Arte y entretenimiento
Detienen a dos sujetos por robo frustrado: Uno tenía una orden vigente	0.4954	0.3788	0.3226
Los Ángeles: Más de mil personas se inocularon contra el Covid-19 durante este fin de semana	0.3724	0.40485	0.4091
La contrarreloj entregó dos bronces a Biobío en Los Juegos de la Araucanía	0.3152	0.3373	0.2380
En Mulchén: Cierran parque educativo el Cisne tras caída de ramas desde un hualle	0.5835	0.3903	0.4915
Los detalles de seguridad de la casa-búnker allanada en villa Los Profesores	0.3588	0.3098	0.3460
Angelinos participaron en concurso internacional de piano	0.2644	0.2296	0.6077
IPC de octubre aumentó 1,3 % respecto al mes anterior.	0.3705	0.2472	0.3748
Acusación constitucional: Diputado Naranjo presenta discurso de 1300 páginas	0.3978	0.2605	0.5168
El cantante Travis Scott es demandando por tragedia en Astroworld	0.3605	0.2577	0.6043

Como se puede observar que en la Tabla 5.3 los valores efectivamente van en el rango entre 0 y 1, donde el valor cercano a 1 implica una mayor similitud con el concepto del tópico. De esta forma, se puede establecer un criterio para la selección del tópico que mejor representa la noticia y correspondería al tópico con el mayor valor y se realiza el proceso de filtrado considerando aquellas noticias que posean al menos un tópico un un valor mayor al umbral previamente designado de 0,65. Esto asegura que la proximidad del Tweet al concepto del tópico sea relativamente alta. Por ejemplo, la última noticia de la Tabla 5.3 tiene una similitud de 0,6 con el tópico de Arte y Entretenimiento considerando que se trata del cantante Travis Scott, pero en el contexto de la noticia el tópico de Accidente también debió haber presentado algún grado de relevancia.

En la Tabla 5.4 se pueden observar las primeras 9 noticias clasificadas (tanto en el tópico específico como el macro tópico) para la cuenta de @cooperativa. Están por orden cronológico con el titular más reciente primero y considerando solamente aquellos que superaron el umbral

fijado. En esta pequeña muestra de noticias se puede notar que efectivamente existen noticias en las que existe una relación entre el tema del titular con el tópico que DDR le asignó. Por ejemplo, la primera noticia trata de un accidente automovilístico y DDR logra detectar este tópico y asignarlo correctamente. Mismo caso ocurre con la séptima noticia, donde se tratan temas deportivos y DDR consigue una buena representación del tema general de la noticia. Sin embargo, existen casos cómo el de la sexta noticia donde se tratan temas que pueden abarcar múltiples tópicos que en este caso sería un bono (una parte del tema Economía) y por otra parte trata de las Bodas de Oro (un aspecto religioso/social). Existen más casos donde dependerá de la persona que lee el titular definir un tópico en específico y se verá influenciado sobre su conocimiento sobre el tema y el balance o importancia que le asigna a cada tema en particular.

Tabla 5.4: Tópicos asignados para algunos titulares en la cuenta de @cooperativa

ID	Texto	Tópico	Macro
1	Un fallecido tras volcamiento en #Tarapacá: Testigos aseguran que conductor trató de esquivar un bache #CooperativaRegiones #CooperativaContigo	Accidente	Catástrofes
2	Gobierno afirma que existen videos de enfrentamiento armado en Cañete #CooperativaContigo	Conflicto armado	Conflicto
3	"#CooperativaContigo Acusación a Piñera: Oposición mantiene opción de la Ley Lázaro y la derecha pide parar el payaseo	Justicia y Derechos	Justicia
4	#CooperativaRegiones Investigadores estudiarán protección de los humedales de la cuenca del Río Queule #CooperativaContigo	Naturaleza	Medio-ambiente
5	¡Final del partido! Huachipato derrotó 2-0 a Unión La Calera en la Fecha 31 del Campeonato Nacional #CooperativaContigo	Evento deportivo	Deporte
6	Extienden plazo de postulación y aumento en el monto del bono Bodas de Oro #CooperativaContigo	Economía	Economía
7	Universidad de Chile tropezó con Alianza Lima en la Copa Libertadores femenina #CooperativaContigo	Evento deportivo	Deporte
8	[Fotos] #CooperativaRegiones Cientos de personas asistieron al funeral de Yordan Lliempi, comunero fallecido en #Cañete #CooperativaContigo	Celebración	Interés Humano
9	¡Rueda el balón y comienza el partido! Unión La Calera y Huachipato cierran la jornada de domingo en la Fecha 31 del Campeonato Nacional #CooperativaContigo	Evento deportivo	Deporte

Dado que para el caso anterior no se logró notar una noticia clasificada de manera comple-

tamente errónea se mostrarán más ejemplos de otra cuenta. En la Tabla 5.5 se observan los tópicos asignados a los titulares para la cuenta @publimetrochile. Se puede observar en esta Tabla que existen noticias cuyo tópico asignado no corresponde al tema del titular donde un claro ejemplo de esto sería la segunda o séptima noticia. La segunda noticia trata sobre temas relacionados a la Justicia más que a Conflicto, por lo que DDR no habría logrado detectar correctamente el tópico. Lo mismo ocurre para la séptima noticia donde evidentemente se habla tanto de un desastre como son los incendios y temáticas relacionadas al medioambiente, sin embargo, DDR le asigna la categoría de Sector Económico. Entre algunos de los motivos por lo que esto puede ocurrir es la forma en que se definieron los diccionarios y el *word embedding utilizado*. En el capítulo de Análisis entrará en más detalle con respecto a este problema. A pesar de que hay titulares que se clasificaron erróneamente siguen habiendo una parte considerable de estos a los que se les asignó correctamente un tópico. Como se mencionó anteriormente, existen ciertos titulares que pueden variar su tópico dependiendo de la persona que los evalúa, y que interpreta de las palabras que representan un tópico específico y es la razón por la que se tuvo que considerar una cantidad no menor de respuestas para la encuesta.

5.1.2. Distribución general de los tópicos

Posterior de tener estos resultados se puede proceder a mostrar la distribución de los tópicos en el conjunto completo de los datos clasificados. En la Figura 5.1 se muestra la distribución en cantidad de noticias por macro tópico. Como se mencionó en la Tabla 5.1 la cantidad de titulares que conforman este gráfico es de *437.209* únicamente un 28 % de los datos. Se puede extraer del gráfico que entre los tópicos principales que cubren los noticieros se encuentran *Economía*, *ACEM*, *Justicia*, *Sociedad* y *Política*. Pueden existir varios motivos detrás de esta distribución de tópicos, como la facilidad de detectar unos por sobre otros y que estos al mismo tiempo sean capaces de superar el umbral. También se debe tener en consideración que las palabras que conforman los diccionarios también afectan la representación final del concepto, al mismo tiempo que las palabras que conforman un titular pueden hacer variar la representación de este en el word embedding.

5.1.3. Comportamiento de las cuentas

Una vez se obtienen los resultados combinando la información de todas las cuentas, se puede proceder a mostrar el comportamiento de cada cuenta por separado. Dada la gran cantidad de cuentas utilizadas sólo se utilizarán 2 como ejemplificación de los resultados obtenidos. Las 2 cuentas utilizadas corresponderán a las mismas que se utilizaron en las Tablas 5.4 y Tabla 5.5 (@cooperativa y @publimetrochile).

En la Figura 5.2 se observa el porcentaje de cobertura de cada tópico realizada por la cuenta de @cooperativa. Para este caso se puede observar la fuerte presencia que posee el tópico de Deporte al inicio del gráfico coincidiendo con las fechas finales de los Juegos Olímpicos ocurridos en 2021 como se mencionó anteriormente. Por otra parte, también se logra notar la cobertura constante de temas políticos lo que tiene sentido considerando el contexto chileno. Es 2021 el inicio del proceso constituyente, además de ser las elecciones presidenciales para el cambio de mando en 2022. Temas de Economía y Justicia no se quedan atrás, sin embargo, el resto de los tópicos posee una menor representación mas no nula. Esto permite entender un poco más el perfil de la cuenta de @cooperativa y los tópicos que abarca considerando que es una cuenta con una gran cantidad de seguidores (3,1 millones

Tabla 5.5: Tópicos asignados para algunos titulares en la cuenta de @publimetrochile

ID	Texto	Tópico	Macro
1	La muerte de la directora de fotografía Halyna Hutchins podría desencadenar cambios en Hollywood	Arte y Entretenimiento	ACEM
2	El sujeto escapó del arresto domiciliario para ir a una comisaría y pedir ser encarcelado.	Prisionero de guerra	Conflicto
3	Una fotografía que se subió en Twitter se ha transformado en un fenómeno viral.	Medios de comunicación	ACEM
4	El hombre confesó haber asesinado a su esposa, durante un viaje por su primer aniversario de matrimonio	Familia	Sociedad
5	Carabineros fue alertado por vecinos, y al llegar al domicilio encontraron a la mujer fallecida	Familia	Sociedad
6	Esta causa comenzó en 2018 cuando concejales presentaron una querella contra Carter por presunta malversación de fondos, apropiación indebida y al...	Magistratura	Justicia
7	El siniestro afecta a unas 0,4 hectáreas de vegetación natural del sector. Bomberos y brigadista de la Conaf ya trabajan en el control del...	Sector económico	Economía
8	Candidatos de Apruebo Dignidad y el Partido Republicano se llevan casi el 60 % de las preferencias de cara a las próximas elecciones.	Elecciones	Política
9	Tribunal de Arica constató en el juicio los ultrajes cometidos por el sujeto por más de una década a su hija...	Magistratura	Justicia

para diciembre de 2021). Se puede observar también que lo mencionado en la sección de Visualización del capítulo 4 se cumple, pues se tiene un gráfico de tiempo con una leyenda con todos los tópicos, la posibilidad para realizar zoom, volver al estado original y además se puede ver el porcentaje exacto de cobertura de un tópico en una fecha en específica. Toda esta información se puede acceder simplemente ubicando el cursor por sobre la línea de la cual se desea conocer información. Otras 2 cuentas de ejemplo se encuentran en la sección de Anexos, y se escogieron de forma completamente aleatoria. El motivo es la gran cantidad de gráficos obtenidos, por lo que mostrar todos es este informe no es factible.

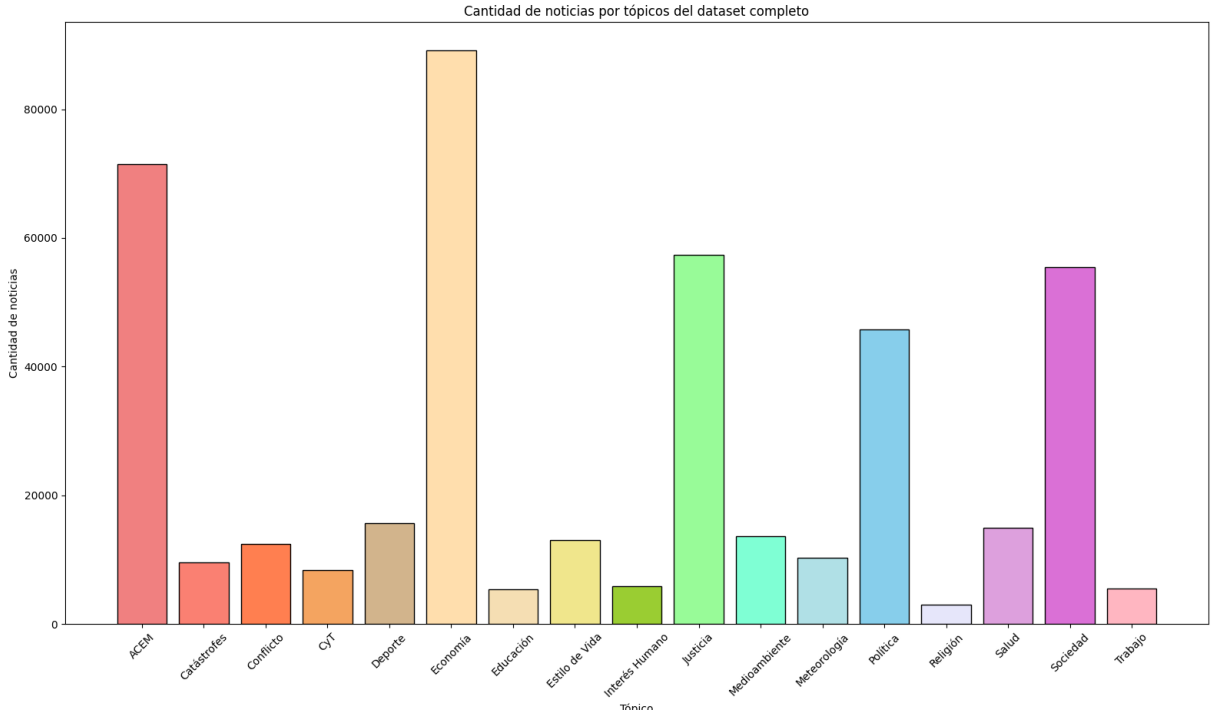

Figura 5.1: Cantidad de titulares por tópico para la totalidad de los datos

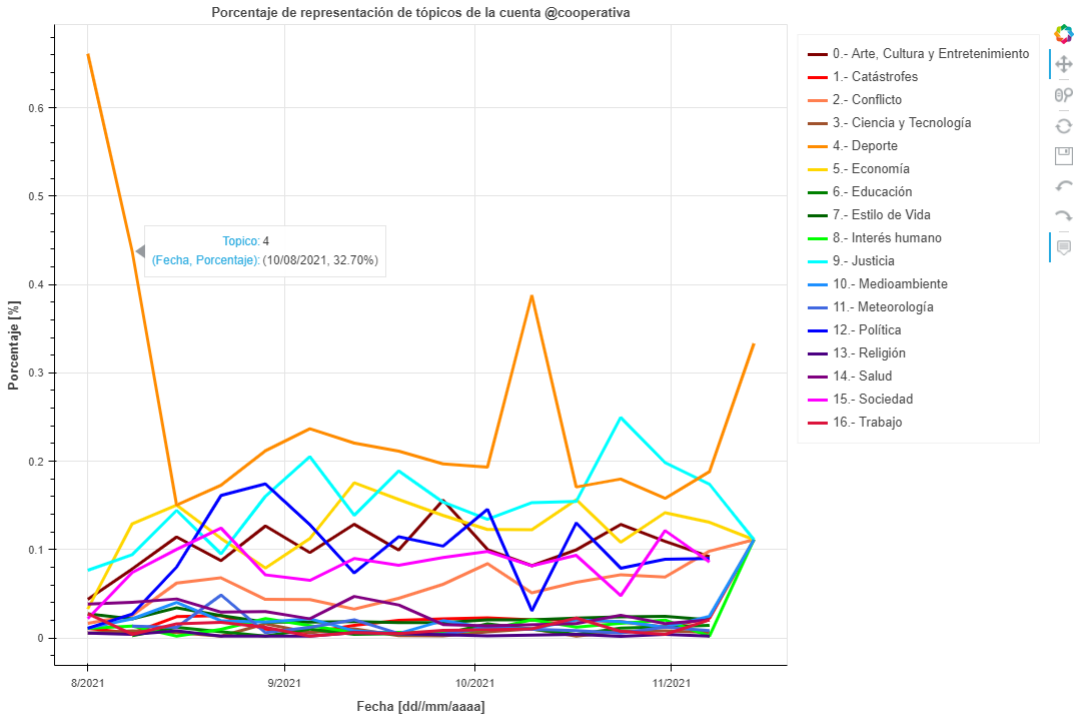

Figura 5.2: Cobertura de tópicos a través del tiempo de la cuenta @cooperativa

5.1.4. Ranking de tópicos y RCA

Ahora se pueda realizar un análisis de ventaja comparativa y el análisis de proximidad entre las cuentas con la información recolectada de las cuentas. Partiendo por RCA, cabe

destacar que al ser 17 tópicos simplemente se mostrarán un par en forma de ejemplo y el resto se encontrará en la sección de Anexos. En la Tabla 5.6 se pueden ver las 10 cuentas con mayor valor de RCA para el tópico de Política. En esta se puede observar que los valores de RCA son bastante altos y que cuentas como @eldemocratacl o @edicionpuntocl se encuentran bastante alto en la lista. Cabe destacar que la cuenta de @gamba_cl se caracteriza por publicar una variedad de noticias (en parte relacionadoas a temas políticos) con un lenguaje más coloquial. Algo importante a destacar es que presentar un valor alto de RCA en este tópico no implica que sea el único de la cuenta, sino que a comparación del resto tienden a publicar más acerca de este tema.

Tabla 5.6: Ranking de las 10 cuentas con mayor valor de RCA para el tópico de Política

Cuenta	RCA
eldemocratacl	20,7015
edicionpuntocl	15,9872
muevomundo	14,0511
radiouniverso	10,9709
diariochile	9,6335
radiomariachile	8,6733
tarapaca_online	8,4283
gamba_cl	8,3148
acciondeongs	7,6051
radioeme	6,5213

En la Tabla 5.7 se observa el ranking con las 10 cuentas para el tópico de Arte, Cultura, Entretenimiento y Medios, donde se pueden observar cuentas como @radiobeethoven, @canalartv, @arteallimite, @radiodisneyla, entre otras, que simplemente a partir del nombre se puede inferir que el enfoque principal de la cuenta está relacionado a este tópico. Inspeccionado en mayor detalle las cuentas resulta que efectivamente varios de los Tweets de estas cuentas se relacionan a dicha temática.

5.1.5. Dendrograma, proximidad y agrupación de cuentas

Ya que se tienen los valores de RCA de cada cuenta, se puede calcular la proximidad entre las cuentas y así agruparlas, formando dendrogramas y *clusters*. El dendrograma con la totalidad de las cuentas resulta difícil de comprender a simple vista dada la gran cantidad de datos que tiene. Debido a esto, se mostrará únicamente una sección aumentada de dicho gráfico. Al igual que con el gráfico de registro temporal de las cuentas, en este se puede realizar zoom y desplazar interactivamente por ser una extensión *HTML*, algo absolutamente necesario con la cantidad de cuentas que se trabajan.

La Figura 5.3 es una sección aumentada del gráfico que se mencionó anteriormente. En esta figura se puede observar como @biobiodeportivo y @deportesarica son agrupados lo que tiene sentido considerando que ambos se encuentran además entre las 10 cuentas con mayor RCA para el tópico deportes como se observa en la Tabla F.4. Bajo ese mismo razonamiento se puede entender que las cuentas de @transmediachile y @wayerless se encuentren bastante cercanas entre sí, ya que ambas se encuentran en la segunda y tercera posición para las 10

Tabla 5.7: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Arte, Cultura, Entretenimiento y Medios

Cuenta	RCA
radiobeethoven	5,7356
canalartv	5,3580
arteallimite	5,3162
contemporaneafm	5,1813
lajugueramag	4,844
canalbangtv	4,7796
etctv_oficial	4,4375
nuevahorizonte	4.4111
ritoquefm	4.3962
radiodisneyla	4.3547

cuentas con mayor RCA para el tópico de Ciencia y Tecnología como se muestra en la Tabla F.2. Este par de ejemplos ayuda a comprender que se pueden formar grupos que tengan sentido utilizando la matriz de proximidad como fuente de información.

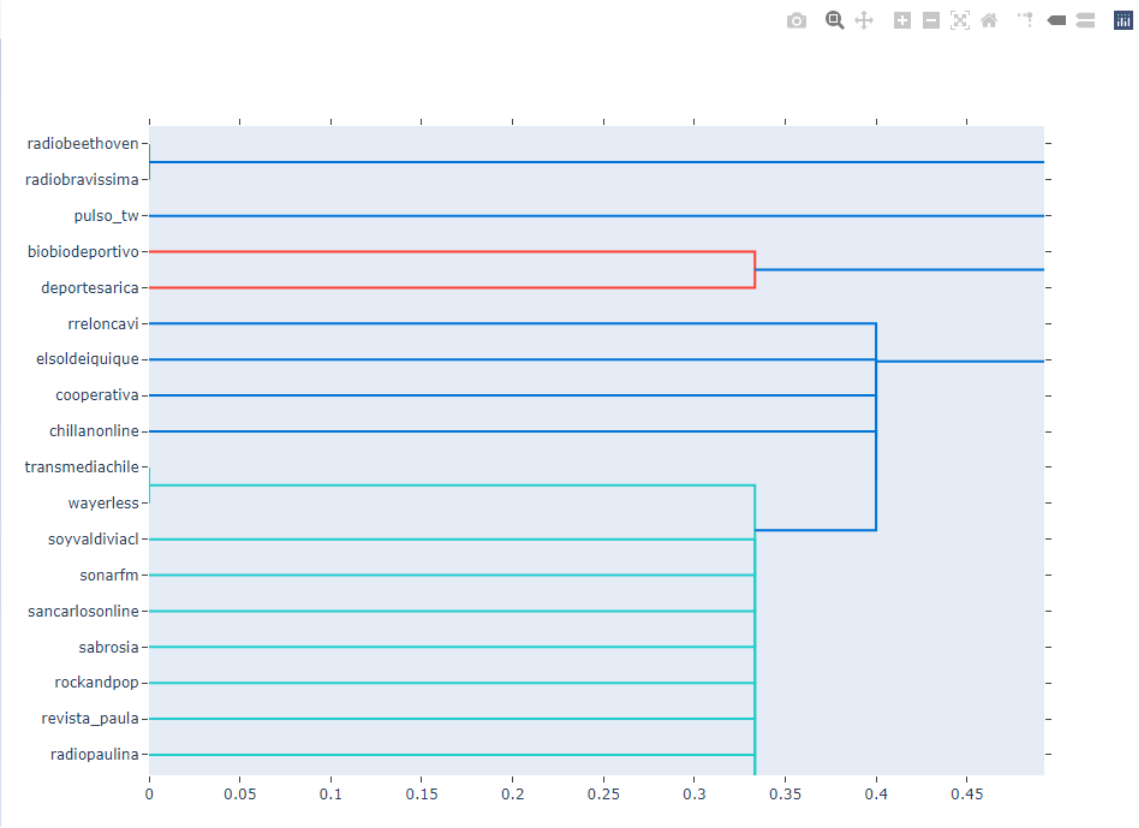

Figura 5.3: Sección del dendrograma de proximidad entre cuentas

También se puede visualizar los distintos grupos formados utilizando la librería *NetworkX* para *Python*. Con esta librería se pueden crear grafos entre distintos objetos con sus enlaces

y atributos respectivos, al mismo tiempo que permite trabajar con datas no estándar. Otra de la cualidades relevantes para este trabajo es su interacción con varias de las librerías de *Python* utilizadas en el trabajo. Antes de utilizar la matriz de proximidad parar crear grafos con *NetworkX*, se debe dejar en valor nulo todo aquel par de cuentas cuya distancia sea mayor a un umbral pre-definido. Ya que se utilizo 0,65 para filtrar los valores que entregaba DDR y así asignar un tópico se utilizó 0,35 como umbral para filtrar las distancias entre cuenta, conservando aquellas que se encontraran efectivamente en un mismo vecindario.

En la Figura 5.4 se observan el resultado de utilizar *NetworkX* en conjunto con *Gephi* la matriz de proximidad filtrada. Los enlaces entre cuentas significan que la distancia entre ellas es de 0.35 o menos y entre más cercanas se encuentran más contenido similar publican (al menos en lo que tópico se refiere). Se observan una gran cantidad de grupos formados diferenciados por color. Este gráfico general resulta complejo de entender debido a la gran cantidad de cuentas que se utilizaron, por lo tanto en los siguientes ejemplos se mostrarán versiones filtradas de este. Algo importante a destacar es la existencia de grupos pequeños, formados por una única cuenta. Estos son cuentas que no poseen mayores similitudes con el resto de las cuentas presentes en el conjunto de datos.

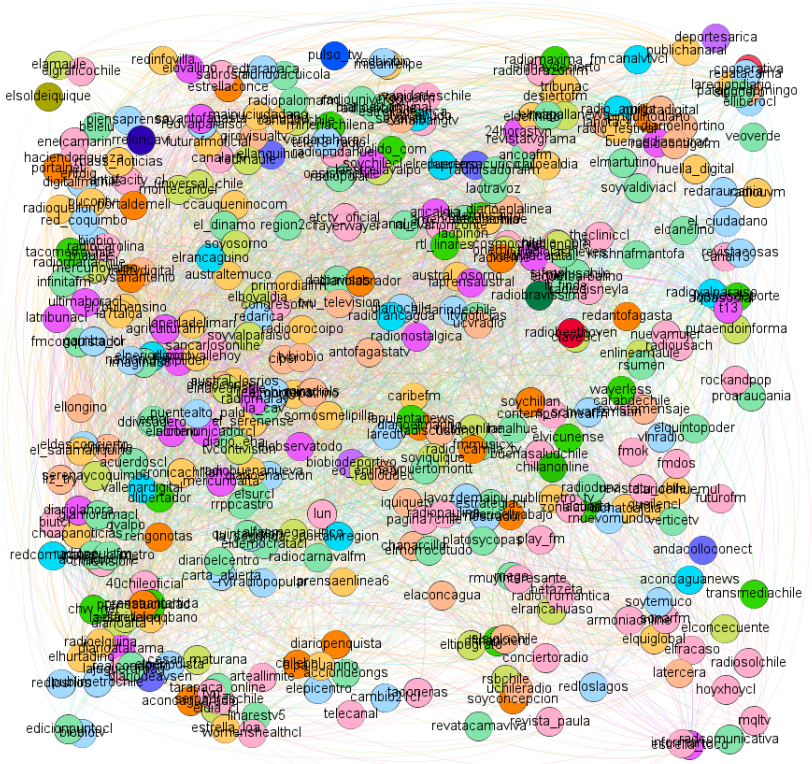

Figura 5.4: Grafo con cuentas similares enlazadas

En la Figura 5.5 Se puede observar el grupo central del grafo, donde se evidencia mejor la similitud entre cuentas. Existen casos interesantes, cómo por ejemplo el grupo formado por las cuentas @ferplei, @elgraficochile y @enelcamarin. En este caso las dos últimas cuentas no

se conectan directamente al resto, por así decirlo, ambas tienen un enlace a @ferplei y esta cuenta se enlaza a @tvn para posteriormente conectarse con 4 cuentas más y así continuar. Del mismo modo ocurre para el caso de las cuentas conectadas a @chw_net que a través de esta y más cuentas pueden generar una conexión con más y más cuentas. Debido a la gran cantidad de cuentas y enlaces resulta complejo visualizar los distintos grupos y cada nombre de cuenta por separado. Con un algoritmo *Fast Greedy* es posible generar comunidades asignando a cada una un color distintivo de esta. Esto facilita la comprensión de los distintos grupos que se pueden generar con las cuentas. Gracias a los algoritmos de agrupamiento es posible generar comunidades dónde sus cuentas son cercanas entre sí. Cabe destacar que para este caso se encontraron 19 comunidades, sin embargo, es posible tanto reducir como aumentar el número de estas dependiendo si los grupos se buscan hacer más generales o más específicos.

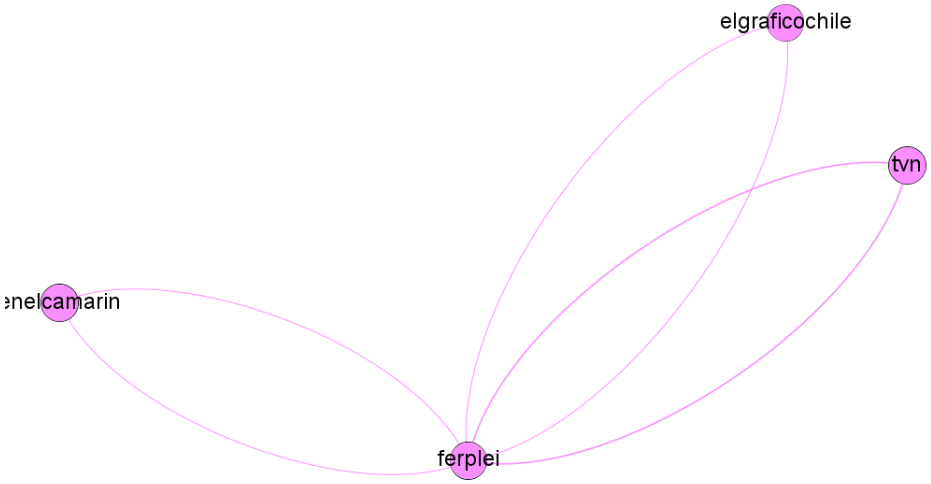

Figura 5.5: Sección del grafo con las cuentas enlazadas a ferplei

Finalmente, en la Figura 5.6 se muestra una de las comunidades obtenidas a partir del grafo, donde se pueden apreciar cuentas cómo @wayerless y @transmedia chile. En esta sección del grafo se puede apreciar de mejor manera que no todas las cuentas que conforman un grupo se encuentran conectadas entre sí. Por ejemplo, @transmediachile no se encuentra conectado directamente a @radiomaxima_fm, pero si a @losriosdeporte que al mismo tiempo se conecta a @radiomaxima_fm. Es por esto que puede darse que dentro de un mismo grupo existan cuentas con una valor de proximidad mayor a 0,35, pero coexisten gracias a las conexiones con otras cuentas cuya proximidad es menor al filtro asignado que también forman parte del

grupo.

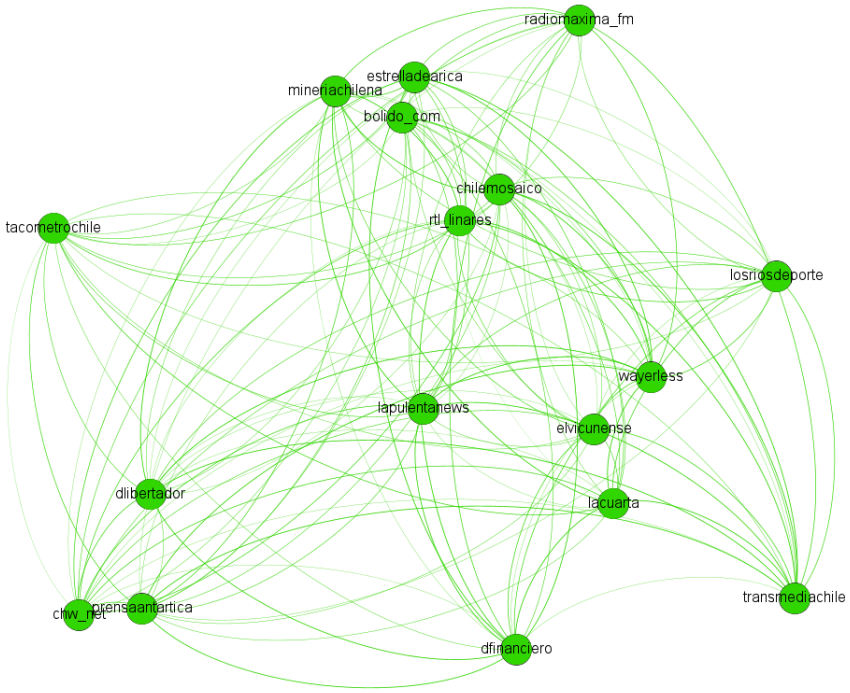

Figura 5.6: Una de las comunidades formadas a partir de las conexiones entre cuentas

Capítulo 6

Análisis de los Resultados

Este capítulo se enfocará en el análisis y discusión de los resultados obtenidos, dificultades encontradas, soluciones propuestas, comportamientos descubiertos, entre otros. El análisis realizado para el método LDA se encuentra en la sección de Anexos D.

6.1. DDR

6.1.1. Encuesta y precisión del modelo

A partir de las encuestas realizadas se puede notar que de la clasificación realizada con DDR se puede obtener una precisión por sobre el 65 % para el tópico general utilizando una cantidad no menor de noticias. Se podría obtener una mejor representación de la precisión del modelo utilizado con una encuesta de mayor escala y con mayor muestra para cada uno de los 88 tópicos. Otra aspecto a qué se puede mejorar de la encuesta es la información que se entrega respecto a cada tópico. Originalmente se deja en manos del encuestado clasificar las noticias con sus conceptos propios acerca de los 88 tópicos utilizados, y lo qué estos abarcan. Sin embargo, la encuesta podría verse beneficiada con una pequeña definición de cada tópico en conjunto con un ejemplo (teniendo en cuenta la definición entregada por el IPTC) en caso de que la persona encuestada no tenga mucho conocimiento respecto a cierto tópico. Con respecto a las clasificaciones obtenidas, en general coinciden con respecto al panorama general de las cuentas. Por ejemplo, en la Tabla 5.4 se pudo observar que varias de las noticias tuvieron una clasificación acorde a uno de los tópicos del titular, pero en la Tabla 5.5 existen titulares que son mal clasificados justificando un 60 % de precisión en la encuesta realizada.

6.1.2. Mejoras al modelo utilizado

Uno de los parámetros que se puede modificar si se buscan obtener mejores resultados en la clasificación es la estructura de los diccionarios. Existe una variedad de palabras que pueden ser clasificadas dentro de la categoría de accidente o catástrofe, cómo de arte y entretenimiento. Hay palabras que pueden ser demasiado específicas para ser útiles en la descripción del tópico (raramente utilizadas en titulares o tweets), o por el contrario, sean demasiado generales para el concepto. Por esto mismo, se deben utilizar aquellas palabras que aportan a la caracterización del concepto y no se sobreponen demasiado con otros tópicos. Dado que para el desarrollo del trabajo no se contaba con el conocimiento acerca de algunos tópicos, la formación de los diccionarios puede recibir mejores y con esto lograr una mejor clasificación de los titulares.

Otro componente que se puede varia es el *word embedding* utilizado. El vector que utilizan las palabras depende netamente del espacio vectorial que se usa, por lo tanto, la distancia existente entre cada palabra depende también del *word embdding*. El espacio vectorial utilizado contiene 1.5 billones de palabras en español de diversas fuentes y al revisar se puede confirmar la presencia de palabras que hacen referencia a distintos aspectos de instituciones o lugares chilenos. Sin embargo, no contiene todo lo respecto a instituciones, sucesos, personajes, lugares o palabras del contexto chileno.Si se llegase a tener un espacio vectorial focalizado en Chile y especialmente noticieros podría ocurrir una mejora a la hora de clasificar noticias, pero al mismo tiempo se perdería un poco la generalidad de éste. Otra forma sería expandir el *word embedding* utilizando los propios titulares de las noticias utilizadas en el trabajo y re-entrenar el espacio vectorial.

Varios de los errores se producen por no considerar el contexto en el que están las palabras, ya que a cada una se le otorga un valor a través del *word embedding* de manera independiente para finalmente promediar todas y obtener el concepto resultante. Esto termina en casos como clasificar con el macro tópico de Medio Ambiente a un titular que menciona *AFP Hábitat*, dado que hábitat forma parte del diccionario de naturaleza y no toma en consideración que en este caso se refiere al nombre de una compañía chilena en vez de un lugar natural. Otros ejemplos que salen de la Tabla 5.5 serían los titulares 4 y 5, donde en ambos se habla sobre un asesinato, sin embargo, debido a la mención de matrimonio, mujer, vecinos, entre otros, son clasificados como Sociedad dado que estas palabras forman los diccionarios de este concepto. Esto también evidencia que todas las palabras poseen el mismo peso a la hora de realizar los cálculos. Por lo tanto, debido a la múltiple mención de palabras de un tópico en específico mueve el concepto del texto hacia este a pesar que el tema central sea otro (algo que puede resultar evidente para una persona). Si existiese alguna forma de incluir el contexto en el que se encuentran las palabras, se podría conseguir una mejor clasificación de las noticias.

6.1.3. Clasificación de tópicos

La distribución de tópicos del conjunto completo de titulares utilizados fue un resultado inesperado, pero del cual se puede extraer información. Como se menciono en el capítulo anterior, las noticias abarcan desde el final de los Juegos Olímpicos, las elecciones primarias, eventos de fútbol, *El Juego del Calamar*, entre otros, en contexto de una pandemia debido al COVID-19. Debido a esto se espera una mayor representación del tópico Deporte y Salud. Solamente se logró capturar las última semana de los Juegos Olímpicos lo que podría explicar que ha pesar de todo no logró una mejor representación. Por su parte, algunas de las noticias relacionadas a la pandemia hacen referencia al mismo tiempo a otro tópico como podría ser entretenimiento (apertura de cines por ejemplo) lo que baja la clasificación de estas noticias. Que el tópico Economía y finanzas tenga tanta representación en las noticias es una sorpresa, pero se debe principalmente al tópico *Sector Económico*. Una diversidad de palabras conforman el diccionario de dicho tópico, y cuando ningún otro tópico se encuentra medianamente presente este tópico consigue superar el umbral establecido, a veces incluso superando al tópico real del titular. Para solucionar esto se debe reducir las palabras que conforman el diccionario y evitar superposiciones con otros conceptos o dividir este en partes más específicas. Además, tendría sentido con lo mencionado anteriormente respecto a la falta de contexto para la representación de las palabras en el espacio vectorial. A pesar de todo se puede ver que el algoritmo no tiene mayores problemas reconociendo titulares referentes al área del entretenimiento, lo que permite una mejor representación de este tópico.

6.1.3.1. Comportamiento de las cuentas

Con los ejemplos mostrados sobre el porcentaje de cobertura de tópico para 3 cuentas distintas se logra comprender que las cuentas pueden tener una temática dedicada o ajustarse a los eventos que ocurren a nivel mundo/país/local. En la Figura 5.2 se puede ver claramente la influencia que tuvieron los Juegos Olímpicos 2021 o las eliminatorias de la Copa Mundial para *Cooperativa* con respecto al tópico de Deportes. Sin embargo, las elecciones primarias hicieron aumentar el porcentaje de noticias relacionadas a la Política por el mes de Julio en la cuenta de @cooperativa. Esto implica que efectivamente los sucesos que acontecen en tiempo real alteran la cantidad de cobertura que se le otorga a un tópico específico. El hecho que las elecciones primarias hayan conseguido un 20 % de cobertura cuando todo lo relacionado a deporte venía dominando es una prueba de esto. Y ocurre nuevamente en Octubre con las clasificatorias los partidos de la Selección Chilena de Fútbol al aumentar la cobertura del tópico Deporte a un 40 %.

Otra característica interesante que se logró confirmar es la existencia de cuentas dedicadas únicamente a uno o dos tópicos, con aumento de algún otro tópico por acontecimientos puntuales. Un claro ejemplo de esto viene siendo la cuenta de *Rock&Pop 94.1*. Esto se ve reflejado en la cantidad de noticias que sacan relacionadas al Arte y Entretenimiento. Por lo tanto, se tienen cuentas que publican una variedad de tópicos y otras cuentas que se dedican principalmente a la cobertura de un tópico. Tiene sentido comparar cuentas con comportamiento similar, por ejemplo, comparar la cuenta de *Rock&Pop 94.1* con otras cuentas presentes en el top ranking del tópico Arte y entretenimiento. El porcentaje de cobertura de @radiobeethoven se puede observar en la Figura 6.1, dónde claramente se aprecia su foco en el tópico de Arte y Entretenimiento. Sin embargo, su principal diferencia con la cuenta de @rockandpop (Figura E.2) es el resto de tópico que publican. La cuenta de @rockandpop sigue publicando de temas relacionados a sociedad, política, trabajo e incluso deporte, a diferencia de @radiobeethoven que casi únicamente publica sobre arte o entretenimiento. Esto también indica que cuentas presentes en el ranking de un tópico efectivamente centran principalmente su cuenta en torno a este, por lo que RCA es una buena herramienta para la comparación entre cuentas en el contexto de este trabajo.

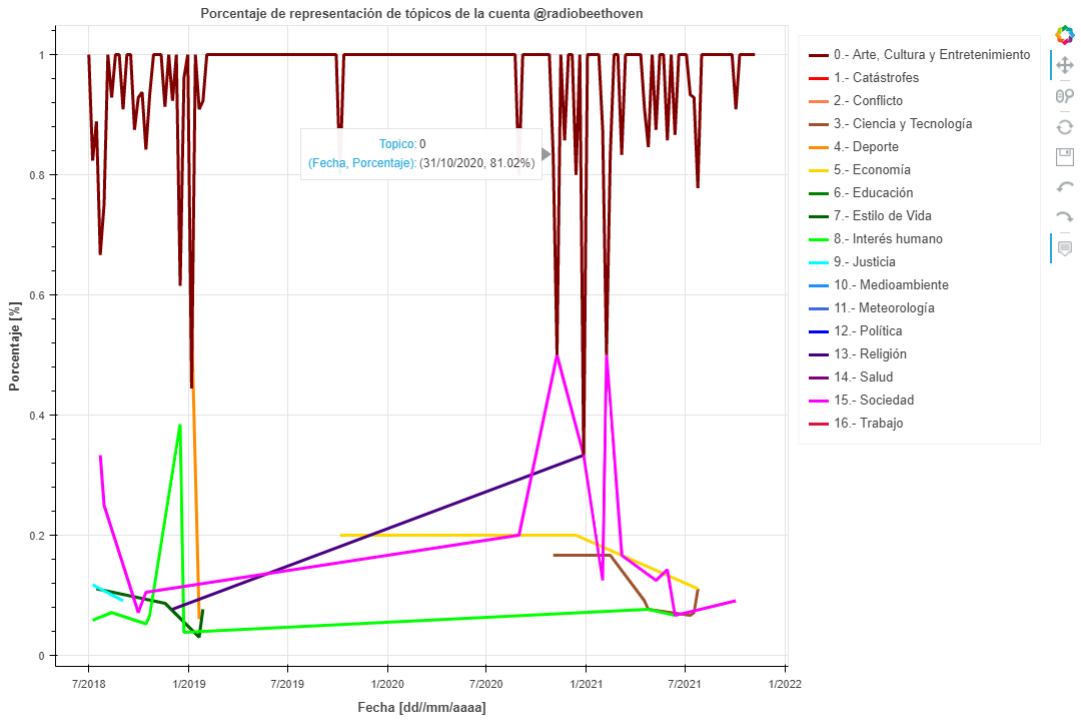

Figura 6.1: Cobertura de tópicos a través del tiempo de la cuenta @radiobeethoven

6.1.3.2. Cantidad de tweets por cuenta

Algo ha tener en consideración es el rango de tiempo de los titulares de la cuenta @radiobeethoven, que va desde el 2018 hasta mediados de Octubre de 2021. Esto se debe a que esta cuenta tiende a publicar una menor cantidad de noticias con respecto al resto. Como se mencionó anteriormente, la extracción de datos se realiza hasta la publicación más antigua que se pueda extraer con la API teniendo en consideración el límite de aproximadamente 3000 *Tweets* (por utilizar una cuenta no premium). Entre menos publicaciones diarias se realizan publicaciones más antiguas pueden ser extraídas. Es decir, existe una diferencia en la cantidad de publicaciones realizadas entre cuentas, donde algunas publican una gran cantidad de tweets por día en comparación a otras que lo hacen en menor medida. Sin embargo, este factor no altera el cálculo del RCA para la asignación de tópico, dado que, RCA cálcula su valor utilizando los porcentajes de tweets realizados por la cuenta con respecto al tópico del total de tweets de dicho tópico.

6.1.3.3. Matriz de proximidad y agrupamiento de cuentas

Para el cálculo de las distancias que conforman la matriz de proximidad no se utilizan los valores de RCA sino que un valor binario con 1 representando un valor de RCA mayor a 1 y 0 en caso contrario. Esto implica que no importa qué tan iguales sean los valores de RCA para un tópico entre 2 cuentas, solamente importa la cantidad de tópicos iguales que publican. Esto se comprueba con las cuentas @transmediachile y @wayerless. Ambas se encuentran en un mismo grupo y forman parte de las cuentas con mayor ventaja respecto a Ciencia y Tecnología. Al observar la evolución temporal en las Figuras 6.2 y 6.3, para las cuentas de @transmediachile y @wayerless respectivamente, se da cuenta de un patrón de publicación

similar. En estas dos cuentas se aprecia la fuerte presencia del tópico Ciencia y Tecnología. Sin embargo, también se puede apreciar que los tópicos de Economía, seguido por los tópicos de Arte y Entretenimiento y Sociedad tienen un grado de presencia no menor. Esto implica que la medida de proximidad es efectivamente una forma viable de generar grupos con cuentas similares entre sí, dado que toma en consideración principalmente la cantidad de tópicos en común que se publican, en lugar de utilizar únicamente el que más publican. Finalmente, esto permite visualizar los grupos formados en un espacio vectorial como se observa en la Figura 5.6. Se puede corroborar que la formación de grupos depende únicamente de los tópicos que se publican frecuentemente, ya que se pueden ver 2 grupos distintos cuentas que se encuentran en el ranking para el Tópico de Deporte. Así mismo se pueden ver que existen cuentas que resultan complejas de agrupar con otras, probablemente debido a la mezcla de tópicos que publican. El grafo obtenido se puede utilizar para recomendar a usuarios visitar los perfiles de cuentas que se encuentren más cercanos a los que frecuenta.

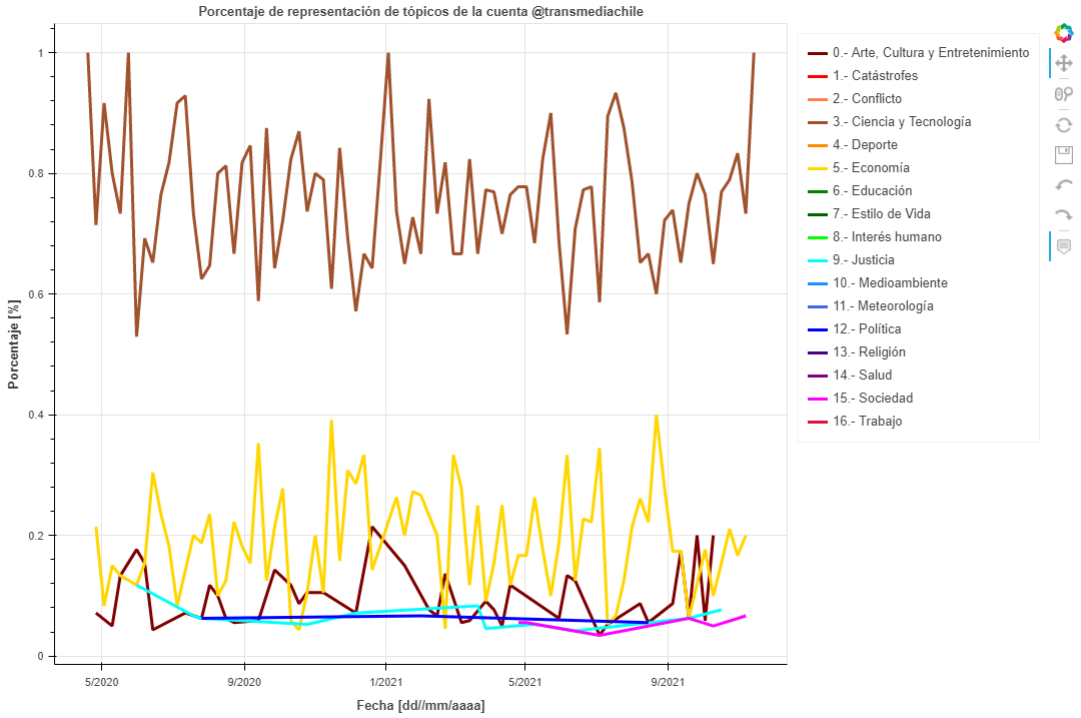

Figura 6.2: Cobertura de tópicos a través del tiempo de la cuenta @transmediachile

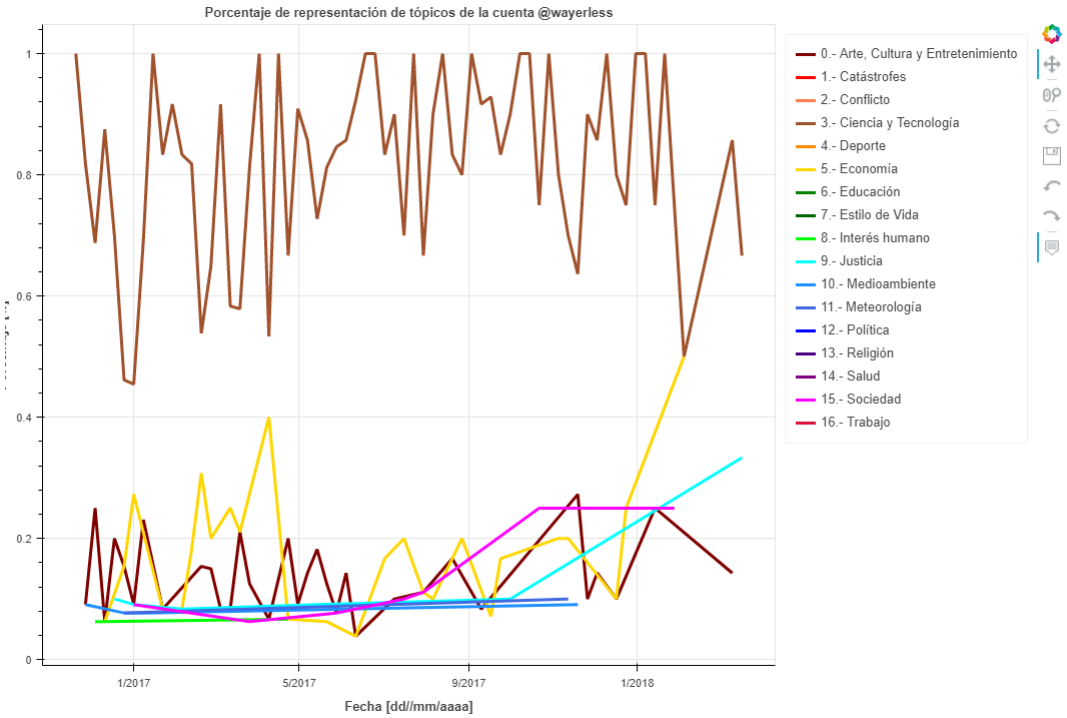

Figura 6.3: Cobertura de tópicos a través del tiempo de la cuenta @wayerless

Un último aspecto importante a considerar es que las conexiones entre cuentas no se encuentran delimitadas a la comunidad que pertenecen. Una cuenta puede estar conectada a cuentas de otras comunidades siempre y cuando se encuentre a una distancia menor de 0,35 de cercanía. Un claro ejemplo de esto se puede observar en la Figura 6.4, donde se muestran todas las cuentas vecinas de @glamoramacl con profundidad 2 (se incluyen vecinos de los vecinos). En este caso se pueden observar 3 cuentas de comunidades distintas de @glamoramacl. Por lo tanto, es posible asumir que ser el vecino del vecino de una cuenta no asegura tu pertenencia a un mismo grupo, y también se considera la similitud general de la comunidad a la hora de ser formadas. También ayuda a generar una gran red de cuentas y es posible llegar a cuentas de otras temáticas a partir de una cuenta específica.

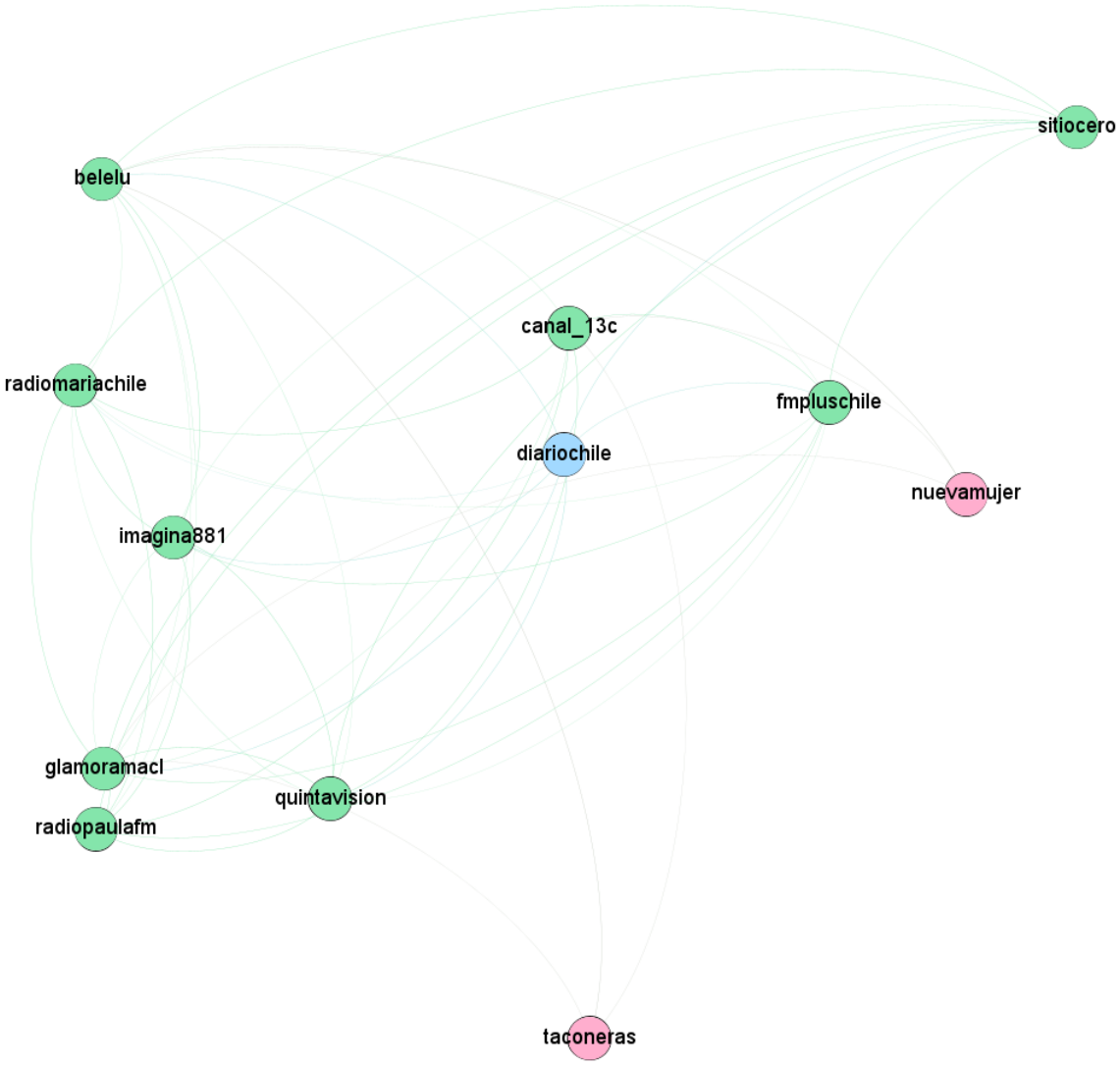

Figura 6.4: Cuentas cercanas a @glamoramacl

Capítulo 7

Conclusión

En el mundo las formas de acceder a información son variadas, y los medios de comunicación de noticias se han adaptado a estas nuevas redes para conectarse con más usuarios. Ya sea *Facebook*, *Instagram* o *Twitter*, los noticieros han conseguido establecerse para poder entregar información y mantenerse conectado a los usuarios. En el contexto chileno, los hechos ocurridos en las manifestaciones ocurridas a finales de 2019 han potenciado el uso de estas redes para la obtención de información y y potenciar noticieros alternativos. Debido a esto, resulta interesante tener una forma de visualizar como distintas cuentas de noticieros (tradicionales y alternativos) cubren los distintos tópicos, y si existen diferencias entre cuentas respecto a la atención que le otorgan a cada tema. Es por esto que se decidió realizar un seguimiento a 375 de noticieros chilenos en la plataforma *Twitter* y observar su comportamiento a través del tiempo, utilizando *Tweepy* y una cuenta de desarrollador para acceder a la información necesaria de estas. De este modo se logra cumplir el objetivo de creación de una base de datos conteniendo los tweets necesarios acerca de una variedad de cuentas de noticias tanto tradicionales como alternativos.

7.1. Metodología

En este trabajo se abarcó la clasificación de titulares en diversos tópicos y el uso de estos para el análisis de cobertura de cuentas noticieras chilenas. No se consideró utilizar técnicas de aprendizaje de máquina supervisadas debido a la gran cantidad de datos que deberían ser etiquetados. Para poder realizar la clasificación de datos se utilizaron métodos inherentemente no supervisados, los que son *Latent Dirichlet Allocation* (LDA) y *Distributed Dictionary Representation* (DDR). Se logró observar que LDA permite conocer los tópicos de los titulares utilizando estos mismos como datos. Sin embargo, se complica a la hora de querer realizar una análisis temporal de los datos. Se optó por utilizar DDR ya que permite pre definir las etiquetas con las cuales se planea clasificar a través de diccionarios y facilita su visualización. Para definir los tópicos se utilizó la taxonomía entregada por el *IPTC*, que esta diseñada para identificar los tópicos de noticias a través de definiciones de diferentes conceptos. Esta es una taxonomía multinivel, por lo tanto, se puede descender por las ramas de un tópico más general (si es que tiene) para encontrar el tema específico de la noticia. Se decidió utilizar el nivel más general de tópicos de la taxonomía para la visualización del porcentaje de cobertura y cálculos de proximidad (y lo que se evalúa finalmente), y el segundo nivel más específico para la clasificación de noticias. Esto permite cumplir los objetivos específicos sobre probar los distintos métodos para la clasificación de noticias y el

establecimiento de etiquetas a utilizar para la clasificación.

7.2. Validación del modelo

Con las encuestas realizadas se pudo comprobar que el método utilizado para asignar los tópicos (DDR) entregó una precisión por sobre el 60 % y se encuentra disponibles a mejoras. Las mejoras se pueden realizar a través de cambios en los diccionarios utilizados, mejor definición de tópicos o cambio del *word embedding* utilizado y una mayor cantidad de noticias clasificadas en el proceso de validación del modelo. La primera de estas correspondería a encontrar el número óptimo de tópicos que se deberían utilizar para clasificar los titulares (en los tópicos no macro tópicos) y cuán específicos deben ser. La segunda forma correspondería a encontrar las palabras adecuadas que forman estos tópicos/diccionarios para una mejor formación del concepto a través de DDR. Una de las últimas formas de conseguir un mejor rendimiento de DDR sería cambiar el *word embedding* utilizado para la representación de las palabras. Al revisar algunos ejemplos de clasificación se puede observar que existen casos donde los titulares son clasificados correctamente, y otros donde se desvía completamente del tópico o lo asocia a un tema secundario dentro de este. Se pudo concluir que este problema deriva principalmente en la falta de mejores diccionarios, un mejor *word embedding* y el hecho que un titular (o tweet) puede no contener las palabras necesarias para realizar una correcta clasificación (ya sea en importancia para el tópico o cantidad).

7.3. Cobertura de tópicos

En cuanto a la cobertura de tópicos a través del tiempo, se pudo ver que efectivamente se logra entregar una forma de visualizar el porcentaje de representación de los tópicos en distintas cuentas. Se logró inferir que existen cuentas que se enfocan casi exclusivamente a ciertos tópicos específicos, o cuentas que se dedican a la publicación de noticias de diversa índole. Estas últimas también pueden verse influenciadas por los eventos de gran importancia que ocurren en el día a día. Esto también ayuda a tener un registro del momento en que ciertos temas comienzan a ser más hablados por los noticieros y relacionarlos con los eventos que ocurrieron en esa ventana de tiempo, como por ejemplo los Juegos Olímpicos o las eliminatorias para la Copa Mundial de Fútbol. Con esto se logra cumplir el objetivo de tener una forma de visualización interactiva para la cobertura de tópicos de las distintas cuentas.

7.4. Ranking y proximidad

En lo que respecta a los valores obtenidos por el cálculo de RCA, se puede corroborar la diferencia entre cuentas. En las tablas respectivas se muestran cuentas que aportan con la mayor cantidad de noticias respecto a un tópico en comparación al resto de publicaciones de estas cuentas. Esto permite entender rápidamente las preferencias que tienen a la hora de publicar noticias y al mismo tiempo poder formar un perfil de la cuenta respecto a los tópicos que publica. También se logró entender que tener un RCA alto en un tópico no impide tener al mismo tiempo un RCA alto en otro, teniendo cuentas que se enfocan en múltiples tópicos. Con esta información se logra establecer una medida de proximidad entre cuentas considerando la cantidad de tópicos en común que publican. Esta medida de proximidad se usa para poder crear grupos de cuentas que se encuentren cercanas y conectadas entre sí. Esto

se puede visualizar de mejor manera en un dendrograma y grafos mostrados en el capítulo de Resultados, donde se forman distintos grupos ya sea de cuentas enfocadas a Deporte, Ciencia y Tecnología o se crean grupos separados debido a su falta de similitud con el resto. Con esto se puede concluir efectivamente la medida de proximidad utilizada es útil a la hora de querer generar agrupar cuentas y comparar que la similitud entre estas. De este modo se tiene una medida para comparar las cuentas que más aportan a un tópico específico, y una medida para comparar cuentas entre sí y ver qué tan similares son (que puede ser utilizado para un sistema de recomendación). Con todo lo anterior, se logró cumplir el objetivo principal del Trabajo de Título, así cómo los objetivos secundarios. Se extraje aprendizaje a partir de los errores y dificultades encontradas, y se tendrá una mejor respuesta en caso de encontrarse con problemáticas similares a la hora de ejercer cómo profesional.

7.5. Trabajo a futuro

Finalmente, una de las mejoras que se puede realizar a partir de este proyecto es la integración de todas las partes en una plataforma web para implementar los resultados en tiempo real. Esto requeriría extraer los publicaciones de las cuentas en intervalos de tiempo fijo, buscando obtener los titulares más nuevos e agregarlos para su clasificación y visualización en el gráfico temporal. Con esto también variaría el cálculo de RCA de algunas de las cuentas (principalmente aquellas que se dedican a publicar de una variedad de tópicos) y provocaría el movimiento de estas dentro del grafo de proximidad (principalmente en eventos de alta cobertura). Realiza esta integración no es una tarea sencilla, pero es algo que se puede lograr y es una buena forma de llevar el trabajo realizado a un mayor nivel. El diseño de esta plataforma depende netamente del uso que ser quiera dar. Por ejemplo, puede ser una página que muestre únicamente la información extraída, con los rankings, grupos, y gráficos obtenidos. Se le puede añadir un sistema de recomendación sencillo, donde únicamente se requiera el nombre de la cuenta de un noticiero y recomiende aquellas cuentas más cercanas. También podría crearse un sistema de recomendación para el usuario, que debería tomar todas las cuentas de noticieros que sigue y encontrar una cuenta que en promedio se encuentre cercana a todas las cuentas que este usuario sigue. Existen muchas formas de implementar la información obtenida a través de este trabajo y es algo que se buscará implementar a futuro.

Bibliografía

[1] Candia C, Uzzi B. Quantifying the selective forgetting and integration of ideas in science and technology. Am Psychol. 2021 Sep;76(6):1067-1087. doi: 10.1037/amp0000863. PMID: 34914440.

[2] Connerton, P. (2008). Seven types of forgetting. Memory Studies, 1(1), 59–71. doi:10.1177/1750698007083889

[3] Real Academia Española.(s.f.). Medio de comunicación. En *Diccionario de la lengua española*. Recuperado en 25 de Octubre de 2021, de https://dle.rae.es/medio#BgOCDE6

[4] Reuters Institue, University of Oxford. (2020) Reuters Institute Digital News Report 2020, p. 92.

[5] Luna J. P., Toro Maureira S. & Valenzuela S. (2021) El ruidoso silencio de los medios tradicionales. Recuperado en 25 de Octubre de 2021, de https://www.ciperchile.cl/2021/03/23/el-ruidoso-silencio-de-los-medios-tradicionales/

[6] Durán, P., Lawrence, T., Fernández, J. E. (2020) La red social Twitter y el proceso constituyente: el caso de las cuentas anómalas. Recuperado en 25 de Octubre de 2021, de https://www.ciperchile.cl/2020/10/17/la-red-social-twitter-y-el-proceso-constituyente-el-caso-de-las-cuentas-anomalas/_ftn4

[7] Candia, C., Jara-Figueroa, C., Rodriguez-Sickert, C. et al. The universal decay of collective memory and attention. Nat Hum Behav 3, 82–91 (2019). Recuperado en 27 de Octubre de 2021, en https://doi.org/10.1038/s41562-018-0474-5

[8] Sáez-Trumper, D., Castillo, C., & Lalmas, M. (2013). Social media news communities: gatekeeping, coverage, and statement bias. Proceedings of the 22nd ACM international conference on Information & Knowledge Management.

[9] De Clercq, O., De Bruyne, L., & Hoste, V. (2020). News topic classification as a first step towards diverse news recommendation. COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS JOURNAL, 10, 37–55.

[10] International Press Telecommunications Council (1965), IPTC is the Global Standards Body of the News Media. Recuperado en 26 de Octubre de 2021, de https://iptc.org/about-iptc/

[11] International Press Telecommunications Council (1965), IPTC Current Members. Recuperado en 26 de Octubre de 2021, de https://iptc.org/participate/membership/current-members/

[12] International Press Telecommunications Council (1965), Media Topics - IPTCs. Recuperado en 26 de Octubre de 2021, de https://iptc.org/standards/media-topics/

[13] Vrandecic, D. & Krötzsch, M. (2014), Wikidata: a free collaborative knowledgebase.

Commun. (ACM). Volume. 57, Pages. 78-85. https://doi.org/10.1145/2629489

[14] International Press Telecommunications Council (1965), NewsCodes Scheme (Controlled Vocabulary). Recuperado en 27 de Octubre de 2021 https://cv.iptc.org/newscodes/mediatopic/?lang=es

[15] Rudnik, Charlotte & Ehrhart, Thibault & Ferret, Olivier & Teyssou, Denis & Troncy, Raphaël & Tannier, Xavier. (2019). Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata.

[16] Tripathy, R.M., Sharma, S.S., Joshi, S., Mehta, S., & Bagchi, A. (2014). Theme Based Clustering of Tweets. CODS.

[17] Tweepy, Tweepy An easy-to-use Python library for accessing the Twitter API. Recuperado en 27 de Ocutbre de 2021, de https://docs.tweepy.org/en/stable/getting_started.html

[18] Tweepy, Tweepy: Twitter for Python!. Recuperado en 28 de Octubre de 2021 , de https://github.com/tweepy/tweepy

[19] Elejalde, E., Ferres, L. & Schifanella, R. Understanding news outlets' audience-targeting patterns. EPJ Data Sci. 8, 16 (2019).

[20] Twitter, Developer Plataform - Get Tweet Timelines. Recuperado en 28 de Octubre de 2021, de https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-home_timeline

[21] Sarica, S. Luo, J.(2020) STOPWORDS IN TECHNICAL LANGUAGE PROCESSING

[22] NLTK. NLTK - Python Module Index. Recuperado en 28 de Octubre de 2021, de https://www.nltk.org/py-modindex.html

[23] Garten, J., Hoover, J., Johnson, K.M. et al. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behav Res 50, 344–361 (2018). https://doi.org/10.3758/s13428-017-0875-9

[24] Marwala, T. (2014) Artificial Intelligence Techniques for Rational Decision Making. Adv. Informat. Knowledge Processing Springer, Cham. p. 8-12.

[25] Doherty, P., Thiebaux, S. Artificial Intelligence. An International Journal. Recuperado en 08 de Noviembre de 2021, de https://www.journals.elsevier.com/artificial-intelligence

[26] R. Saravanan and P. Sujatha. A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification, 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 945-949, doi: 10.1109/ICCONS.2018.8663155.

[27] P. Ongsulee, Pariwat, "Artificial intelligence, machine learning and deep learning," ICT and Knowledge Engineering (ICT&KE), 2017 15th International Conference on. IEEE, 2017

[28] Nasteski, Vladimir. (2017). An overview of the supervised machine learning methods. HORIZONS.B. 4. 51-62. 10.20544/HORIZONS.B.04.1.17.P05.

[29] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, 2001.

[30] N. Friedman, D. Geiger, and M. Goldszmidt. "Bayesian network classifiers," Machine

learning, vol. 29, pp. 131-163, 1997.

[31] Lee, C, Woo, S., Linear classifier design in the weight space, Pattern Recognition, Volume 88, 2019, Pages 210-222, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2018.11.024.

[32] Caruana. R., Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (ICML '06). Association for Computing Machinery, New York, NY, USA, 161–168. DOI:https://doi.org/10.1145/1143844.1143865

[33] Shah, K., Patel, H., Sanghvi, D. et al. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. Augment Hum Res 5, 12 (2020). https://doi.org/10.1007/s41133-020-00032-0

[34] Tang, Duyu & Wei, Furu & Yang, Nan & Zhou, Ming & Liu, Ting & Qin, Bing. (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. 1. 1555-1565. 10.3115/v1/P14-1146.

[35] Ghannay, Sahar & Favre, Benoit & Estève, Yannick & Camelin, Nathalie. (2016). Word Embeddings Evaluation and Combination. Language Resources and Evaluation.

[36] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[37] Dy, J. G., & Brodley, C. E. (2004). Feature selection for unsupervised learning. Journal of machine learning research, 5(Aug), 845-889.

[38] Khanam, Memoona & Mahboob, Tahira & Imtiaz, Warda & Ghafoor, Humaraia & Sehar, Rabeea. (2015). A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance. International Journal of Computer Applications. 119. 34-39. 10.5120/21131-4058.

[39] Hofmann, T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning 42, 177–196 (2001). https://doi.org/10.1023/A:1007617005950

[40] Zhang, Q., Yang, Y., Liu, Y., Wu, Y. N., & Zhu, S. C. (2018). Unsupervised learning of neural networks to explain neural networks. arXiv preprint arXiv:1805.07468.

[41] Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. International Journal of Advanced Research in Artificial Intelligence, 2(2), 34-38.

[42] Ko, Y., & Seo, J. (2000). Automatic text categorization by unsupervised learning. In COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics.

[43] David, J. (2016). Sentiment and topic classification of messages on Twitter : and using the results to interact with Twitter users.

[44] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet allocation: extracting topics from software engineering data," art Sci. Anal. Softw. data, pp. 139–159, 2015

[45] Putri, I. R., & Kusumaningrum, R. (2017). Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia. In Journal of Physics: Conference Series (Vol. 801, No. 1, p. 012073). IOP Publishing.

[46] C. Văduva, I. Gavăt and M. Datcu, "Latent Dirichlet Allocation for Spatial Analysis of Satellite Images,ïn IEEE Transactions on Geoscience and Remote Sensing, vol. 51, no. 5, pp. 2770-2786, May 2013, doi: 10.1109/TGRS.2012.2219314.

[47] Sanandres, E. & Llanos, R. & Madariaga, C. (2018). Topic Modeling of Twitter Conversations.

[48] Q. Chen, L. Yao and J. Yang, "Short text classification based on LDA topic model,"2016 International Conference on Audio, Language and Image Processing (ICALIP), 2016, pp. 749-753, doi: 10.1109/ICALIP.2016.7846525.

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

[50] Lee, Junseok & Kang, Ji-Ho & Jun, Sunghae & Lim, Hyunwoong & Jang, Dongsik & Park, Sangsung. (2018). Ensemble Modeling for Sustainable Technology Transfer. Sustainability. 10. 2278. 10.3390/su10072278.

[51] Yang, S., & Zhang, H. (2018). Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. International Journal of Computer and Information Engineering, 12(7), 525-529.

[52] Guo, Lei & Vargo, Chris & Pan, Zixuan & Ding, Weicong & Ishwar, Prakash. (2016). Big Social Data Analytics in Journalism and Mass Communication: Comparing Dictionary-Based Text Analysis and Unsupervised Topic Modeling. Journalism & Mass Communication Quarterly. 93. 10.1177/1077699016639231.

[53] Cardellino, C.(2016) Spanish Billion Words Corpus and Embeddings, https://crscardellino.github.io/SBWCE/

[54] Cañete, J. & Chaperon, G. & Fuentes, R. & Ho, J. & Kang, H. & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. PML4DC at ICLR 2020.

[55] Ortiz, O.(2020). Who do we spend time with across our lifetime?. Our World in Data. Recuperado en 08 de Abril de 2022, de https://ourworldindata.org/time-with-others-lifetime

[56] Bokeh Development Team (2018). Bokeh: Python library for interactive visualization. urlhttp://www.bokeh.pydata.org.

[57] Balassa, B. (1965), Trade Liberalisation and Revealed Comparative Advantage, The Manchester School, 33, 99-123.

[58] Hausmann, Ricardo & Hidalgo, Cesar, 2014. "The Atlas of Economic Complexity: Mapping Paths to Prosperity"MIT Press Books, The MIT Press, edition 1, volume 1, number 0262525429.

[59] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

[60] Molina, Carlos & Gibon, François-Marie & Julio, Pinto. (2008). Estructura de macroinvertebrados acuáticos en un río altoandino de la Cordillera Real, Bolivia: variación anual y longitudinal en relación a factores ambientales. Ecología Aplicada. 7. 10.21704/rea.v7i1-2.365.

[61] Lee, Kathy & Palsetia, Diana & Narayanan, Ramanathan & Patwary, Md. Mostofa Ali & Agrawal, Ankit & Choudhary, Alok. (2011). Twitter Trending Topic Classification. 251-258. 10.1109/ICDMW.2011.171.

Anexos

Anexo A. Tópicos no utilizados/fusionados

Tabla A.1: Lista de los tópicos del IPTC que no se utilizaron o fueron fusionados con otros para su uso.

Tópicos no utilizados	Aprendizaje social, Aniversario Asociaciones de padres de alumnos Calificación educativa, Celebridades, Ceremonia Ciencia biomédica, Cumpleaños, Desgracia humana Entrenamiento deportivo Estudiante Examen de admisión, Líder religioso Malla curricular, Organización no gubernamental Organización de salud, Plantas, Profesores Propiedad y administración deportiva Relaciones entre las instituciones religiosas y el Estado Transferencia deportiva
Tópicos fusionados	Crisis política y Disidencia política Emigración e Inmigración Gente y Celebración, Infraestructura deportiva Institución deportiva y Organización deportiva Institución científica, Investigación científica y Normas Récord y Premio

Anexo B. Diccionarios utilizados

B.1. Arte, Cultura, Entretenimiento y Medios

Tabla B.1: Diccionarios utilizados para el tópico Arte, Cultura, Entretenimiento y Medios

Diccionario	Palabras
Arte y entretenimiento	arte dibujo escultura foto pintura caricatura anime animación teleserie serie video stream pop musica cine película concierto festival drama baile actor cantante celebridad kpop
Cultura	tradición cultura biblioteca museo libro festividad lenguaje idioma Monumento_Nacional Patrimonio_Cultural exposición literatura cultural memorial monumento
Medios comunicación	internet diario revista periódico radio redes stream televisión boletín blog reportaje publicación noticia desinformación audiovisual comunicación documental propaganda

B.2. Catástrofes y accidentes

Tabla B.2: Diccionarios utilizados para el tópico Catástrofes y accidentes

Diccionario	Palabras
Accidente	accidente derrumbe explosión colisión ahogar volcar atropello colapso desplome descarrilamiento colapso hundimiento
Catástrofe	desastre deslizamiento erupción inundación sismo terremoto incendio avalancha huracán tornado tormenta maremoto tsunami catástrofe
Socorro/Emergencia	urgencia socorro evacuación emergencia auxilio rescate evacuar

B.3. Ciencia y tecnología

Tabla B.3: Diccionarios utilizados para el tópico Ciencia y tecnología

Diccionario	Palabras
Ciencia natural	astronomía estrella galaxia biología fisiología cosmología física energía electromagnetismo partículas geología oceanografía química átomos materia bioquímica
Ciencia social	ensayo antropología arqueología sociología psicología lingüística historia geografía filosofía derecho sintaxis fonética Ciencia_Social etnología metafísica
Institución/Investigación científica	institución laboratorio experimento innovación Investigación_Científica tesis académico ensayo Innovación_Productiva academia
Matemática	matemática álgebra geometría trigonometría lógico-deductivo ecuación integral derivada promedio estadística media aritmética topología
Tecnología	tecnología maquinaria estructura industria fabricación construcción irrigación app computadora Inteligencia_Artificial robótica software hardware biometría radiofrecuencia consola Steam

B.4. Conflicto, guerra y paz

Tabla B.4: Diccionarios utilizados para el tópico Conflicto, guerra y paz

Diccionario	Palabras
Atentado	atentado bomba bioterrorismo explosivo incendiario tiroteo terrorista explosión asalto asesinato balacera
Conflicto armado	guerra conflicto armado guerrilla sabotaje secuestro intervención ocupación civil tropas Guerra_Civil
Disturbios	disturbio manifestación desorden rebelión revolución protesta manifestantes revuelta alboroto
Golpe de estado	derrocamiento Golpe_de_Estado dictador Dictadura_Militar dictatorial Grupo_Militar destitución Estado_De_Excepción opresión Orden_Público motín tiranía sublevación
Matanza	masacre matanza genocidio exterminio aniquilación purga inmolación lesa atrocidad
Proceso de paz	pacificación paz mediador amnistía desarme indulto excarcelación absolución desmovilización desmilitarización
Reconstrucción	reconstrucción restauración desarme desminar rehabilitación remodelación
Prisión de guerra	preso rehén prisionero guerra

B.5. Deporte

Tabla B.5: Diccionarios utilizados para el tópico Deporte

Diccionario	Palabras
Deportes de competición	ajedrez alpinismo escalada marciales atletismo fútbol balonmano boxeo ciclismo gimnasia patinaje paracaidismo voleibol básquetbol deporte
Disciplina deportiva	sanción disciplinaria arbitro réferi mediador
Dopaje	rdopaje esteroides doping antidopaje anabólicos
Evento deportivo	Campeonato_Mundial cammpeonato torneo Campeonato_Internacional copa Juegos_Olímpicos partido campeonato
Organización deportiva	cancha estadio Estadio_Municipal hipódromo polideportivo Estadio_de_Fútbol organización Asociación_Deportiva instalación
Logro deportivo	medalla medallas récord campeón bicampeón campeona Récord_Mundial supercopa

B.6. Economía, negocios y finanzas

Tabla B.6: Diccionarios utilizados para el tópico Economía, negocios y finanzas

Diccionario	Palabras
Economía	economía aranceles banco Banco_Central bono comercio consumidor balanza litigio crédito deflación deuda empresa cooperativa empresariado empleo exportación fondos hipoteca inversión mercado precio presupuesto recesión tasa criptomoneda productividad
Información de empresas	marketing compra contrato empresa sociedades adquisiciones fusiones multinacionales mercado patente subcontratación dividendo beneficios contabilidad auditoría financiamiento previsión quiebra bolsa inversión despido
Mercado de bolsa	compra venta metales Metales_Preciosos préstamo deuda activos ahorrantes stock commodities recambios postventa mercado finanza capitales divisas cambiario
Sector económico	agricultura Sector_Económico acuicultura pesqueros ganado plantación vinicultura artesanía alimento lujo bebida electrónicos juguetes vestuario computación Bienes_Raíces inmobiliaria diésel gasolina gas combustible servicio transporte metro transantiago turismo Centro_Comercial tiendas almacenes Mercado_Municipal

B.7. Educación

Tabla B.7: Diccionarios utilizados para el tópico Educación

Diccionario	Palabras
Crianza	crianza criar tutor pupilo educación enseñanza prebásica preescolar primaria
Educación primaria	Educación_Infantil Educación_Primaria Enseñanza_General_Básica colegio liceo estudiante alumno preescolar profesor secundaria instituto politécnico Colegio_Público Escuela_Pública Colegio_Particular Jardín_Infantil clases educación Simce
Educación secundaria	universidad academia instituto estudiante egresado graduado facultad Universidad_Privada Universidad_Estatal admisión acreditación evaluación clases Prueba_de_Selección_Universitaria PSU Prueba_de_Aptitud_Académica educación especialización

B.8. Estilo de vida y tiempo libre

Tabla B.8: Diccionarios utilizados para el tópico Estilo de vida y tiempo libre

Diccionario	Palabras
Fitness	calistenia rutina gimnasio trotar aeróbicos entrenar fitness running ejercitación crossfit pilates musculación
Estilo de vida	alimentación vegetariano vegano vegan jardín patio planta plantas hogar decoración piercing tatuaje dieta ornamentación pircing
Ocio	recreativo recreacional buceo caza club juego Aire_Libre lotería bar café parque restaurante pasatiempos colección vacaciones viaje ocio recreo gameplay zoológico cafetería hobbies entretenimientos

B.9. Interés humano, animales, insólito

Tabla B.9: Diccionarios utilizados para el tópico Interés humano, animales, insólito

Diccionario	Palabras
Animales	mascota gato perro reptil ave animal hámster cuy loro tortuga hurón pájaro pez especie roedor hervíboros
Celebración	aniversario cumpleaños boda ceremonia graduación funeral memorial conmemoración Día_de_la_Madre San_Valentín Navidad
Premio	premio récord logro distinción galardón Óscar ganador

B.10. Mano de obra

Tabla B.10: Diccionarios utilizados para el tópico Mano de obra

Diccionario	Palabras
Desempleo	despido paro desempleo cesantía desocupación desempleado pensionista
Empleo	aprendiz autoempleo capacitación profesional empleado empleador oficio profesión salario trabajo empleo oficinista autoocupación
Jubilación	jubilación pensión Previsión_Social cotización fondo subvención compensación jubilar pensiones
Legislación laboral	regularización Legislación_Laboral Secretaría_de_el_Trabajo Secretaría_de_Trabajo Ministerio_de_Trabajo Seguridad_Laboral Seguro_Médico Salud_de_los_Trabajadores
Mercado de trabajo	trabajo salario puesto vacante Entrevista_De_Trabajo Mercado_Laboral laboral salariado Salario_Mínimo
Relación laboral	contrato huelga paro protesta Relación_Laboral pre-contrato manifestación Control_Empresarial
Sindicato	sindicato colectivo trabajador gremio Unión_de_Trabajadores sindical

B.11. Medio ambiente

Tabla B.11: Diccionarios utilizados para el tópico Medio ambiente

Diccionario	Palabras
Cambio climático	Calentamiento_Global Efecto_Invernadero Cambio_Climático temperatura emisión desforestación sequía deshielo emisiones desertificación derretimiento
Contaminación ambiental	contaminación smog residuo polución vertidos Contaminación_Ambiental Contaminación_Industrial desecho corrosivo Sustancias_Tóxicas derrames
Naturaleza	ecosistema especie nativo ambiente naturaleza biodiversidad hábitat medioambiente biotopo humedales cordillera lago
Política ambiental	Política_Ambiental ecosostenible sustentabilidad ecológica sustentable ecoeficiente hipocarbónicas
Preservación	preservación extinción flora fauna biodiversidad depredación endémicas especies Monumento_Natural Paisaje_Protegido Espacio_Natural_Protegido
Recursos naturales	agua río océano pantano cuenca Energía_Renovable fotovoltáica eólica geotérmica energético bosque montaña Recursos_Naturales Desarrollo_Sustentable Recursos_Hídricos Ministerio_de_el_Ambiente biomasa

B.12. Meteorología

Tabla B.12: Diccionarios utilizados para el tópico Meteorología

Diccionario	Palabras
Alerta meteorológica	prealerta alertas meteorológico tormentas vientos
Estadística meteorológica	humedad Presión_Atmosférica temperatura Metros_Cúbicos Velocidad_De_el_Viento
Fenómeno meteorológico	polar sequía tornado tormenta tifón nieve niebla viento ciclón Tormenta_Eléctrica neblina
Pronóstico	llover grados celcius temperatura chubasco soleado nevar nublado precipitación calor frío pronóstico clima

B.13. Policía y justicia

Tabla B.13: Diccionarios utilizados para el tópico Policía y justicia

Diccionario	Palabras
Autoridades	arresto investigación desaparición policía carabineros vigilancia detención Policía_de_Investigaciones allanamiento autoridad detención gendarmería uniformados Gendarmería_de_Chile Dirección_de_Inteligencia_Nacional Inspectoría_General
Criminalidad	agresión delincuencia corrupción soborno extorsión crímen narcotráfico contrabando delito infracción malversación tráfico fraude violación evasión homicidio asesinato maltrato robo secuestro piratería a_mano_armada altercado vandalismo
Justicia/Derechos	derecho norma regla Derecho_Penal Código_Civil Código_de_Procedimiento_Penal libertad Derecho_Civil Código_Penal cláusula ley
Magistratura	corte Corte_Suprema Corte_Suprema_de_Justicia Tribunal_Supremo juez juicio acusado litigio cárcel multa suspensión judicial testigo abogado fiscal magistratura víctima

B.14. Política

Tabla B.14: Diccionarios utilizados para el tópico Política

Diccionario	Palabras
Derechos fundamentales	censura derecho Derecho_Fundamental Vida_Privada Derecho_Civil libertad Libertad_De_Expresión integridad honra Derecho_Constitucional honor
Crisis/Disidencia política	Crisis_Política oposición opositor disturbio alboroto manifestación protesta huelga disidencia
Elecciones	candidato debate elecciones revocación referendo Sistema_Electoral votación presidenciales parlamentarias votaciones votos electoral campaña
Gobierno	comisión comité parlamento congreso Partido_Político constitución Asamblea_Constituyente Consulta_Pública ministro ministerio Fuerzas_Armadas presidente Presidente_de_la_República diputado senador destitución gobierno constitucional constituyente
Legislación	regulación corte tribunal extradición legislación ley normativa norma Ley_Nº Ley_Orgánica
Política	cabildo Sistema_Político política Historia_Política dirigente militante sindicalista democracia dictadura Partido_Político lobby constitución
Política de gobierno	impuestos mediador reglamento nacionalización privatización Empresa_Estatal Seguridad_Ciudadana Participación_Ciudadana Readaptación_Social tributos estatización Políticas_Internas Medidas_Agrarias Publicaciones_De_Información_General Pensiones Subsidios
Relaciones internacionales	tratado Ayuda_Internacional disputa frontera fronteriza Deuda_Externa Asilo_Político refugiado diplomado diplomacia Tratado_Internacional Relaciones_Internacionales Ciencias_Políticas Estudios_Internacionales

B.15. Religión y culto

Tabla B.15: Diccionarios utilizados para el tópico Religión y culto

Diccionario	Palabras
Conflicto religioso	Intolerancia_Religiosa islamofobia cristianofobia cruzadas antisemitismo
Creencias	taoísmo sikhismo parsismo masonería masón judaísmo judío islam hinduismo secta cristiano cristianismo budismo animismo
Evento religioso	ritual navidad pentecostés ramadán Yom_Kippur bautismo rito Hanukkah Jánuca Día_de_la_Cruz Pésaj Shavuot Pascua_Judía
Lugar de culto	iglesia templo mezquita capilla sinagoga parroquial catedral basílica santuario
Texto religioso	biblia corán tora Tanakh El_Corán tafsir

B.16. Salud

Tabla B.16: Diccionarios utilizados para el tópico Salud

Diccionario	Palabras
Enfermedad	sida epidemia pandemia peste virus coronavirus cáncer Enfermedad_Mental lesión obesidad intoxicación enfermedad contagio síntoma infección
Enfermedad no humana	parásito plaga pulga garrapata zoonótico Trichinella Cryptosporidium Borrelia Cyclospora
Establecimiento de salud	hospital clínica Centro_De_Salud Centro_Médico Hospital_Regional psiquiátrico internado Hospital_General Ambulancia Emergencia
Política de salud	Salud_Pública Seguro_De_Salud salud política Servicio_Médico
Profesión médica	médico cirugía farmacología geriatría medicina obstetricia ginecología odontología oftalmología oncología ortopedia pediatría psiquiatría radiología Personal_Médico enfermero
Tratamiento	tratamiento dieta suplemento medicamento medicina vacuna cirugía terapia prescripción vacunado pacientes vacunación fármaco antibiótico

B.17. Sociedad

Tabla B.17: Diccionarios utilizados para el tópico Sociedad

Diccionario	Palabras
Asistencia social	beneficencia Cuidado_Infantil Vivienda_Social vulnerabilidad Ayuda_Social Asistencia_Social Bienestar_Social Protección_Social
Comunidades	club comunidad pueblo comunidades sociedad población etnias aldeas pueblos indígenas
Condición social	pobreza pobre Condición_Social refugio Sueldo_Mínimo indigencia
Demografía	población censo habitantes demografía demográfica Censo_Nacional pobladores habitantes ciudadanos
Discriminación	discriminación racismo sexismo homofobia xenofobia misoginia Integración_De_Los_Derechos_Humanos_De_La_Mujer_Y-_La_Perspectiva_De_Género racial estigmatización intolerancia discriminatorias transfobia
Familia	adopción madre padre casa matrimonio noviazgo pololo hijo familia aborto contracepción anticoncepción hogar
Humanidad	adolescente niña adulto bebé discapacitado homosexual Orientación_Sexual Identidad_de_Género minoría étnia indígena LGBT
Inmigración	emigrante extranjero inmigrante emigración inmigración Inmigrante_Ilegal migrante indocumentado
Problema social	acoso acoso_sexual adicción Delincuencia_Juvenil Abuso_Sexual esclavitud prostitución funa manipulación Embarazo_Adolescente servidumbre pedofilia
Valores	antivalores muerte pornografía ética eutanasia suicidio moralidad decencia honestidad aborto

Anexo C. Latent Dirichlet Allocation (LDA)

Se partió utilizando LDA para todos los Tweets de las cuentas combinadas, como para los Tweets de las cuentas por separado. En ambos casos se especificó que la cantidad de clusters que el programa debía formar era de 17, haciendo relación a la estructura diseñada por el IPTC. Esto no implica que sea la forma más efectiva de utilizar LDA, dado que existen formas de optimizar el modelo como también encontrar el número óptimo de clusters. Sin embargo, sirve como una primera aproximación para observar el comportamiento de los datos, dado que no se planeaba utilizar esta técnica para resolver el problema final. Los resultados obtenidos para los tópicos con la cantidad total de los datos se pueden ver en la Tabla C.1.

También se obtuvieron los clusters de las cuentas por separado y así ver como varían dependiendo del enfoque principal que tiene el medio. En este sentido, la cantidad de Tweets influye en la representación para los tópicos por lo que cuentas con una mayor cantidad de publicaciones se pueden asemejar a los tópicos obtenidos en la Tabla C.1. Los resultados obtenidos para dos cuentas aleatorias utilizadas se pueden ver en las Tablas C.2 y C.3.

Tabla C.1: Palabras de los 17 tópicos encontrados utilizando la cantidad total de Tweets del dataset

Tópico	Palabras frecuentes
0	ohiggins, son, primer, parque, trabajo, estamos, traves, mujeres, esta, aire
1	aca, mas, muy, cuenta, hacer, le, hay, tener, porque, alcalde
2	mejores, regional, vallenar, bomberos, hospital, gracias, 12, mar, 2021, noticias
3	via, carabineros, entrada, ruta, recuerda, iquique, santa, mujer, hombre, jornada
4	años, desde, hasta, primera, asi, mas, sabado, dias, domingo, siguiente
5	zona, sector, vivir, casa, santiago, centro, valparaiso, comuna, familia, esta
6	casos, covid19, nuevos, region, salud, 1, 2, 3, fallecidos, 4
7	san, viernes, todo, estos, alto, rancagua, 22, valdivia, puerto, busca
8	nuevo, capitulo, esta, e, todas, cuarentena, vida, conversacion, junto, musica
9	proyecto, aqui, lunes, tus, mas, nuestra, uno, mundo, edicion, toda
10	mejor, gobierno, quieres, maule, cerca, director, cuando, miercoles, sera, julio
11	nueva, junto, nuestro, programa, constitucion, proyectos, disfruta, temuco, fernando
12	solo, mira, presidente, tambien, historia, puedes, compra, piñera, 11, visita
13	descubre, dos, lugar, entre, chile, mercado, concepcion, esta, tres, partido
14	semana, fin, paso, cambio, coronavirus, plan, comunas, vacunacion, norte, proxima
15	gran, pandemia, total, puede, 2230, durante, fueron, mes, final, sistema
16	mas, barrio, curico, caso, personas, mil, han, estas, araucania, millones

Se puede observar en la Tabla C.1 que existen tópicos a los que se les puede encontrar algún tipo de coherencia, como por ejemplo, el número 14 que habla principalmente del coronavirus, el paso a paso y la vacunación donde todo se puede resumir en el tópico de COVID-19. Sin embargo, en casos para el tópico 1 puede ser confuso encontrar la temática principal del clusters. Esto se debe a que las palabras que lo conforman suelen ser utilizadas en una diversidad de contextos y no necesariamente en un caso particular. Otro ejemplo de caso en el que se puede encontrar una coherencia más o menos general a lo largo del cluster sería el número 5, que utiliza palabras como *vivir*, *sector*, *casa* y *comuna*, además de ciudades, para referirse a vivienda y ubicaciones. Una de las mayores complicaciones para formar clusters que hagan sentido vendría siendo la gran cantidad de palabras que no aportan mayor información a la temática de la noticia como se mencionó anteriormente.

Se puede observar en la Tabla C.2 que los datos evidentemente corresponden a la cuenta de *@elmostrador*, dado que en los tópicos 0, 1 y 2 es la palabra más influyente. En el mismo tópico 0 se puede observar que también se tocan temas políticos debido a la presencia de las palabras política, gobierno y director. En este mismo sentido los tópicos 1 y 2 hablan de temas económicos y de salud respectivamente. Para esta cuenta se puede observar una mayor cohesión en lo que respecta a los clusters formados, debido a que corresponden a los datos de una sola cuenta en lugar de una combinación de varias. Sin embargo, siguen existiendo casos como el cluster 16 en el que es complejo encontrar una temática en particular, además de la existencia de palabras dentro de un clusters que son complejo de relacionar el tópico general sin el contexto apropiado como en el caso de Marta para el tópico 0.

Para el caso de la cuenta de *@vallenardigital*, se puede observar que efectivamente trata

Tabla C.2: Palabras de los 17 tópicos encontrados para la cuenta @elmostrador

Tópico	Palabras frecuentes
0	elmostradorenlaclave, política, lagos, marta, gobierno, director, mori, puede
1	elmostradorenlaclave, político, analista, retiro, va, cuarto, 10, economista
2	elmostradorenlaclave, pandemia, salud, dosis, personas, vacunación, vacunas
3	jorge, carahue, carabineros, crisis, dice, ataque, querella, ética, muerte
4	generación, m, últimas, mundo, premio, 2021, interés, primer, daza, retorno
5	nueva, convención, constituyente, constitución, constitucional, ahora, mesa
6	ministro, reglamento, ahora, mercados, paris, impacto, gobierno, advierte
7	pueblo, lista, +, candidato, destacado, josé, presidencial, sociales, mapuche
8	paso, elmostradorenlaclave, plan, votación, chile, economista, participación, 24
9	ddhh, región, derechos, políticas, sur, desarrollo, dentro, humanos, estallido
10	nuevo, ramón, san, clave, mostrador, dos, cultura, panoramas, historia
11	presidente, braga, proyecto, piñera, ahora, mujeres, ley, votos, diputados
12	agenda, opinión, país, chile, cómo, estudio, china, tercera, niños, braga
13	cultura, presos, estallido, cambio, chilena, investigación, festival, acusa, chileno
14	elecciones, 2021, boric, jadue, sichel, provoste, ahora, presidencial, yasna
15	covid19, mundo, casos, ahora, nuevos, minsal, nivel, fallecidos, positividad
16	cuba, video, editar, chile, años, país, mundo, menos, casi, variante

primariamente de tópicos referentes a Vallenar o al menos incluye la palabra Vallenar en varios de sus Tweets. Ignorando la presencia de esta palabra, se puede ver que existente tópicos donde la mayoría de las palabras apuntan a un tema específico como es el caso del cluster 9 donde se tienen palabras relacionadas a política. O por ejemplo en el caso del cluster 6 que trata de temas policiales, accidentes y catástrofes.

Tabla C.3: Palabras de los 17 tópicos encontrados para la cuenta @vallenar-digital

Texto	Palabras frecuentes
0	vallenar, freirina, huasco, aguas, chañar, ecológico, atacama, sociales, través
1	nueva, entrada, publicada, vallenardigital, vallenar, chollay, atacama, pdi
2	alto, carmen, vallenar, aluvión, afectados, disfrutaron, ayuda, tránsito, casa
3	vallenar, mujer, mujeres, éxito, versión, día, carabineros, deportes
4	vallenar, educación, freirina, programa, realizará, infantil, red, sexual
5	niños, freirina, millones, entrega, huasco, atacama, pesos, niñas, vallenar
6	copiapó, fiscalía, amarilla, tierra, sujeto, policial, ocurrido, incendio, muerte
7	pdi, detiene, vallenar, personas, dos, copiapó, tres, hallazgo, chile, investiga
8	atacama, nan, región, regional, gobierno, postular, damnificados, recursos
9	provoste, vallenar, yasna, diputada, ruta, 5, norte, ley, vecinos, atacama
10	censo, social, agua, huasco, participar, 2017, vallenar, potable, proceso, rural
11	salud, bomberos, atacama, 8, enero, lamenta, realizados, 20, realizará
12	vallenar, liceo, alcalde, cristian, tapia, gobierno, llamado, plaza, inclusión
13	alerta, comunas, preventiva, diego, temprana, atacama, almagro, declara
14	atacama, intensidad, regiones, semana, paseo, menor, familias, justicia, mes
15	huasco, provincia, vallenar, hospital, provincia, chañaral, trabajo, nuevo, año
16	vallenar, canta, festival, seguridad, verano, sábado, comuna, concurso, noche

Anexo D. Análisis de resultados - LDA

En esta sección se realizará un análisis de los resultados obtenidos, características encontradas en las cuentas, comportamiento de los algoritmos utilizados, entre otros. Iniciando con los resultados entregados por LDA, se puede observar que existe una diferencia al utilizar todos los titulares como un solo conjunto de datos, a separar los titulares por sus cuentas respectivas. Una mayor cantidad de datos implica una mayor cantidad de tópicos, también ayuda a entender los temas más relevantes que hay en el conjunto completo de los datos, y al reducir a 17 grupos se tienen palabras de tópicos que se hablan frecuentemente. Por otro lado, al utilizar LDA separado por cuenta se puede representar de mejor manera los tópicos de las cuentas, incluyendo palabras que hacen referencia a esta misma como lo sería *elmostradorenclave* para el caso de la cuenta @elmostrador como se ve en la Tabla C.2. Se vería complicado el seguimiento de tópicos similares en distintas cuentas debido a las diferencias existentes entre cuentas al momento de abarcar estos. O por otro lado, una misma cuenta habla de temas similares pero en contextos distintos como podrían ser las elecciones. Por lo tanto, se debería encontrar una forma de asignar un nombre para cada tópico tanto nuevo como antiguo, un color respectivo y formas de diferenciarlos de nuevos tópicos que surjan con palabras similares que conforman éstos (lo que llevaría a una gran cantidad de tópicos). También se complicaría el análisis de proximidad realizado a falta de tópicos definidos y estandarizados para cada tópico emergente. Al igual que en el caso del conjunto completo, LDA o BERT (otro algoritmo para la detección de tópicos) se pueden encontrar la totalidad de tópicos que existen en los titulares. Se logró notar que reducir a solamente 17 tópicos

implica que los grupos formados puedan tener palabras sin mucha relación al resto. A pesar de todo, utilizar LDA o BERT o cualquier algoritmo que extraiga los tópicos directamente de los datos podría conseguir una representación más o menos acertada de estos (obviando los problemas que traen consigo las técnicas de aprendizaje no supervisado), y son útiles para entender antes de iniciar un trabajo cuál es el panorama general de los datos a utilizar.

Por esto mismo, logra comprobarse la particularidad de los métodos no supervisados y específicamente LDA que se explicó en el capítulo *Marco teórico*. La función principal de LDA es encontrar tópicos en los datos y clasificar cada dato utilizando los tópicos encontrados. Sin embargo, no interpreta que representa dicho tópico y no es el objetivo para el problema planteado. Se necesita que el tópico asignado represente un aspecto general e invariante en el tiempo, dado que nuevos datos (noticias) surgen y nuevos temas son abarcados. Con esta nueva ola de datos LDA crearía nuevos tópico para poder clasificar los nuevos datos (en caso de ser necesario), lo que volvería aún más complicada la representación temporal que se busca. Es por esto que LDA se utiliza en un inicio solamente como método de exploración de los datos, para obtener una idea cuáles son los temas tratados por diferentes cuentas. DDR por otro lado permite una mayor generalización de los tópicos de clasificación y el uso de palabras características de cada tópico permiten captar de mejor manera tópicos futuros. La mayor desventaja que posee DDR es su falta de contextualización de los textos clasificados, la necesidad de diccionarios bien definidos y un word embedding apto para este tipo de problema.

Anexo E. Ejemplos de gráficos de cuentas

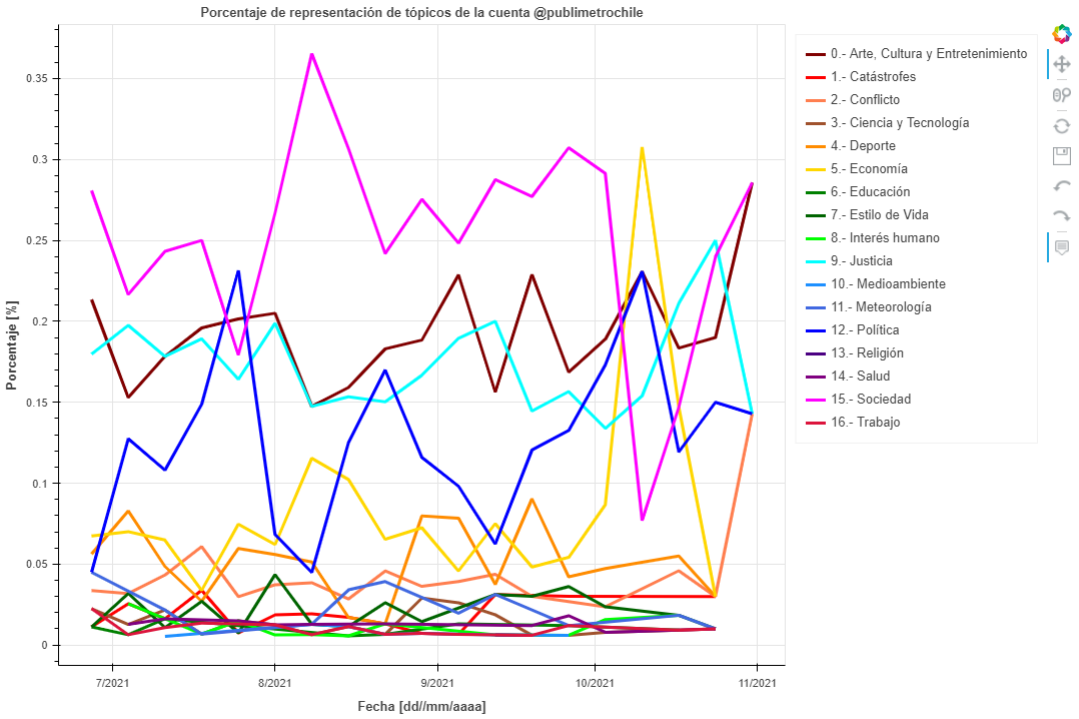

Figura E.1: Cobertura de tópicos a través del tiempo de la cuenta @publimetrochile

Enla Figura E.1 muestra la cobertura de la cuenta de @publimetrochile. En este caso

se puede observar que los tópicos de Sociedad, Arte y Entretenimiento, Justicia y Política comparten gran parte de las publicaciones de la cuenta. Se observa un *peak* de Política cercano al mes de Julio coincidiendo con las elecciones primarias. Varios tópicos se encuentran cercanos al 0 % de representación como podría ser Meteorología, Conflicto o Salud. Esto puede deberse a que varias de las noticias que se referían a algunos de estos tópicos no lograron pasar el filtro designado. Otro motivo de la baja representación pueda deberse a que el enfoque del medio no sean aquellos tópicos que no se cubren.

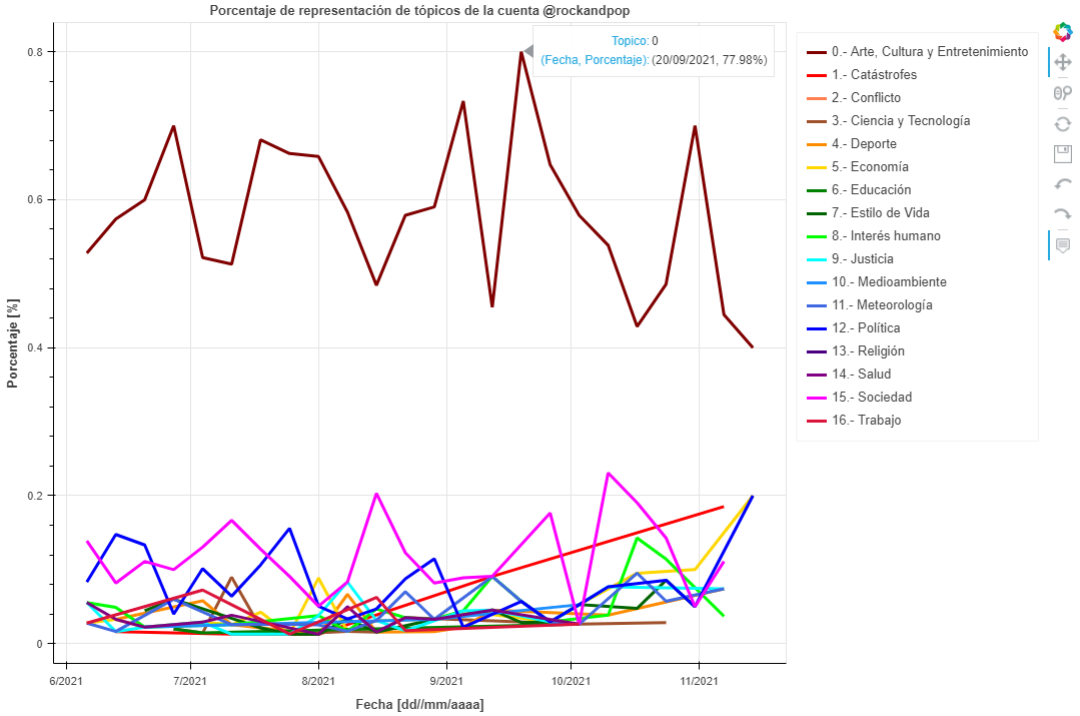

Figura E.2: Cobertura de tópicos a través del tiempo de la cuenta @rockandpop

En la Figura E.2 se observa la evolución temporal de la cuenta de @rockandpop. Como se puede observar esta cuenta claramente se enfoca en lo relacionado al Arte, Entretenimiento, Cultura y Medios de Comunicación. Esto tiene sentido, ya que, la cuenta @rockandpop corresponde a la de la radio 94.1 FM con el mismo nombre he incluso en su descripción detallan que se enfocan principalmente en música. Para este caso se puede notar que DDR logra capturar correctamente el perfil de esta cuenta, donde también se observan una presencia algo considerable del tópico de Sociedad.

Tabla F.1: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Catástrofes y accidentes

Cuenta	RCA
onemichile	12,9635
dichatoaldia	5,5498
laestrellaiqq	5,9171
el_bulnensino	5,8506
serenaycoquimbo	5,3477
soyiquique	5,2318
aconcaguanews	4,9125
publimetro_tv	4,5809
laciudad	4,1874
radioquellon	4,1718

Tabla F.2: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Ciencia y tecnología

Cuenta	RCA
chw_net	40,0363
transmediachile	34,0123
wayerless	28,1487
betazeta	25,0490
fayerwayer	20,8753
niubie_com	19,6991
radiocorazonfm	5,8190
tacometrochile	5,2630
bolido_com	4,8419
quillotadigital	4,7129

Anexo F. Tablas de RCA de los tópicos utilizados

Tabla F.15: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Sociedad

Cuenta	RCA
vanidadeschile	61,4479
imagina881	48,0540
acciondeongs	31,3200
radio_romantica	29,2115
rrppcastro	25,1264
armoniaonline	20,5982
chilevision	20,4572
muevomundo	20,2075
serpradreschile	19,8123
elgraficochile	18,0451

Tabla F.3: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Conflicto, guerra y paz

Cuenta	RCA
verdadahoracl	6,9278
rvfradiopopular	4,3163
puentealto_pald	4,1544
piensaprensa	3,8823
dirariochile	3,5992
s_schwartzmann	3,2803
gamba_cl	3,2494
soyconcepcion	3,1386
biobio	3,1331
soytemuco	3,1033

Tabla F.4: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Deporte

Cuenta	RCA
enelcamarin	39,9488
biobiodeportivo	37,4006
elgraficochile	31,9639
ferplei	25,9379
losriosdeporte	21,7533
deportesarica	19,8975
cooperativa	6,5710
tarapaca_online	5,6598
rreloncavi	5,0122
adnradiochile	4,4076

Tabla F.5: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Economía y finanzas

Cuenta	RCA
zonapublimetro	27,5925
haciendoriqueza	23,3905
estrategiacl	21,8897
americaeconomia	17,4725
aquasocial	16,7545
mundoacuicola	16,4397
la_cav	16,3484
tacometrochile	15,5055
elcanelino	13,1092
bolido_com	12,4704

Tabla F.6: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Educación

Cuenta	RCA
radioubbchile	16,6519
somosmelipilla	7,4425
uchileradio	5,9237
redinfovilla	5,2616
misanfelipe	5,0269
rengonotas	4,9109
portalviregion	4,5903
radiomaxima_fm	4,5593
diarioelcentro	3,8913
austral_losrios	3,4097

Tabla F.7: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Esilo de vida y tiempo libre

Cuenta	RCA
sabrosia	58,2833
womenshealthcl	36,2920
canal_13c	31,3773
lt_finde	22,9312
platosycopas	15,2961
taconeras	11,8186
vanidadeschile	10,2212
enelcamarin	8,0025
la_cav	7,9985
cesar_maturana	7,9561

Tabla F.8: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Interés humano, animales e insólito

Cuenta	RCA
radiodisneyla	18,2192
tacometrochile	6,3856
publimetro_tv	6,3822
pucontv	6,1078
veoverde	5,7176
muyinteresante	5,3640
prensaenlinea6	5,2691
lapulentanews	5,2004
ferplei	4,8779
elpaihuanino	4,6928

Tabla F.9: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Mano de obra

Cuenta	RCA
chilebcl	5,0013
mega	4,7620
estrella_toco	3,2725
liz_try	2,7870
redlosrios	2,7546
tvbiobio	2,7376
redtarapaca	2,6982
redloslagos	2,6507
redarica	2,6148
redatacama	2,5627

Tabla F.10: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Medioambiente

Cuenta	RCA
enlineamaule	11,6689
prensaantartica	10,939
veoverde	9,0586
elcanelino	4,8613
zonapublimetro	3,7424
diario_eha	3,5845
elmontepatrino	3,3797
laotravoz	3,3633
clave9cl	3,3476
elquiglobal	3,2283

Tabla F.11: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Meteorología

Cuenta	RCA
onemichile	10,1132
radio_festival	8,3968
radiocarnavalfm	7,5774
publichanaral	6,7051
radiomaxima_fm	6,6325
serenaycoquimbo	6,4992
publiemtro_tv	6,4774
dlibertador	6,1686
elpaihuanino	5,4352
el_bulnesino	5,0259

Tabla F.12: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Policía y justicia

Cuenta	RCA
carabdechile	59,8548
verdadahoracl	32,4501
ciper	16,5912
radioeme	14,7241
muevomundo	12,5310
diariochile	12,0499
ferplei	10,8499
diariolider	10,1838
radiouniverso	10,0515
acciondeongs	8,7928

Tabla F.13: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Religión y culto

Cuenta	RCA
caribefm	12,8299
prensaenlinea6	11,3331
elpaihuanino	8,9985
maipuciudadano	7,4153
elmontepatrino	6,6352
chilemosaico	6,0505
chiloealdia	5,9269
prensatuciudad	5,6540
tribunac	5,5492
elhurtadino	5,4487

Tabla F.14: Ranking de las 10 cuentas con mayor valor de RCA para el tópico Salud

Cuenta	RCA
carta_abierta	30,9034
gvalpo	19,4893
fayerwayer	13,3534
hoyxhoycl	11,7477
tarapaca_online	11,6445
buenasaludchile	10,7464
radiouniverso	10,0575
fmpluschile	8,9564
quintavision	6,8835
muyinteresante	6,6404