UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

ANÁLISIS DEL DIÁLOGO SOBRE POLÍTICOS LATINOAMERICANOS EN TWITTER

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

MIGUEL ÁNGEL FERNANDO ZÚÑIGA GONZÁLEZ

PROFESOR GUÍA:
AIDAN HOGAN

MIEMBROS DE LA COMISIÓN:
CLAUDIO GUTIÉRREZ GALLARDO
PABLO GONZÁLEZ JURE

SANTIAGO DE CHILE
2020

Resumen

Twitter es una aplicación de microblogging que permite enviar mensajes de texto plano de corta longitud, con un máximo de 280 caracteres, llamados tweets. Durante el transcurso de los años, esta aplicación se ha convertido en una de las redes sociales más utilizadas alrededor del mundo, contando con más de 328 millones de usuarios activos. Esto genera miles de millones de tweets diarios con información del pensar de los usuarios.

Utilizando Wikidata como fuente de información, y una base de datos del Instituto Milenio de Fundamentos de los Datos compuesta del 1 % de los tweets emitidos de manera global durante los últimos 6 meses del año 2017, el desafío de este trabajo es; enlazar conjuntos de publicaciones en Twitter referentes a políticos latinoamericanos con Wikidata de tal manera que permita entender mejor como se manifiestan las conversaciones políticas de la región latinoamericana, añadiendo factores como el género, el partido político, el lugar de procedencia del político al que se le hace referencía, entre otros. Este resultado permitirá el desarrollo de proyectos y análisis futuros sobre la conversación política en Twitter, tanto nacional como a escala continental, permitiendo profundizar o investigar temas del día a día.

Durante la realización de este trabajo se generó una base de datos correspondiente a un filtro de los tweets anteriormente mencionada; dicha base de datos incluyó los tweets que mencionan usuarios de Twitter asociados a políticos latinoamericanos, tweets emitidos por políticos latinoamericanos, y tweets que mencionan el nombre de los políticos latinoamericanos mas no necesariamente su usuario de Twitter. Esta información vino acompañada de los meta-datos de Wikidata. Además de lo anteriormente señalado, se mostró y caracterizó, a modo de prueba de concepto, el análisis de sentimiento efectuado sobre los tweets.

La generación de la base de datos terminó permitiendo no tan sólo hacer una caracterización del pensamiento general de los usuarios de Twitter hacia los políticos de manera exitosa, incluyendo la presencia de estos en la red social (proceso hecho con una metodología replicable en otros ámbitos siempre que se cuente con un conjunto de meta-datos similar), sino que también genera la posibilidad de caracterizar a los propios usuarios; proceso que puede servir para extraer bases de datos específicas de ciertos paises y así analizarlos en mayor detalle.

Cómo podemos votar por un sistema que habla de dignidad y acto seguido atropella lo más sagrado de la dignidad humana; que es la libertad de conciencia.

- Mario Moreno 'Cantinflas'

Agradecimientos

Agradezco profundamente a todos los que me han acompañado en este arduo proceso. Destacando a mi familia, por sobre todo mis padres José Miguel Zúñiga Rodriguez y Gloria Beatriz González Tejos.

Tambien aprovecho de agradecer a aquellos amigos de la Universidad que me fui haciendo en el camino Jorge Vega, Vicente Silva, Daniel Segovia, Luis Felipe Castro, Jorge Quiroz, Solange Vivanco, Juan Millán, Tomás Morales, Camilo Utrera, Nicolás Sepulveda y por sobretodo a mi novia Florencia Baez que siempre fue un apoyo en este arduo proceso.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Desde hace una década ha habido un incremento exponencial en la cantidad de datos disponibles en internet; dichos datos generan una información abrumadora con un potencial infinito. Año tras año la accesibilidad a la internet es mayor, permitiendo acceso en tiempo real a millones de personas. Esto permite también que figuras del mundo del cine, la música, la política y cualquier referente pueda desplegar su opinión en distintas plataformas, llegando a una audiencia de magnitud de millones, de manera instantánea.

Corresponde un desafío convertir esos datos en información comprensible para las personas: información que nos ayude día a día a tomar mejores decisiones.

1.1. Situación actual

En la actualidad existen múltiples aplicaciones y plataformas que permiten a personas de todo el mundo interactuar entre sí. Dichas plataformas son capaces de almacenar millones de datos que, no hace mucho tiempo, no podían ser registrados de igual manera. A raíz de esto se genera la oportunidad de trabajar con millones de datos, así como procesarlos para ser capaz de concluir un análisis a raíz de éste.

Un claro ejemplo de lo antes señalado es la aplicación Twitter, la cual consta con millones de usuarios asociados que publican diariamente. Aunque la aplicación permite a sus usuarios compartir noticias y otro contenido de una forma muy flexible, dichas publicaciones han generado durante los años una gran cantidad de mensajes de odio, tales como extremismo político, misoginia y denigración de personas.

De todo esto existe una gran cantidad de información, la cual podría servir para reconocer patrones en diferentes entidades públicas. Específicamente se pretende generar una base de datos con los tweets que hagan referencia a los políticos latinoamericanos, para que de esta manera obtener una herramienta que enriquezca el estudio de diferentes materias de las ciencias sociales y políticas, como por ejemplo el descontento social que hay por parte de la ciudadanía con sus representantes.

Además, debido a que los datos entregados por Twitter son netamente texto, estos no poseen

muchos metadatos sobre los usuarios, situaciones y contextos en general a los que se les hacen referencia. Hoy en día un tweet como "Renuncia hoy, presidente Piñera" no es capaz de determinar a quién hace referencia; esto se debe porque el dato es sólo texto y no se utiliza ninguna referencia en el tweet (a pesar de nombrar a la figura pública por su apellido).

Hoy en día no existe ningún tipo de red social masiva de microblogging o de cualquier índole que genere una base de datos con los detalles de las entidades. En el caso específico de las menciones a políticos latinoamericanos, determinar que en un tweet hace mención a un político sin hacer referencia directa a su usuario no es una tarea sencilla, esto debido a las diferentes formas de mencionarlo, como también la coloquialidad del lenguaje empleado.

También es necesario generar algún tipo de conexión con otras fuentes de datos para obtener información sobre el lugar de procedencia del político, su género, el partido al cual está afiliado, entre otras características. Por lo tanto, se puede determinar que los datos que puede entregar Twitter no son suficientes por sí mismos como para realizar un análisis sobre las publicaciones de esta índole o similar.

1.2. Solución propuesta

Por medio de varios conjuntos de datos recolectados de Twitter por parte del Instituto Milenio de Fundamentos de Datos (IMFD)[1], los cuales corresponden al 1 % de tweets emitidos durante el año 2017, siendo estos miles de millones de publicaciones genéricas. Se desea crear y trabajar con un subgrupo de éstos, orientados a políticos latinoamericanos. Todo esto con la intención de asociar los tweets que mencionan a políticos con los metadatos de estos mismos; como la nacionalidad, el sector político al cual representan, el género, la edad, entre otros.

Esto se llevaría a cabo usando "entity linking", lo cual se puede definir como la conexión de las entidades mencionadas en un texto con referencias correspondientes en una base de conocimientos. Por lo tanto, con sólo saber el nombre de la figura pública a la que se hace referencia, se podrían extraer, al menos en teoría, los datos antes señalados.

Utilizando la base de conocimientos Wikidata se pretende extraer un set de carácteristicas de los metadatos previmente señalados, incluyendo los políticos latinoamericanos. Una vez relacionados los datos, se desea establecer múltiples análisis para evaluar las características de los datos obtenidos. Dentro de los análisis a ejecutar se encuentra la evaluación de los sentimientos asociados a las publicaciones. De esta manera un tweet como "Renuncia hoy, Piñera incompetente" podrá catalogarse como un tweet asociado a sentimientos negativos hacia la figura pública. Además usando entity linking con Wikidata podemos indentificar que esta figura es el presidente actual de Chile, un hombre ex miembro de Chile Vamos, ex alumno de la Pontificia Universidad Católica de Chile, etc.

Se espera no tan sólo poder implementar a cabalidad el uso óptimo de las herramientas y técnicas ya señaladas, sino también poder ayudar a los múltiples expertos en ciencias sociales del IMFD a crear bases de datos más amigables para trabajar y sacar conclusiones que evidencie parte de la actual cultura latinoamericana y permita hacer un análisis exhaustivo

[1]Recolectado por la Profa. Bárbara Poblete, Hernán Sarmiento, entre otros.

de los tweets.

1.3. Objetivos

1.3.1. Objetivo General

Enlazar conjuntos de publicaciones en Twitter referentes a políticos latinoamericanos a una base de metadatos de tal manera que permita entender mejor cómo se manifiesta las conversaciones políticas de la región latinoamericana, añadiendo factores como el género, el partido político, el lugar de procedencia del político al que se le hace referencia, entre otros.

1.3.2. Objetivos específicos

1. Recolectar gran número de datos correspondiente a publicaciones de usuarios en Twitter.

2. Crear subgrupo de datos a partir de los previamente recolectados que tengan un mayor porcentaje de publicaciones de latinoamérica y/o asociadas a políticas de esta región.

3. Identificar qué publicación hace referencia a qué político, tanto directamente, que puede ser refiriéndose a la persona por su usuario de Twitter, como indirectamente, que puede ser refiriéndose a la persona por su nombre en texto plano..

4. Validar la calidad de los datos resultantes y la precisión de los métodos ocupados.

5. Realizar el análisis de los datos, que permita sacar conclusiones sobre las temáticas y sentimientos asociados a dichas publicaciones en la región de Latinoamérica.

1.4. Conocimientos requeridos y desafíos del trabajo

Para el desarrollo del trabajo en cuestión, dados los objetivos propuestos, se requerirá hacer uso de diversas técnicas computacionales, las cuales son:

- **Conocimiento de bases de datos relacionales**: Es requerido un conocimiento de las bases de datos relacionales, en particular, un buen manejo de SQL, esto debido a que la extracción de los datos correspondientes a usuarios de Twitter y tweets que se ocuparon en el desarrollo de este trabajo estaba almacenado en este modelo de base de datos.
- **Manejo de grafos de conocimiento**: Es esencial el conocimiento del funcionamiento de la base de conocimientos, para extraer la información referente a los políticos por medio de consultas. Es requerido tener buen manejo de SPARQL y ser capaz de reconocer entidades, para así generar la extracción de la misma.
- **Manejo de herramientas para procesar grandes cantidades de datos**: Dada la gran cantidad de datos que se tuvo desde un principio (siendo estos de la escala de millones de datos), se requiere poseer capacidades para el manejo de los mismos. Durante el transcurso de estre trabajo se ocupó la librería Pandas de Python.

- **Comprensión de técnicas de reconocimiento y detección de entidades**: Para el cumpliento del trabajo se requerirá poder trabajar con alguna herramienta que permita el reconocimiento de entidades, evaluando su comportamiento y diagnosticando cuáles condiciones pueden generar un mejor desempeño en el caso particular de textos pequeños. El propósito de esta técnica es buscar entidades mencionadas en el texto que hagan referencia a personas del mundo de la política latinoamericana, para luego vincular estas menciones de entidades en texto con sus entidades correspondientes en la base de conocimiento (Entity Linking).

1.5. Contribuciones

Este trabajo entrega las siguientes contribuciones:

- Una base de datos a partir de la extracción de tweets, que hagan mención a algún político latinoamericano o que hayan sido publicados por algún político latinoamericano.
- Caracterización general de los tweets emitidos por políticos latinoamericanos.

Se espera que la información extraída a partir de este trabajo pueda servir para futuros trabajos de investigación de múltiples disciplinas.

1.6. Metodología

La metodología implementada busca cumplir los objetivos ya mencionados. Dicha metodología puede dividirse en múltiples etapas.

La primera de dichas etapas corresponde a la extracción de los datos de las entidades de políticos latinoamericanos por medio de alguna base de conocimientos. Dichas entidades deben corresponder a políticos que sigan vivos, y que entreguen tanto la información de su nombre y país, como también (en el caso de que la entidad cuente con la información correspondiente) el partido político al cual representa, su género y cuenta de Twitter. Para el caso del desarrollo de este trabajo se ocupó la base de conocimientos Wikidata, para extraer el conjunto de entidades antes señalados, para esta esta primera sección del proceso se requirió tener un manejo sobre grafos de conocimiento acompañado de un buen manejo de SPARQL (lenguaje con el que se requerían hacer las consultas de los datos).

La segunda parte está relacionada con la extracción de los datos de tweets, obteniendo información como el texto del mismo, el usuario asociado y el identificador del tweet en cuestión. Para el presente trabajo se solicitó acceso a una base de datos de tweets previamente recopilados por el Instituto Milenio de Fundamentos de los Datos (IMFD). La institución otorgó acceso a los datos, la cual correspondía a una base de datos relacional, con múltiples tablas separadas por día de emisión, con datos de usuario, tweet y otras características. Cada una de estas tablas poseía cientos de millones de datos.

Como tercera parte se espera poder filtrar los datos, dejando sólo aquellos que estén asociados con tweets en español. Luego de este primer filtro se seleccionarán aquellos que hagan mención directa o indirecta de los políticos, es decir los que escriben el nombre del

usuario de Twitter del político, como por ejemplo "Eres un gran líder @PepeMujicaDice" habla directamente al ex presidente de Uruguay José Mujica puesto que lo menciona ocupando su usuario. Por otro lado una mención indirecta vendría siendo el sólo nombrar al político pero sin usar su usuario de Twitter, como por ejemplo "Maduro es el presidente de Venezuela". Se requirió conocimiento de consultas SQL y manejo de base de datos relacionales para extraer los datos correspondientes, filtrando por sólo aquellos tweets generados en español. Además, cabe recalcar que, los datos en su gran mayoría se encontraban incompletos.

Para detectar la mención directa e indirecta se ocuparán diferentes formas:

- **Mención directa :** Se leerá el texto, y se extraerá de este los usuarios mencionados por el mismo; esta tarea resulta sencilla si es que se cuenta con una lista de usuarios referentes a políticos latinoamericanos, ya que se puede detectar el usuario de Twitter al ser una palabra que empieza con el símbolo "@". Para conseguir dicha lista se utilizarán las consultas de Wikidata seleccionando aquellos políticos con cuentas de Twitter asociadas. Si bien la tarea no resulta ser de gran envergadura, esta se complica al manejar un conjunto de datos de la escala antes mencionada, por lo mismo se optimizó el código ocupando la librería de manejo de datos de Python; Pandas.
- **Mención Indirecta :** Para detectar la mención indirecta se ocupará una librería de Python llamada pyspothlight, la cual corresponde a un Python wrapper de DBpedia Spotlight, de la cual se hablará más adelante. Dicha librería permitirá (a grandes rasgos) detectar las entidades asociadas al tweet, está tarea resulta más compleja debido al ruido generado en el tweet, ya sea por que la entidad no es claramente mencionada o por algún tipo de falla ortográfica, para estos se ajustarán los parámetros de la librería de Python con tal que, según experimentos previos, funcione con mayor precisión. Luego de tener las entidades del tweets se revisará si alguno de aquellos resultados corresponden a uno de los políticos previamente extraídos.

Como quinta parte, se filtrarán los tweets que procedan de un ID que corresponda a alguno de los usuarios de Twitter asociados con el grupo de políticos latinoamericanos con el que se está trabajando. Estos tweets corresponden a publicaciones emitidas por el mismo político.

La sexta parte, corresponde a interconectar los datos extraídos tanto de la base de conocimientos, como de la base de tweets en cuestión, a su vez asignando a cada uno de los datos un análisis de sentimientos del mismo. Inicialmente se pretendía hacer el análisis de sentimiento como parte del trabajo, pero los datos de los cuales se extrajeron los tweets ya tenían una tabla de análisis de sentimientos asociados, por lo cual no fue necesario este procedimiento.

Por último se busca realizar experimentos con los datos extraídos, evidenciando de forma clara como están representados los políticos latinoamericanos en Twitter, y a su vez mostrando como es que ellos mismos se expresan en la popular plataforma cibernética.

1.7. Resultados esperados

El principal resultado que se espera obtener de esta memoria es la creación de una base de datos que asocie tweets con políticos latinoamericanos de una base de conocimientos. Este resultado permitirá desarrollo de proyectos y análisis futuros sobre la política tanto nacional

como a escala continental, permitiendo profundizar o investigar temas del día a día.

Se espera también obtener resultados caracterizando la base de datos, como la comparación numérica de tanto tweets totales con los tweets de la base de datos, como tweets totales con las menciones directas/indirectas de políticos, o la distribución por género de la cantidad de políticos mencionados, entre otros resultados. Por otra parte se espera poder extraer del análisis de sentimientos ya hecho, resultados que caractericen de mejor manera a los políticos; se pretende agrupar los sentimientos emitidos hacia una figura pública del mundo de la política y evidenciar los sentimientos que este logra generar en tanto en el común de la gente, como entre sus pares.

Capítulo 2

Estado del Arte

En el siguiente capítulo se espera contextualizar sobre los avances contemporáneos a este trabajo de investigación, los cuales puedan apoyar y servir para el desarrollo del mismo.

Cabe destacar que al ser la mayoría de las investigaciones hechas en tweets en inglés, no todos los trabajos son directamente comparables o aplicables al trabajo en cuestión. No obstante, dichos trabajos forman una base sólida sobre el cual construir el trabajo.

2.1. Twitter

Twitter es un servicio de microblogging; lo que significa que es una forma de comunicación o sistema de publicación que consiste en el envío de mensajes cortos de texto. En este caso en particular los mensajes del autor están limitados a 280 caracteres o menos [1]. Los mensajes se conocen como tweets. Twitter posee la opción de seguir a alguien, lo que significa que, si es que sigo a una persona en Twitter, obtengo sus actualizaciones en mi línea de tiempo personal, convirtiéndose en su seguidor. Por el contrario si alguien te sigue, ellos son tus seguidores. En el caso de que ambos se sigan mutuamente, entonces pueden intercambiar mensajes directos privados, referido como DMs [2]. Los hashtags en Twitter han sido desarrollados como una forma para que las personas categoricen sus tweets. El tema de un tweet generalmente se denota colocando un "#" delante del tema, como por ejemplo "#GOT".

Twitter en la actualidad almacena millones de publicaciones que hacen referencia a variados temas; se pretende extraer de esta plataforma el conjunto de datos en el cual se realizará el análisis, puesto que consta con millones de usuarios alrededor del mundo que publican constantemente.

Una vez extraído una base de datos de Twitter se espera filtrar los tweets por su idioma y, de ser posible, si es que estos fueron emitidos desde un país latinoamericano. Para este tipo de filtrado de Twitter ya existen estudios que proveen herramientas para la clasificación del tweet [3] que permiten no sólo identificar el lenguaje, sino también poder hacer esto mientras el tweet mencione el nombre del político en cuestión.

2.2. Bases de conocimiento

Se define una base de conocimientos (o KB por sus siglas en inglés) como una tecnología utilizada para almacenar información compleja tanto estructurada como no estructurada, utilizada por un sistema informático. Dentro de las bases de conocimiento más ocupadas actualmente se encuentran Wikidata [4] y DBpedia [5].

2.2.1. Wikidata

Wikidata es una base de conocimientos que se distingue por ser libre y colaborativa. Esta tiene por propósito gestionar a nivel mundial los datos anexados a Wikipedia de la mejor manera posible, puesto que la naturaleza de Wikipedia no permite gestionar de manera eficiente los 30 millones de artículos con datos en 287 idiomas diferentes que hoy en día hay en la herramienta. Actualmente Wikidata contiene 71,611,020 items, las cuales son revisadas y editadas por alrededor de 24.186 personas que corresponden a los usuarios activos. Los datos de Wikidata corresponden en su gran mayoría a artículos escolares (31.5 %), seguidas de entidades sin categoría (25.5 %) y referencias a seres humanos (8.9 %); este último valor corresponde a 6.376.879 entidades [6].

Se busca utilizar Wikidata como herramienta para agregar aquellos datos faltantes a los tweets, tales como a quién está referido, género, visión política de la figura pública a la que se alude, entre otros. De esta manera, por ejemplo; un tweet como "Me encanta el alcalde Jadue" se puede vincular con datos propios de la persona a la que se hace referencia como cargo público; se provee una muestra de estos datos en la Figura 2.1, en donde los datos completos incluyen la nacionalidad del político, los cargos que ha tenido, la fecha de su nacimiento, los partidos políticos en los cuales ha militado, su género, entre muchos otros atributos.

Wikidata Query Service

Wikidata Query Service es un servicio de consultas de la base de conocimientos de Wikidata. El Servicio de Consultas de Wikidata (WDQS) proporciona una forma para que las herramientas accedan a los datos de Wikidata, a través de una API SPARQL [7] . Dicha API además posee múltiples ejemplos que sirve como guía para la generación de consultas. Por algunos de estos ejemplos se encuentran; " Cantidad de humanos que se encuentran instanciados en Wikidata", "Políticos que hayan muerto de algún tipo de cáncer", "Seres humanos sin descendencia", entre otros.

En la Figura 2.2 se puede ejemplificar la forma en que se hace la consulta en Wikidata Query Service, haciendo mención directa a las propiedades (por ejemplo wdt:P31) y las entidades de la base de conocimientos (wd:Q6256).

Por otra parte la Figura 2.3 es una muestra del resultado obtenido por la consulta anterior. Dicha respuesta es una vez más una ejemplificación del trabajo que se espera realizar ocupando este servicio.

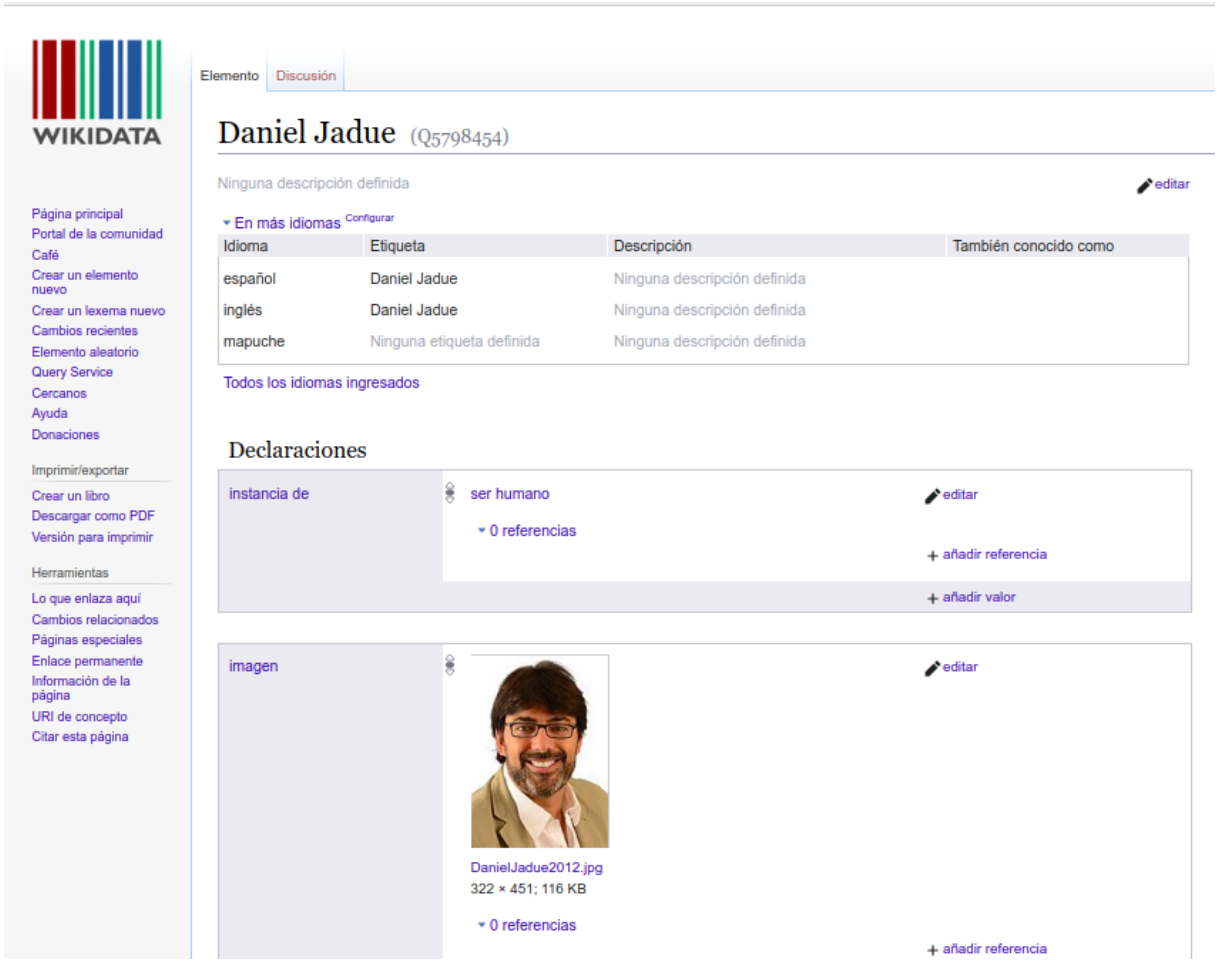


Figura 2.1: Perfil de político en Wikidata

politician	politicianLabel	country	countryLabel	gender	genderLabel	party	partyLabel
wd:Q173238	Carlos Alberto Reutemann	wd:Q414	Argentina	wd:Q6581097	masculino	wd:Q1053668	Partido Justicialista
wd:Q178649	Romário	wd:Q155	Brasil	wd:Q6581097	masculino	wd:Q2054789	Partido Socialista Brasileño
wd:Q179265	Julia Carabias Lillo	wd:Q96	México	wd:Q6581072	femenino		

Figura 2.3: Resultado consulta Wikidata

Ocupando Wikidata Query Services se espera extraer un conjunto de datos que contenga políticos latinoamericanos, con los atributos suficientes para determinar el género, la edad, el país en el que viven, su nombre, si es que tienen usuarios de Twitter, si es que están afiliados o estuvieron a algún partido político, y si es que están muertos.

2.2.2. DBpedia

DBpedia es un proyecto comunitario que extrae conocimiento estructurado y multilingüe de Wikipedia y lo hace libremente. Este se encuentra disponible en la Web utilizando tecno-

```
SELECT ?politician ?politicianLabel ?username ?cause ?date ?deathPlace
?gender ?genderLabel ?party ?partyLabel ?country ? countryLabel
WHERE
{
   #Latin American country
   ?country wdt:P31 wd:Q6256.
   ?country wdt:P361 wd:Q12585.

   #Politician of a Latin American Country
   ?politician wdt:P31 wd:Q5.
   ?politician wdt:P27 ?country.
   ?politician wdt:P106 wd:Q82955.

   #Have Twitter account
   OPTIONAL {?politician wdt:P2002 ?username }

   #They are dead
   OPTIONAL {?politician wdt:P509 ?cause}
   OPTIONAL {?politician wdt:P570 ?date}
   OPTIONAL {?politician wdt:P20 ?deathPlace}

   #Their gender
   OPTIONAL {?politician wdt:P21 ?gender}

   #Their party
   OPTIONAL {?politician wdt:P102 ?party}

   #Gets labels
   SERVICE wikibase:label { bd:serviceParam wikibase:language "es,en" }
}
```

Figura 2.2: Consulta Wikidata Query Service sobre todas las personas de latinoamérica que sean políticos

logías de la Web Semántica y Datos Vinculados. El proyecto extrae, de forma automática, conocimientos de diferentes ediciones de Wikipedia en diferentes idiomas.

La Figura 2.4 muestra el perfil de las entidades en DBpedia, en este caso el político chileno Sebastián Piñera; al igual que su contraparte en Wikidata, este perfil muestra múltiples propiedades de la entidad, acompañados de una breve introducción al personaje, estas propiedades incluyen el nombre, la fecha de nacimiento, los cargos públicos que ha ejercido, entre otros.

Se pretende (como alternativa a Wikidata) utilizar esta base de conocimientos para agregar meta-datos a los tweets, vinculándolos con información relevante para cuando se desee hacer los análisis. Se va a preferir usar inicialmente Wikidata para el trabajo, debido a que la extracción de datos de DBpedia es automatizada y esto puede repercutir en que existan errores en la extracción que sean más difíciles de detectar.

Muchas de las herramientas utilizadas hoy en día para la realización de entity linking se

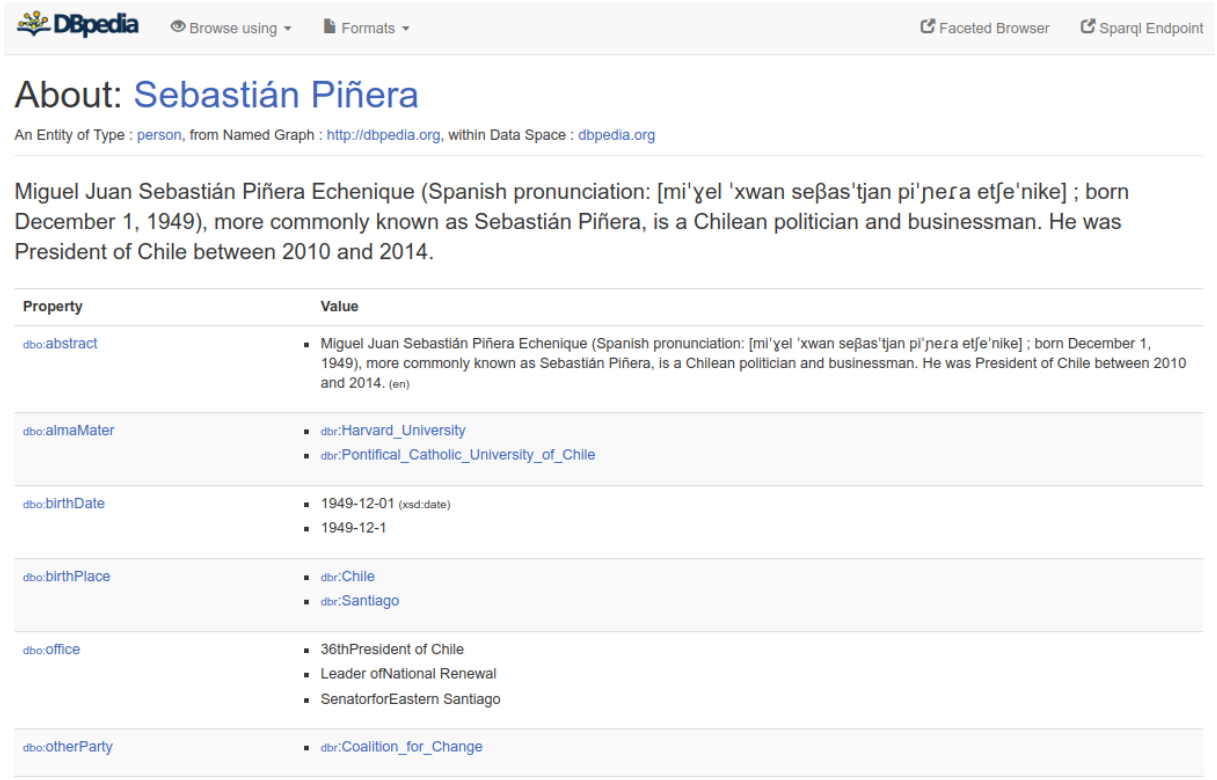


Figura 2.4: Perfil de político Sebastián Piñera en DBpedia

usan utilizando como base de conocimientos DBpedia; si bien no se ocupará de manera directa DBpedia, esta resulta ser un recurso indirecto en el trabajo, ya que esta base de conocimientos es esencial para la herramienta DBpedia Spotlight (del cual se hará mención a mayor profundidad en los siguientes puntos), la cual será la herramienta a utilizar para la detección de entidades en los tweets.

2.3. Uso de Entity Linking

Entity linking es la tarea de vincular las menciones de entidades en texto con sus entidades correspondientes en una base de conocimientos. Las aplicaciones potenciales incluyen extracción de información, recuperación de información y población de base de conocimientos. Sin embargo, esta tarea es un reto debido a las variaciones de nombre y la ambigüedad de la entidad [8]. En la Figura 2.5 se puede ver un ejemplo de mención directa de la cuenta de Twitter de una figura pública, mientras que la Figura 2.6 hace referencia a esta misma figura por medio del apellido del mandatario (de forma indirecta). El uso de Entity Linking en estos textos podrá detectar el nombre de Sebastián Piñera y conectar dicha referencia a la entidad Q306 de Wikidata [9] y de está manera poder tener acceso a múltiples atributos tanto del mandatario como los demás políticos de latinoamérica.

Hoy existen diversos métodos para la generación de entity linking, los cuales son orientados generalmente a campos específicos de textos.

Uno de dichos métodos es ADEL [10], método que se puede ocupar para datos tanto genéricos

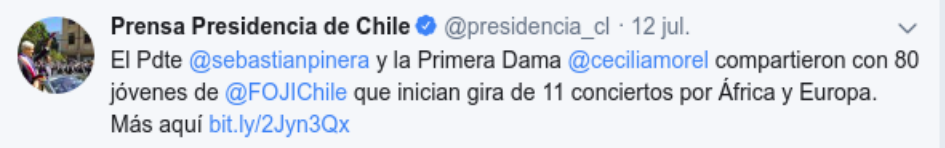


Figura 2.5: Mención directa del presidente Sebastián Piñera en Twitter

Figura 2.6: Mención indirecta del presidente Sebastián Piñera en Twitter

como específicos. ADEL está basado en un enfoque híbrido que combina diversos métodos de extracción para mejorar el nivel de reconocimiento y un proceso eficiente de indexación de la base de conocimientos para aumentar la eficiencia del paso de enlace [11]; cabe destacar que ADEL puede usarse tanto en DBpedia como MusicBrainz.

Babelfly corresponde a otro método comúnmente usado, que ocupa desambigüedad del sentido de las palabras [12] (WSD por sus siglas en inglés); por esta razón, el método es capaz de reconocer entidades sin mayor contexto y generar Entity Linking.

Se desea utilizar un método actual en un subconjunto específico de tweets latinoamericanos, para unir los datos de las publicaciones extraídas de Twitter con las bases de conocimientos, vinculando de esta manera las publicaciones asociadas a políticos con las entidades correspondientes en la base de conocimientos.

2.3.1. Entity Linking en Twitter

Twitter se ha convertido en una de las plataformas más grandes de comunicación de la última década, lo que lo convierte en una herramienta que provee datos claves para diversos usos, como detección de tendencias o monitoreo de marcas, entre otros. Pero existen varios problemas recurrentes en el contexto de Entity Linking, siendo el más importante el hecho de que muchas veces los tweets hacen referencia a personas o situaciones de forma implícita, como por ejemplo, en vez de hablar de "Sebastián Piñera" se hace alusión a esta persona como "El presidente", cita que se infiere por el contexto del emisor del mensaje y del texto que rodea la mención. Dicho problema se acentúa en Twitter ya que el contexto dado en una

red de microblogging es aun menor [13].

Otros problemas asociados con procesar las publicaciones de Microblogging son las jergas propias de cada persona o cultura, las abreviaturas de las palabras y la manera informal en que se suele hablar por este medio. Debido a esto, se han creado múltiples herramientas para poder hacer Entity Linking con textos de poca cantidad de caracteres. De hecho, se ha llegado a hacer un desafío dentro de la comunidad científica para hacer Entity linking de tweets en inglés a DBpedia con entidades que fuesen mencionadas explícitamente dentro del tweet [14].

Entity Model Networks corresponde a un método que toma tanto el conocimiento en sí de la entidad, como el contexto asociado a esta y a partir de eso generar un modelo de identidad integrado [15]. Se ha establecido que generando un modelo de entidades integradas estos son capaces de reconocer entidades a las que se le hace mención implícitamente [13].

2.3.2. DBpedia Spotlight

DBpedia Spotlight es un sistema para anotar automáticamente documentos de texto con URI de DBpedia. DBpedia Spotlight permite a los usuarios configurar las anotaciones a sus necesidades específicas a través de la base de conocimientos de DBpedia y medidas de calidad como prominencia, pertinencia tópica, ambigüedad contextual y confianza de desambiguación. DBpedia Spotlight se comparte como código abierto y se implementa como un servicio web disponible gratuitamente para uso público [16] .

Dado que el código de esta herramienta está abierto se utilizará dicha DBpedia Spotlight para detectar entre los millones de tweets todos aquellos que, hiciesen mención de una persona que según la tabla de datos ya extraída, corresponda a alguno de los políticos con los que se plantea trabajar.

2.4. Análisis de sentimiento

El análisis de sentimiento (también llamado opinion mining) se refiere a la aplicación del procesamiento del lenguaje natural, la lingüística computacional y el análisis de texto para identificar y clasificar opiniones subjetivas. En términos generales, el análisis de sentimiento tiene como objetivo determinar la actitud de un escritor con respecto a algún tema o la polaridad contextual general de un documento [17]. La actitud puede ser su juicio o evaluación, su estado afectivo (es decir, el estado emocional del autor al escribir), o la comunicación motivacional deseada.

Se propone utilizar esta aplicación de procesamiento para detectar los sentimientos asociados a los datos extraídos; cabe señalar que esta práctica ya ha sido hecha [18], trayendo consigo resultados positivos del uso de la aplicación.

2.4.1. Clasificación de polaridad

Una de las formas más populares de aplicación de análisis de sentimiento es usar la clasificación de polaridad (sentiment/polarity classification). Dicha clasificación se traduce en, según

las palabras contenidas en el texto, determinar la orientación del sentimiento observado en el texto.

En otras palabras, el texto se determina si es positiva o negativa con respecto a un tema específico. Sin embargo, este no es un tema simple de resolver, ya que dependiendo del contexto hay palabras que pueden expresar tanto una opinión positiva como negativa. En el caso de tener 2 polaridades, que es más usado en la literatura, cada mensaje puede ser catalogado como positivo o negativo. En este método, se incluyen estudios sobre distintos 8 tópicos, como lo son las reseñas de películas o libros ("bueno" o "malo"), opinión de productos ("me gusta" o "no me gusta") o en elecciones políticas ("va a ganar", "no va a ganar"). Además, se han considerado comúnmente 6 emociones universales [19]: enojo, disgusto, miedo, alegría, tristeza y sorpresa. De esta manera es posible catalogar los tweets de acuerdo a estas emociones de manera de determinar su polaridad y el grado de ésta, lo cual puede ser de gran utilidad para diferenciar mensajes en una mayor cantidad de categorías y no sólo positivo-negativo. De la misma forma, es posible hacer una escala gradual entre positivo-negativo, pudiendo tener 7 grados (3 positivos, 1 neutro y 3 negativos, variando de muy negativo a muy positivo) o 11 grados (-5 al 5, siendo -5 muy negativo, 0 neutro y $+5$ muy positivo) [20].

2.4.2. SentiStrength

SentiStrength corresponde a una librería de Java utilizada para generar análisis de sentimiento. Esta herramienta ha sido clasificada como un sistema de detección de fuerza de sentimiento rápido en su uso sobre tweets en español [21]. Si bien esta herramienta no se usará directamente en este trabajo, se utilizará una base de datos que fue previamente analizada con esta herramienta.

Capítulo 3

Extracción y filtración de datos

Este capítulo detalla el procedimiento correspondiente a la extracción de los datos utilizados, tanto los datos extraídos de Wikidata como los tweets, y el proceso de filtrado que se utilizó.

3.1. Extracción de datos de Wikidata

Lo primero en desarrollarse en este trabajo de memoria fue la extracción de los datos de políticos latinoamericanos de Wikidata. Si bien los datos de la base de conocimiento cuentan con variados atributos, se decidió ocupar el servicio de consultas Wikidata Query Service para extraer aquellas personas que se definan en la página como políticos. En el portal mencionado se consultaron aquellos entidades que estuviesen en la base de conocimientos, que se tratasen de seres humanos y que adicionalmente tuviesen como una de sus características ser políticos.

Además de esto, se decidió trabajar netamente con políticos que en el inicio de la memoria (Mayo 2019) estuviesen vivos. Esta decisión se hizo principalmente por dos propósitos; primero, para evitar que las entidades actuales fuesen comparados con personajes que vivieron hace más de dos siglos, y segundo, para no generar ruido con nombres que hoy en día se pudiesen ocupar para otro propósito, como es el caso de muchos de los libertadores de latinoamérica, con universidades, plazas o calles, como es el caso de Diego Portales, quien puede ser reconocido como político y universidad en la base de conocimientos como se muestra en la Figura 3.1.

Lamentablemente Wikidata no cuenta con un atributo que directamente indique si la persona está fallecida pero ofrece diversos posibles atributos por los cuales se entiende que ha muerto, como lo son; fecha de muerte, causa de muerte y lugar de muerte.

Otro detalle importante de mencionar sobre la extracción de los datos en Wikidata es la incompletitud de muchos de los datos; al ser una base de conocimientos colaborativa, la cantidad de información que abunda en los datos varía dependiendo de la entidad. Debido a esto se optó por tomar como atributo obligatorio solamente el país y la entidad del político, dejando como requisito opcional el o los partidos por los cuales el político ha militado, el género con el cual es identificado actualmente el político y el usuario de Twitter con el cual es asociado. Este último atributo llegó a ser esencial para el trabajo de la memoria, permitiendo

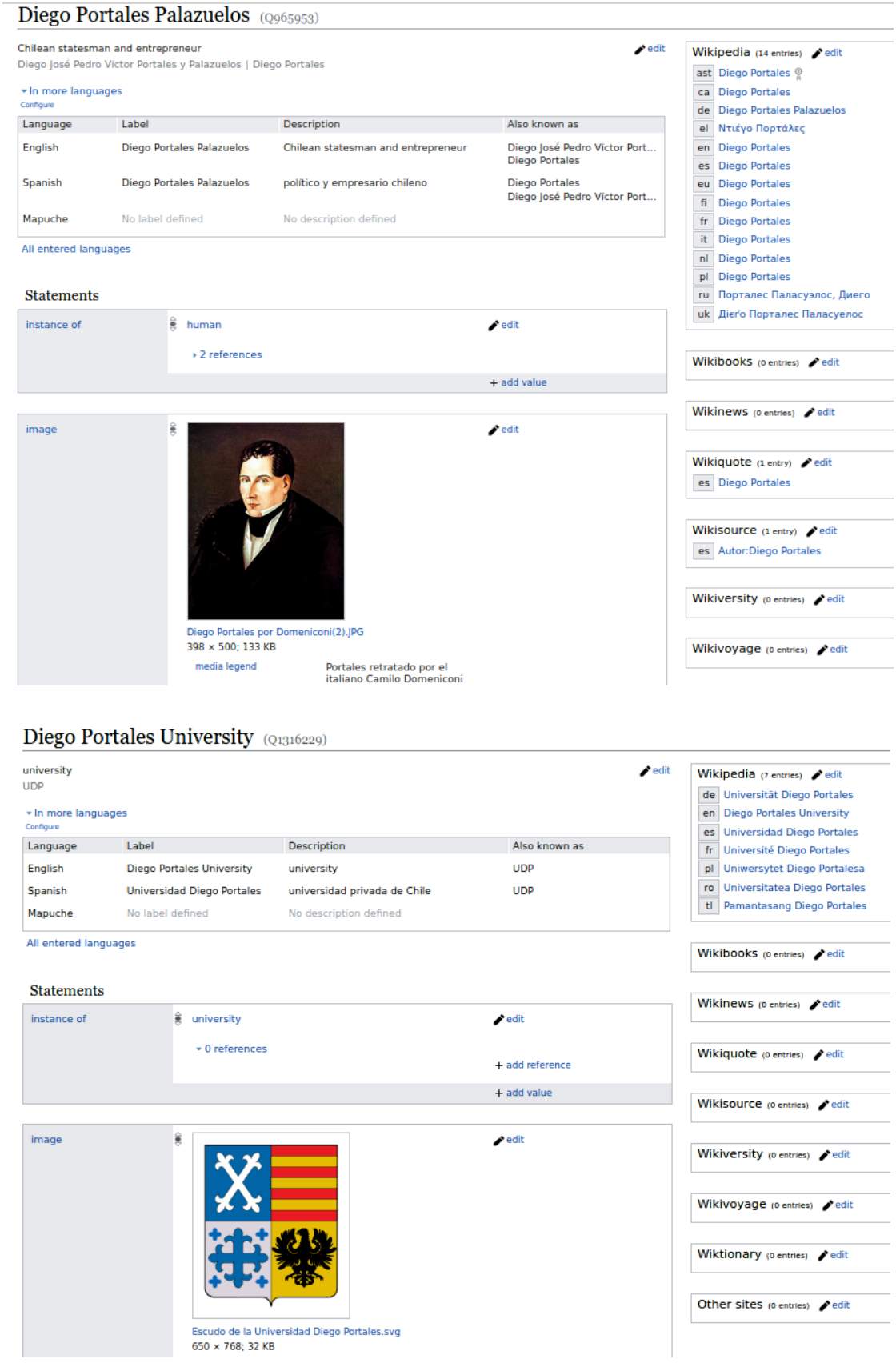


Figura 3.1: Página del Político y la universidad Diego Portales en Wikidata

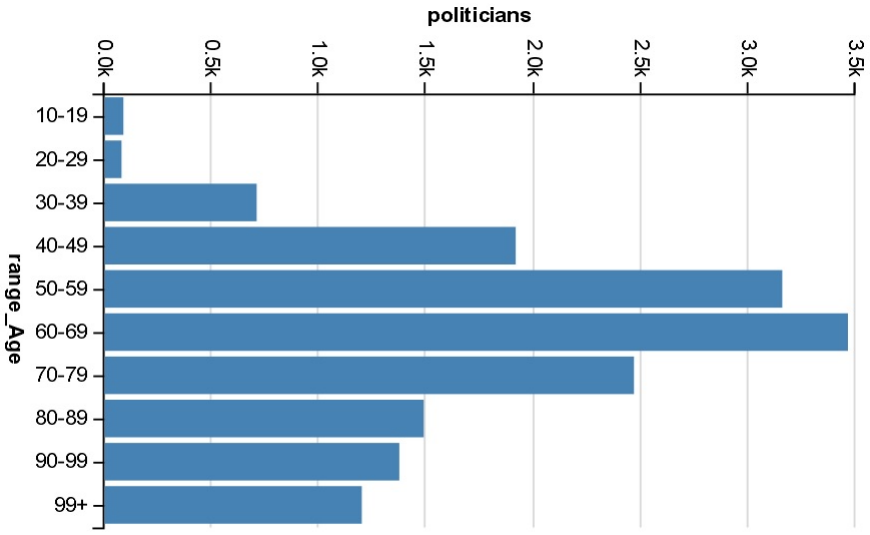


Figura 3.2: Distribución de edad de políticos en Wikidata

comparar la mención directa al usuario del político con la mención simple del nombre.

Luego de hacer la extracción se extrajeron 18.620 datos correspondientes a políticos latinoamericanos. No obstante, fue necesario definir una edad límite, ya que al hacer la distribución de políticos por edad, se comprobaba que un amplio número de estos (alrededor de tres mil políticos) sobrepasaba la edad de los cien años, como se puede apreciar en la Figura 3.2. Este detalle pudo deberse a que, a pesar de que muchos de estos políticos se encontrasen fallecidos, no existe esa información en la base de conocimientos.

Dado este percance en la muestra de los datos se decidió establecer una edad máxima, para así filtrar los datos. Dicha edad fue los 85 años, ya que es una cota mayor a la esperanza de vida de cualquier país en el mundo [22]. De esta manera se redujo la muestra a 11.398 políticos.

A modo de ejemplo los resultados de los datos se pueden encontrar en la Tabla 3.1

político	**país**	**género**	**partido**	**usuario**
Caio Koch-Weser	Brasil	m	PSA	
José María Figueres Olsen	Costa Rica	m	PLN	
Alberto Fujimori	Perú	m	NPP	AlbertoFujimori
Alberto Fujimori	Perú	m	Alianza por el futuro	AlbertoFujimori
...	...	...	...	...

Tabla 3.1: Ejemplo de tabla de datos extraída de Wikidata

Por último es importante señalar que de los más de 11 mil datos, una muestra no mayor a mil políticos latinoamericanos constaba con una cuenta de Twitter asociada a la base de conocimientos. Este dato es muy útil para el trabajo puesto que se ocupó en una de las etapas de la filtración de los datos.

3.2. Extracción y filtración de tweets

Con respecto a la extracción de los tweets ocupados en el presente trabajo, se utilizó una base de datos con tweets previamente extraídos por el Instituto Milenio de Fundamentos de los datos, por medio de la API de Twitter. La base de datos del IMFD constaba con el 1 % de los tweets emitidos a escala global de múltiples días que van desde el 27 de junio de 2017 hasta el 28 de diciembre de 2017. También se contó con los datos del período comprendido entre el 2 de diciembre del año 2018 y el 3 de marzo del año 2019.

Cada día correspondía a 8 tablas con información relevante a los tweets extraídos ,los usuarios que emitían dichos tweets e incluso un análisis de sentimiento hecho a los tweets.

En total se utilizaron tablas correspondientes a 186 días, generando un aproximado de 75 millones de tweets en español.

3.2.1. Extracción de tweets de la base de datos

Las tablas de días que se ocuparon para el trabajo de esta memoria, fueron desde el 28 de junio del año 2017 hasta el 27 de diciembre del mismo año, completando un total de 186 días. Si bien estos no eran todos los datos disponibles, fue este rango de fecha el único considerado principalmente por dos razones:

- Se trataba de un rango continuo de 6 meses, lo cual facilitaba el examinar el cambio de los datos a través del tiempo.
- El resto de los datos disponibles correspondían a diciembre del 2018 hasta el 3 de marzo del 2019, lo cual al mezclarse con datos del 2017 podrían entorpecer la muestra, ya que en muchos de los políticos de 2017 tenían otros cargos y funciones a finales de 2018 y principios de 2019.

Cabe destacar que, dada la magnitud de las tablas, las cuales eran de cientos o miles de millones de datos por cada día, se decidió inicialmente extraer solamente los tweets latinoamericanos, esto con la intención de trabajar con el menor número de tweets irrelevantes lo más pronto posible. Si bien esto era virtualmente posible ya que la tabla de tweets contaba con múltiples atributos que determinaban el lugar de procedencia, como por ejemplo el nombre del país, muchos de esos atributos se encontraban vacíos. Dadas estas circunstancias se seleccionó otro atributo de la tabla que estaba constantemente en los datos de la tabla para hacer el filtro previo a la extracción: el lenguaje del tweet.

Se decidió trabajar netamente con tweets en español y no el conjunto de español, portugués y francés, que son los principales idiomas que se hablan en latinoamérica y que aparecen dentro del conjunto de nacionalidades de los políticos extraídos de la base de conocimientos. Esta decisión se hizo principalmente por tres razones:

- La gran mayoría de países en latinoamérica son de habla hispana.
- La selección de tweets en español evita la presencia de muchos otros países ajenos a latinoamérica. Se tiene noción de que los datos van a contar con algunos tweets de españa, que lamentablemente no van a ser filtrados, no obstante luego pasará por un proceso de filtrado para seleccionar aquellos tweets que mencionen políticos latinoame-

ricanos, y con esto se espera presumir que la base de datos corresponderá a tweets en la mayoría de usuarios latinoamericanos.
- Las entidades que identificará el software encargado de hacer Entity Linking, van a estar directamente relacionadas al lenguaje en el cual se setea.

En total se extrajeron múltiples tablas de datos, las cuales fueron combinadas en una sola tabla de más de 9Gb de tamaño, contando con más de 75 millones de tweets con información de correspondiente al tweet, incluido el texto del mismo.

Por último es importante señalar que las tablas de usuarios y de análisis de sentimiento de los tweets se utilizaron posteriormente en el trabajo, así que también ocurrió una extracción de estas; una para filtrar los datos emitidos por los usuarios de Twitter que estaban asociados a algún político, y la otra para extraer el análisis de sentimientos de los tweets filtrados respectivamente.

3.2.2. Filtración de tweets

Una vez extraída la tabla de datos, se utilizó la librería pandas de Python para generar un proceso de filtrado en los datos; este proceso de filtrado corresponde a la eliminación de todos aquellos tweets (y sus datos asociados) que no hablen o no sean de políticos latinoamericanos. Para ello el proceso se subdividió en tres partes:

- Filtración según mención directa.
- Filtración según mención indirecta.
- Tweets emitidos por los políticos.

Es importante señalar que cada una de estas partes se hizo de manera independiente.

Filtración de datos para encontrar menciones directas

Después de haber extraídos los datos se decidió generar el filtro por menciones. Para este proceso se requirieron los datos de los usuarios de Twitter de los políticos, como la tabla recién creada.

Nombre	**Usuario**
Amazonino Mendes	AmazoninoAM
Michaëlle Jean	MichaelleJeanF
Alberto Fujimori	AlbertoFujimori
...	...

Tabla 3.2: Tabla de Políticos Latinoamericanos con su nombre de Usuario

Para esta parte del trabajo lo primero que se hizo fue un proceso de filtrado ocupando la librería pandas de Python. Con dicha librería se seleccionaron todos aquellos tweets que contacen por lo menos con un carácter "@". Luego de esto se seleccionaron todos aquellos tweets que tuviesen al menos uno de los usuarios de la tabla de políticos latinoamericanos (Tabla 3.2), almacenando en una columna extra el o los políticos que hubiesen sido mencionados en el tweet.

Filtración de datos para encontrar menciones indirectas

En el caso de las menciones indirectas se decidió ocupar Entity Linking. En primera instancia se tenía la intención de usar la herramienta Babelfy para la detección de identidades, debido a su buen rendimiento en estudios ya comprobados con datos en múltiples lenguajes [23]. Pero debido a que la herramienta permite un número limitado de consultas diarias y para el desempeño de este trabajo se requiere hacer una consulta or tweet, generando de esta manera más de 75 millones de consultas, se prefirió optar por la alternativa siguiente en rendimiento, DBpedia Spotlight, la cual trabaja con entidades de la base de conocimientos de DBpedia, es Open Source, y tiene una librería para trabajar directamente en Python [24].

Se configuró DBpedia Spotlight en idioma español con los parámetros support en 20 y confidence en 0.4. DBpedia Spotlight no sugiere ningún valor para los parámetros ya mencionados; en la librería de Python ambos parámetros tienen por valor inicial 0. No obstante, se decidió ocupar los valores ya mencionados debido a resultados obtenidos por experimentos iniciales.

Es importante mencionar que mientras se hacían los experimentos iniciales del trabajo, para evaluar el funcionamiento de DBpedia Soptlight, se identificó que la herramienta no era capaz de detectar las entidades de personas si estas no se encontraban escritas con mayúscula inicialmente. Dado esto, se tuvo que hacer un preprocesamiento de los tweets.

Tomando un diccionario de palabras en español, se creó un programa que leyese cada palabra del tweet, evaluará si dicha palabra se encontraba en el diccionario y en caso de no encontrarse, cambiase la palabra por la misma, solamente que esta vez con la inicial en mayúscula. Esto permitió que apellidos y nombres que no se encontraban en el diccionario y que originalmente se encontraban escrito con letra minúscula, fueran evaluados como posibles entidades de personas. De esta manera tweets como “me encanta el spaggetti” o “que grande maradona”, retornarían “me encanta el Spaggetti” y “que grande Maradona” respectivamente.

Luego de hacer la configuración se almacenaron los tweets en los cuales DBpedia Spotlight consideraba que hubiesen personas en él; esto se pudo hacer gracias a que DBpedia Spotlight entrega una serie de atributos de la entidad, en formato JSON, como se puede apreciar en la Figura 3.3.

Además de esto se añadían los datos entregados por DBpedia Spotlight, datos con los cuales se pudo hacer el nexo con las tablas de datos extraídas previamente, ya que contaban con un link a la entidad de DBpedia.

Al trabajar DBpedia Spotlight con los datos de DBpedia y haber trabajado hasta el momento con solamente datos extraídos de Wikidata, se requirió hacer una conexión de ambas partes. Dicha conexión no resultó ser tan difícil, esto gracias a que tanto DBpedia como Wikidata extraen información de Wikipedia. Esto generaba que, el link al cual hacían referencia los resultados de DBpedia Spotlight fuesen de tipo “http://dbpedia.org/page/nombre_apellido” mientras que la entidad extraida de Wikidata correspondía al mismo Nombre y apellido. Como por ejemplo si el nombre en wikidata correspondía a “Daniel Jadue”, al ser detectado en DBpedia Spotlight la entidad entregada era “http://es.dbpedia.org/resource/Daniel_Jadue”.

```
{
  "@text": "Piñera pesao",
  "@confidence": "0.4",
  "@support": "20",
  "@types": "",
  "@sparql": "",
  "@policy": "whitelist",
  "Resources": [
    {
      "@URI": "http://es.dbpedia.org/resource/Sebastián_Piñera",
      "@support": "1455",
      "@types": "Http://xmlns.com/foaf/0.1/Person,Wikidata:Q5
      ,Wikidata:Q24229398,Wikidata:Q215627,
      DUL:NaturalPerson,DUL:Agent,Schema:Person,
      DBpedia:Person,DBpedia:Agent",
      "@surfaceForm": "Piñera",
      "@offset": "0",
      "@similarityScore": "0.9902596096163403",
      "@percentageOfSecondRank": "0.008890256620676714"
    }
  ]
}
```


Figura 3.3: Ejemplo uso de DBpedia Spotlight en texto:“Piñera pesao”

Considerando esto, se leyeron cada uno de los links entregados por DBpedia Spotlight que hacían mención a personas. Ya con este conjunto de datos, las referencias a personas se filtraron por sólo aquellas que fuesen menciones directas a la muestra de políticos extraída previamente.

Selección de tweets emitidos por políticos

Por último se filtraron de la base de datos de tweets en español, todos aquellos tweets que tuviesen por usuario alguno de los 1024 políticos que figuraban con cuenta de Twitter en los datos extraídos en Wikidata.

Si bien se tenían los datos del nombre de usuario en Twitter, fue necesario primero extraer de las tablas de usuarios ya mencionadas, los ids de los usuarios en cuestión. Se requirió hacer esto, puesto que en la base de datos de tweets se contaba sólo con esta información.

Cabe señalar que al ser una extracción del 1 % de los tweets mundiales, y luego ser filtrados en español, no todos los usuarios de políticos latinoamericanos que se encontraban en la tabla de datos, aparecían emitiendo algún tweet. Además al ser sólo tweets en español, muchos de los políticos de Brasil, Haití y otros países latinos con otra lengua predominante, no figuraron entre los usuarios encontrados. Una vez teniendo los ids de usuario de los políticos latinoamericanos, se filtro la tabla de datos de tweets según este parámetro.

Finalmente las tres tablas resultantes se agruparon netamente en una gran tabla con los datos de del id del tweet, id del usuario, políticos a los que menciona directamente, políticos a los que menciona indirectamente, fecha y nombre del político que emitió el tweet de haberlo hecho.

3.3. Extracción tabla de análisis de sentimiento

Luego de generar una base con los datos ya descritos se procedió a descargar de la tabla de análisis de sentimiento de los tweets. Dicha tabla constaba con una fila correspondiente al id del tweet, y las demás correspondiente a los valores ya evaluados por el IMFD usando la librería Sentistrength de Java. Dicho trabajo almacenó valores neutros, positivos y negativos encontrados en el tweet como también un valor llamado polaridad, el cual corresponde a la suma de los valores positivos y negativos.

Una vez extraída de las diferentes tablas, todos los datos de sentimiento de los tweets, se procedió a trabajar por separado con las tablas de tweets emitidos por políticos, tweets que mencionan a políticos directamente y tweets que mencionan a políticos indirectamente. Esta decisión se llevó a cabo debido a que las menciones directas e indirectas corresponden a columnas distintas con una lista de posibles políticos y para esta parte del trabajo se prefirió dividir de tal manera que cada dato tuviese la mención de solamente un político.

Por último se anexaron los datos del político correspondiente a cada dato de las menciones directas, indirectas y emisiones.

Capítulo 4

Características de los datos

En el siguiente capítulo se profundizará en mayor medida sobre las características principales de los datos obtenidos una vez filtrados la base de datos de tweets, como también los metadatos extraídos de Wikidata.

4.1. Datos extraídos de Wikidata

Los primeros gráficos generados corresponden al detalle de los datos extraídos de Wikidata, generando una vista general de lo que vendría siendo los datos de los políticos en cuestión.

4.1.1. Distribución por genéro

Con respecto a la distribución de género, como se puede apreciar en la Figura 4.1, la mayoría de los políticos de la base de conocimiento pertenecen al género másculino, siendo estos 9.094 datos, los cuales corrsponden al 78.5 % de la muestra. El género femenino posee 2.299 datos de la muestra, lo cual corresponde al 19.8 % de los datos. Ádemas existen 5 políticos que corresponden a mujeres transgénero. Es importante destacar que, al ser el género un atributo opcional, 178 políticos de la base de conocimientos no poseen un género determinado y por lo tanto se decidió omitirlos en el gráficos.

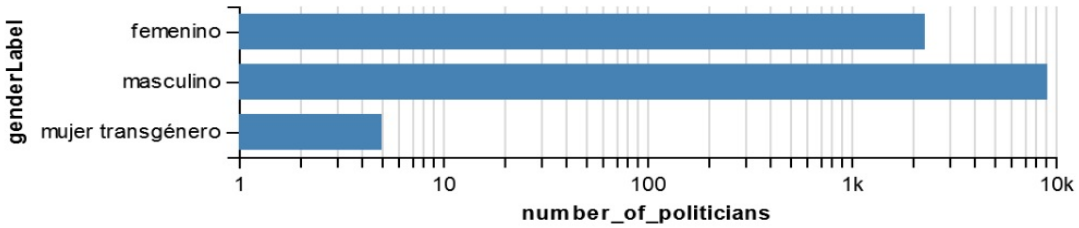


Figura 4.1: Cantidad de políticos latinoamericanos por género en escala logarítmica

Por otra parte, como se evidencia en la Figura 4.2, el promedio de edad de los políticos latinoamericanos que se encuentran en la base de conocimientos es, tanto en hombres como mujeres, superior a los 55 años, siendo 57 en caso de las mujeres y 61 en el de los hombres.

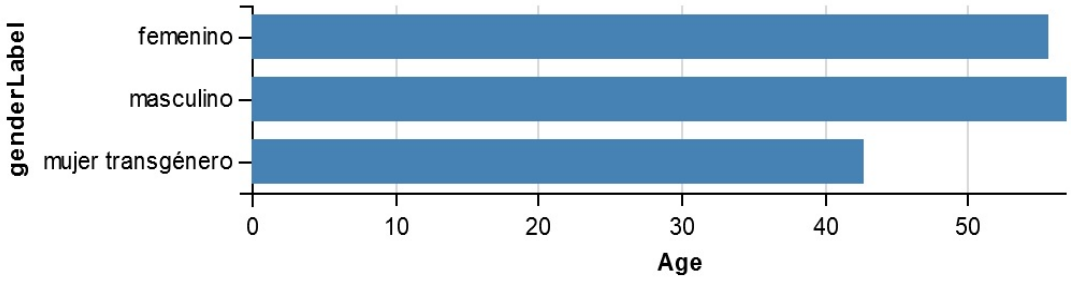


Figura 4.2: Promedio de edad de políticos latinoamericanos por genéro

La proporcionalidad latinoamericana que hay entre los géneros muestra una mayor superioridad de presencia masculina; esto también se ve reflejado en cada país de manera individual, tal como se puede apreciar en la Figura 4.3, la cual muestra en mayor detalle ambas distribuciones de género para cada país.

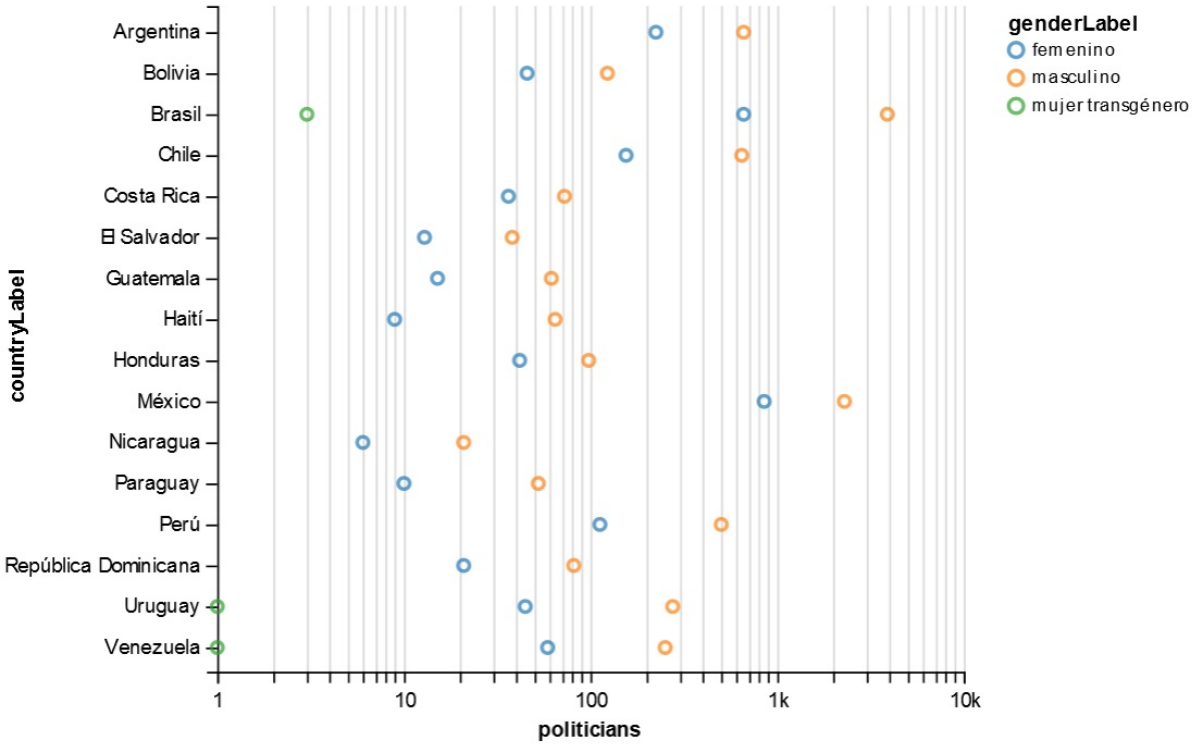


Figura 4.3: Cantidad de políticos latinoamericanos por género en escala logarítmica de cada país

De la misma manera, se comparó el promedio de edad de los políticos por país y género, evidenciando que en la mayoría de estos el promedio de edad que poseen las mujeres es menor al de los hombres exceptuando el caso de República Dominicana, como se puede apreciar en la Figura 4.4.

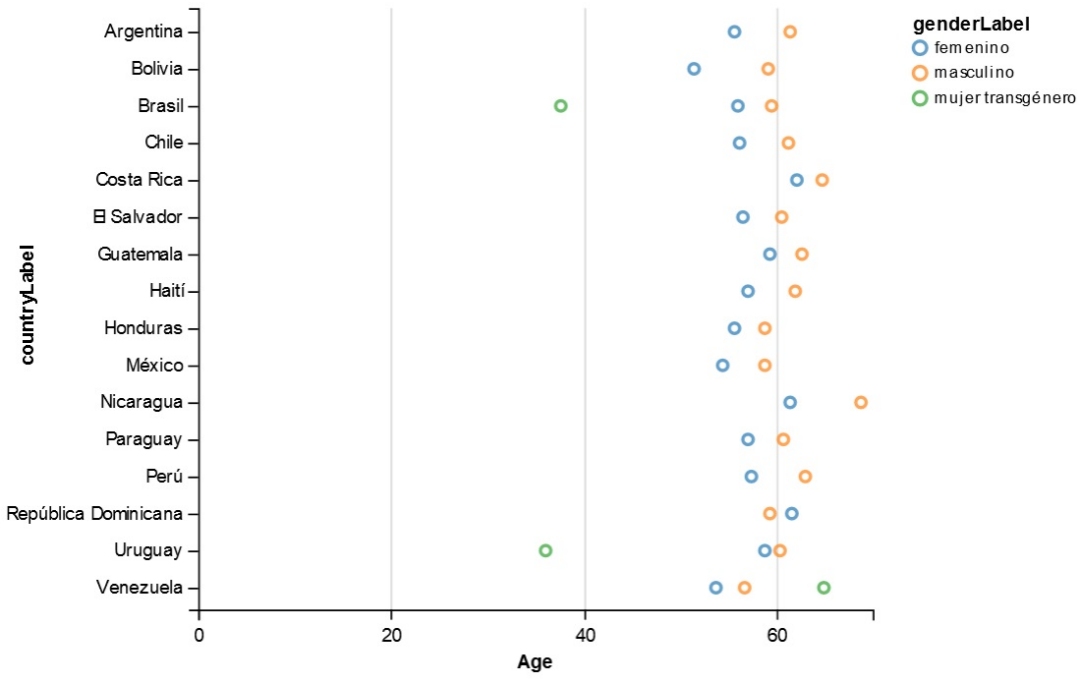


Figura 4.4: Promedio de edad de políticos latinoamericanos por país y género

Políticos con usuarios de Twitter disponibles

Del conjunto de políticos latinoamericanos que tienen asociado dentro de la base de conocimiento un usuario en Twitter, la distribución de estos con respecto a su género fue la que se puede apreciar en la Tabla 4.1, la que evidencia que de los 11.398 datos sólo el 10,4 % tiene un usuario de Twitter asociado .

Género	Número de políticos	Número de usuarios	Porcentaje[%]
Masculino	9.094	856	9,4
Femenino	2.299	338	14,7
Mujer transgénero	5	1	20

Tabla 4.1: Distribución por género de políticos de Twitter

4.1.2. Representatividad de cada país

Se decidió generar resultados que contrapusieran el número de políticos de cada país con la población de los mismos, esto con la intención de tener una idea de los países que se encontraban mejor representados en la muestra obtenida. Para esto se calculó la división entre el número de políticos de un país y su población; luego este número se multiplicó por un millón para dar el valor de representatividad que se muestra.

Dicho proceso se hizo tanto por género como general, dando por resultado a Uruguay como el país con mayor representatividad en este aspecto y Nicaragua el país de menor representatividad. Estos datos se pueden apreciar en la Figura 4.5.

Se puede ver la amplia representación de Uruguay por parte de sus políticos. Siendo la representatividad de las mujeres mayor a la representatividad general de cinco países (El

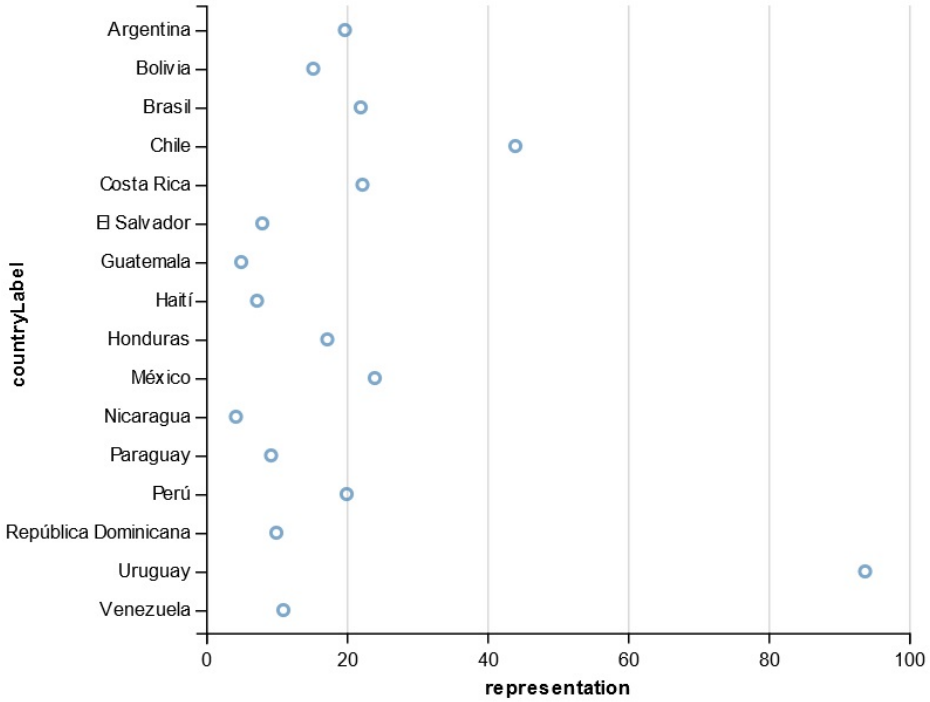


Figura 4.5: Representación de cada país

Salvador, Guatemala, Haití, Paraguay y República Dominicana). La representatividad tan alta de Uruguay en comparación a los otros países puede ser gracias a que este país solo consta con 3.456.750 habitantes según Wikidata [25], siendo el país con menos población de toda la muestra, seguido por Costa Rica.

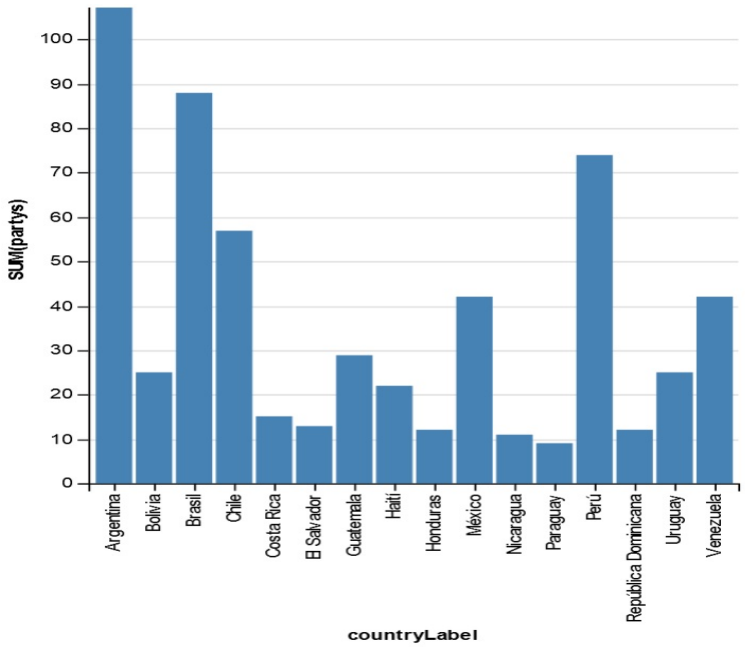


Figura 4.6: Número de partidos políticos por país

4.1.3. Partidos políticos por país

Se hizo el cálculo del número de partidos que hay en cada país, siendo Argentina el país con mayor cantidad de partidos políticos distintos, con 110 de estos, y Paraguay el país con menor cantidad, contando solamente con 9 en la base de conocimientos. Dicho resultado se puede apreciar en la Figura 4.6.

4.2. Características generales base de datos resultante

De la filtración de la tabla general de datos, se conformaron tres tablas de datos con los filtros ya mencionados. El número de datos totales filtrados fue 1.845.420, de los cuales se pueden subdividir en los siguientes conjuntos:

- **Menciones directas:** 967.982 datos. De los 1.197 políticos con usuario anexado a Wikidata se mencionaron 702 políticos.
- **Menciones indirectas:** 834.694 datos. De los más de 11.348 políticos de la muestra extraída de Wikidata solo 1.063 fueron encontrados con una mención indirecta. Cabe señalar que se leyeron sobre 200 muestras del resultado entregado, obteniendo una precisión cercana al 93 %.
- **Tweets emitidos por políticos:** 87.549 datos. La muestra contaba con 394 políticos distintos que emitian tweets de los 1.024 extraídos utilizando Wikidata.

Dentro de los tweets de la muestra existieron algunos casos que tanto eran mencionados por políticos y hacian menciones directa y/o indirectamente; la magnitud de este tipo de conjuntos se puede apreciar en la Figura 4.7.

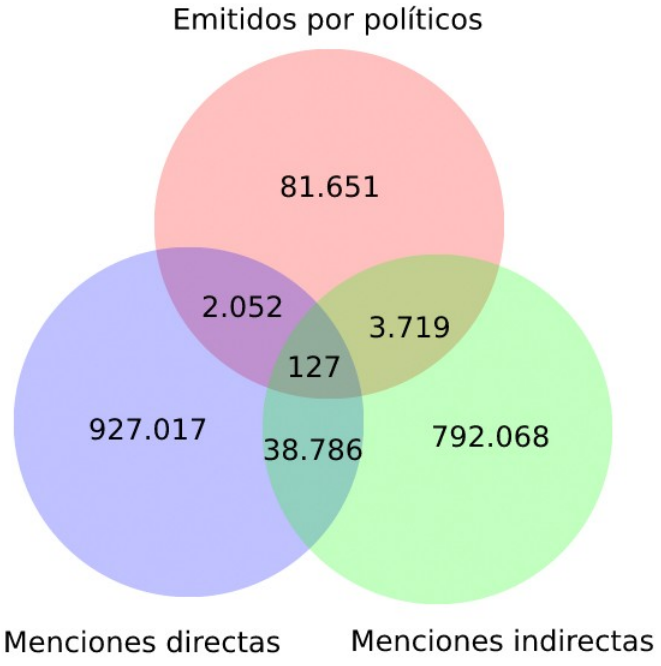


Figura 4.7: Subconjuntos tabla de datos resultante

4.2.1. Distribución menciones directas

- **Género**: De manera general las menciones directas a los políticos latinoamericanos se distribuyeron en alrededor de 90 % para los políticos del género masculino y 10 % para el género femenino. Además hubieron 105 menciones a una política venezolana que es clasificada como mujer transgénero según la base de conocimientos (Tamara Adrian). Un mayor detalle sobre la distribución se puede apreciar en la Figura 4.8.

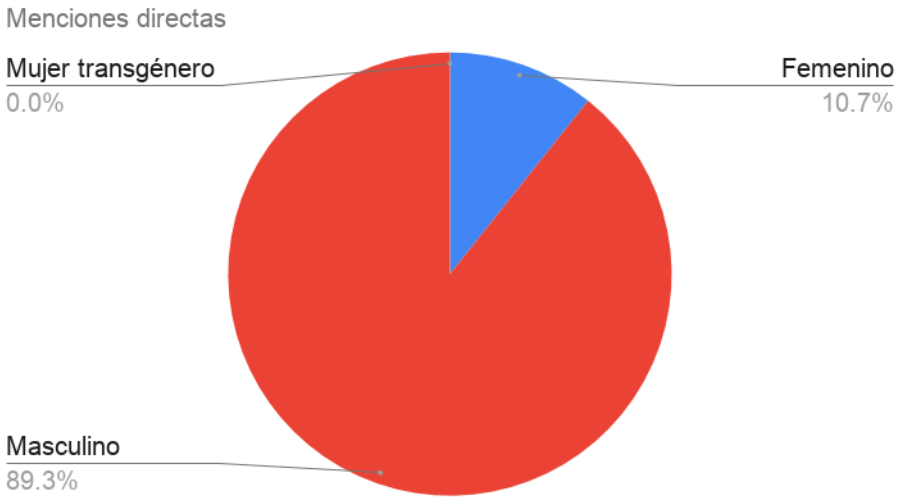


Figura 4.8: Distribución menciones directas por género

- **País**: Por otra parte la distribución por país muestra a Venezuela como el país mayormente mencionado de todos los países latinoamericanos.

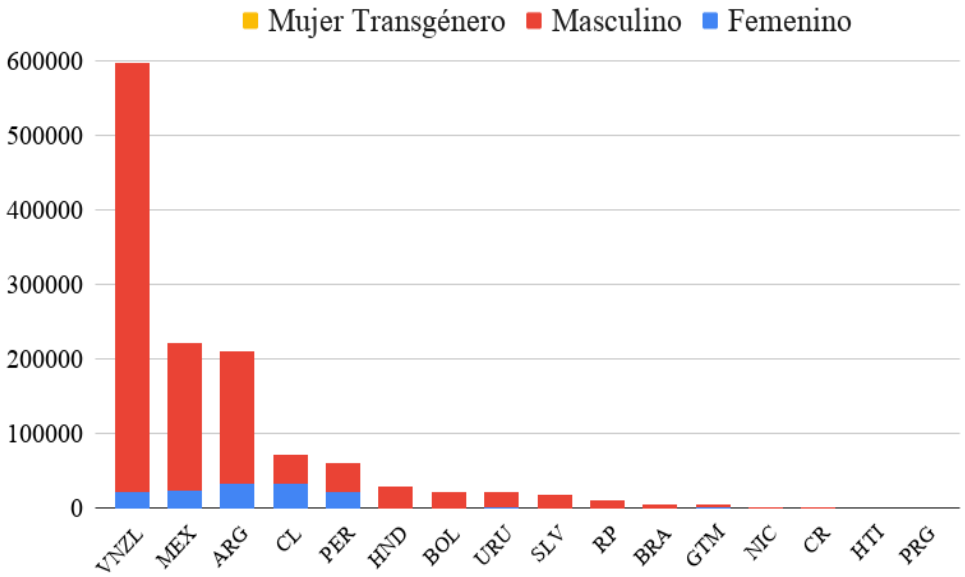


Figura 4.9: Distribución menciones directas por país

- **Distribución de género por país**: El gráfico de esta distribución corresponde al de la Figura 4.10. Si bien el caso de Paraguay es inusual pero al ser solo 40 datos estos podrían vincularse al retweet de alguna noticia.

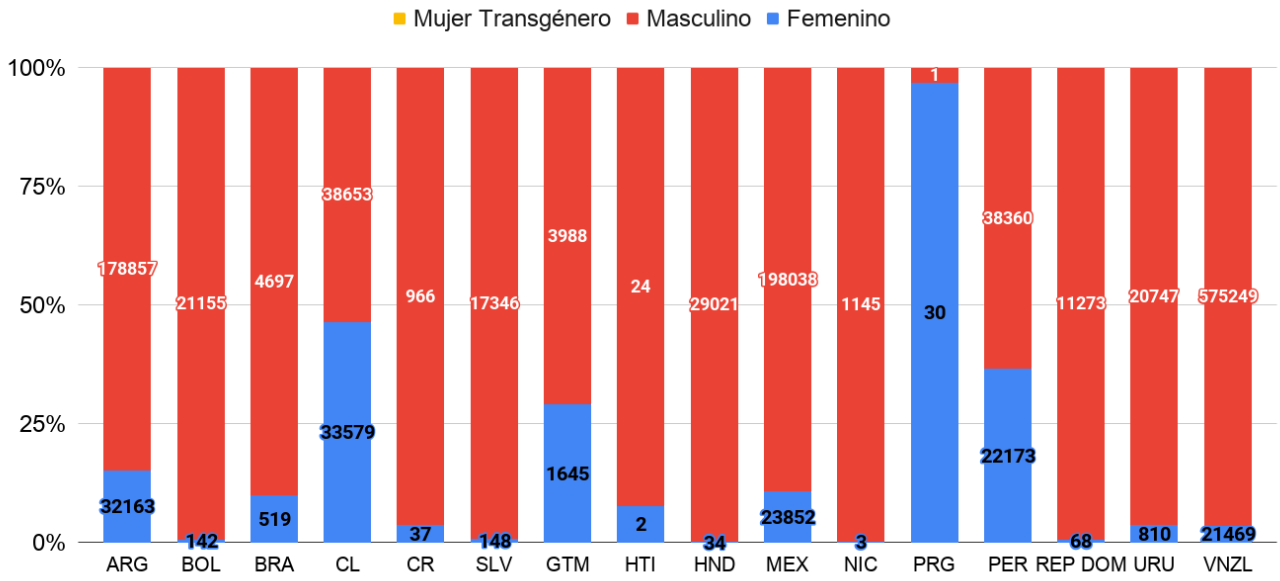


Figura 4.10: Distribución menciones directas por género y país

4.2.2. Distribución menciones indirectas

- **Género**: Alrededor del 12 % de las menciones indirectas corresponden a políticos de género femenino; el porcentaje restante lo abarca el género masculino tal como se puede apreciar en la Figura 4.11.

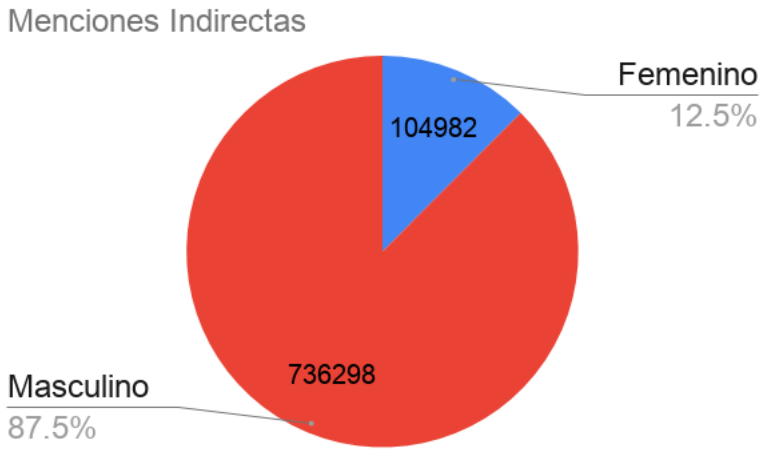


Figura 4.11: Distribución por género, menciones indirectas

- **País**: En el caso de la distribución por país, venezuela es el mayormente mencionado.

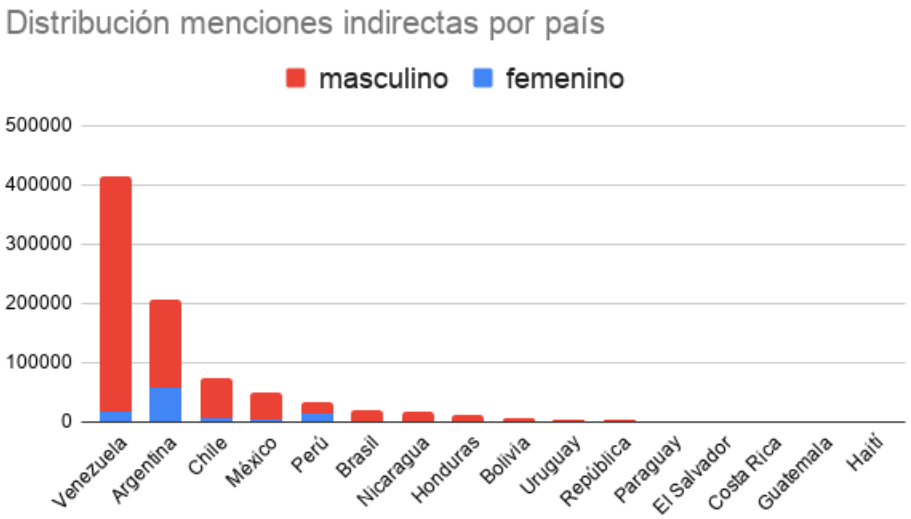


Figura 4.12: Distribución menciones indirectas por país

- **Distribución de género por país**: De igual manera que con las menciones directas, se agruparon los datos tanto por género y por país, mostrando a su vez la distribución porcentual de la misma. En este gráfico destaca Guatemala al ser el único país con un mayor número de menciones indirectas hacia mujeres.

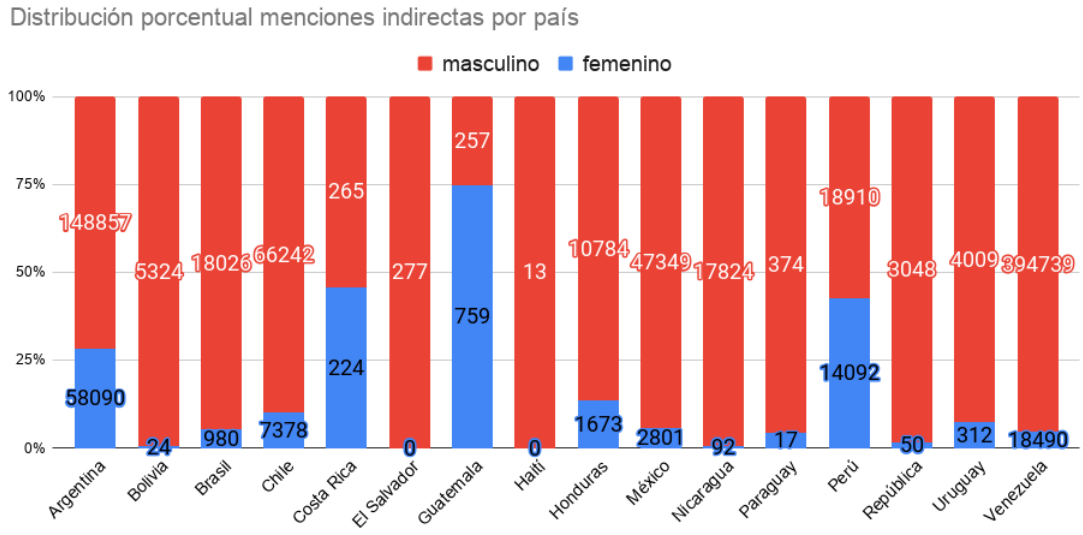


Figura 4.13: Distribución porcentual menciones indirectas por país

4.3. Curva de Lorenz

Con la intención de detectar la distribución de las cantidades totales de la base de datos entre los políticos se decidió aplicar tres curvas de Lorenz [26]. Consecuentemente con el trabajo ya empleado se implementaron curvas correspondientes tanto a las menciones directas como indirectas; sin embargo en este caso se decidió agregar una curva de Lorenz que refleje la distribución de los tweets emitidos por políticos.

Las curvas de Lorenz de los tweets emitidos por políticos, las menciones directas a políticos y las menciones indirectas corresponden a las Figuras 5.14, 5.15 y 5.16 respectivamente. La línea azul de las curvas representa la homogeneidad en la distribución, mientras que la línea roja es la distribución real.

La curva de Lorenz es una forma de representar la distribución de los resultados obtenidos. Por lo tanto, si por ejemplo la línea roja se encuentra en el punto (30,3) esto quiere decir que el 30 % de los usuarios de Twitter emitieron 3 % de los tweets de la muestra.

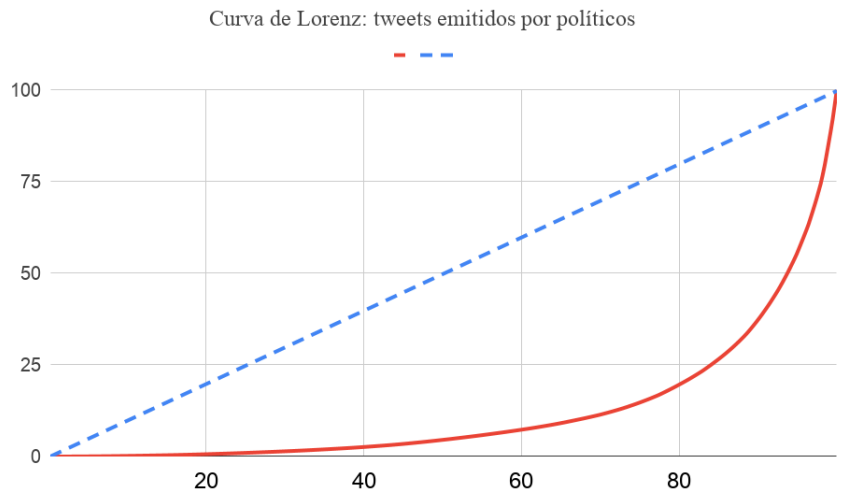


Figura 4.14: Curva de Lorenz de tweets emitidos por políticos

Es importante destacar que las curvas de Lorenz se hicieron entre los políticos que eran mencionados al menos una vez en los tweets; siendo 387 políticos distintos en el caso de los tweets emitidos por los mismos políticos, 1.058 en el caso de los políticos mencionados indirectamente y 669 en el caso de los políticos mencionados directamente.

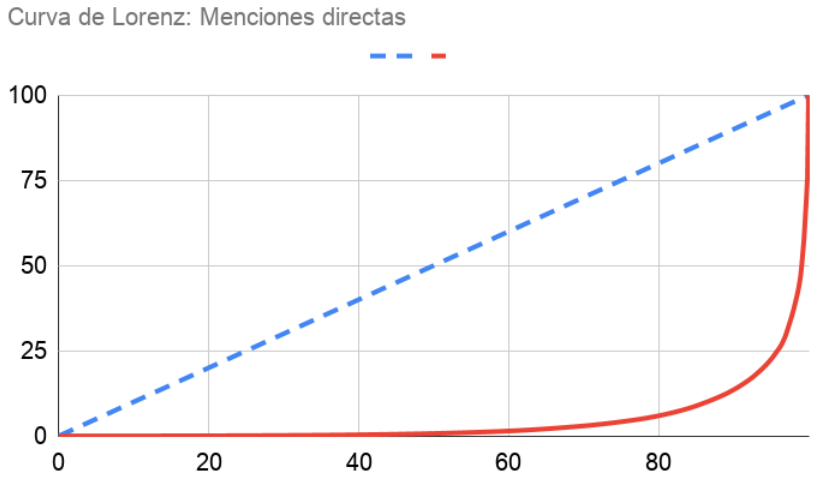


Figura 4.15: Curva de Lorenz de menciones directas

Figura 4.16: Curva de Lorenz de menciones indirectas

Al hacer las diferentes curvas de Lorenz se pudo apreciar que en las menciones directas el 96 % de los políticos son mencionados 25 % de las veces, esto significa que de los 669 políticos que fueron mencionados directamente 34 de ellos fueron mencionados 725.987 veces de las 967.982 total. El caso de las menciones indirectas es más drástico aun, llegando a ser el 2 % de los políticos más mencionados los que se distribuyen el 75 % de las menciones. Esto quiere decir que hay 22 políticos que son mencionados al menos 626.025 veces según la muestra.

Con respecto a los tweets emitidos, si bien la curva es menos pronunciada, la distribución de los tweets emitidos por los políticos sigue siendo más equitativa aglomera el 75 % de los tweets entre el 16 % de los políticos que más usan la plataforma.

4.4. Usuarios frecuentes en la muestra

De la base de datos creada a partir de los tweets que mencionan a políticos latinoamericanos se determinó que los 1.845.420 tweets eran emitidos por no más de 512.000 usuarios donde la gran mayoría de estos (por sobre el 75 % de los usuarios de la muestra) no aporta con más de 1 tweet por usuario. De hecho el 90 % de toda la muestra tiene como máximo 3 tweets por usuario y el 99 % de los usuarios tiene como máximo 37 tweets por usuario. No obstante, dentro de este 1 %, específicamente hablando del 0,1 % más activo (512 usuarios), se logra llegar a usuarios con cientos e incluso miles de tweets emitidos.

La Figura 4.17 muestra el número de tweets por usuario contando solo este 1 % mientras que la tabla 5.2 muestra el nombre y número de tweets emitidos en la muestra de los 10 usuarios que más publicaron.

Es importante señalar que, si bien el top 10 de usuarios de Twitter de la muestra que más publican con respecto a políticos latinoamericanos corresponden a plataformas web de medios de prensa, dentro del 0,1 % que más publica también se encontraron personas naturales, estando incluso entre los 100 primeros.

Usuario	**Número de tweets emitidos**
@DolarToday	6300
@AlbertoRodNews	5647
@ElNacionalWeb	5374
@NTN24ve	4360
@eldestapeweb	3792
@presidencialVen	3777
@ConElMazoDando	3628
@VTVcanal8	3290
@CaraotaDigital	3250

Tabla 4.2: Top 10 usuarios de Twitter que más hablan sobre políticos latinoamericanos según la muestra

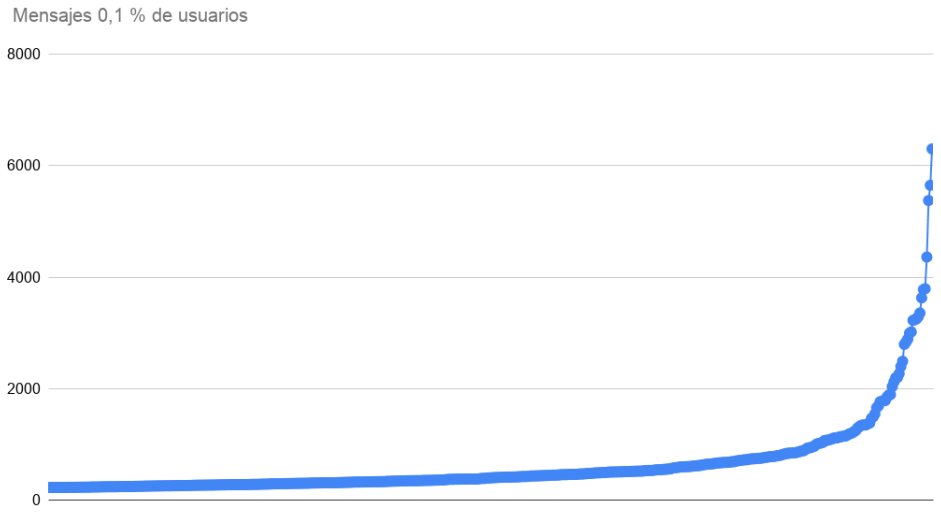


Figura 4.17: Número de mensajes emitidos por el 0,1 % de los usuarios más activos

Capítulo 5

Análisis de sentimiento

Los principales trabajos hechos con la muestra de datos fueron dados por la unión de los tweets con el análisis de sentimiento preexistente. Como una prueba de concepto, el siguiente capítulo presentará los resultados de un análisis de sentimiento, esto fue hecho ya que el objetivo del trabajo en sí mismo es extraer datos sobre políticos latinoamericanos que pueden ser usados en el contexto de estudios por expertos de las ciencias políticas y sociales.

5.1. Características análisis de sentimiento

Con respecto a las características de la muestra de datos asociada a los sentimientos, se hicieron distribuciones similares. El análisis de sentimientos entregaba tanto valores para alusiones negativas, como positivas o neutras, por separado dentro del mismo tweet. Sin embargo para el estudio se consideraron solamente los parámetros de número de tweets y polaridad, el cual consistía en la suma de los valores tanto negativos como positivos. Se hablará de polaridad promedio como el resultado de la polaridad dividida por el número de tweets emitidos por los usuarios.

5.1.1. Menciones directas en análisis de sentimiento

- **Género**: Evaluando por género se evidenció de manera general que aquellos usuarios mencionados directamente de género femenino tienen una polaridad menor a los masculinos, teniendo ambos una polaridad positiva. Como ya se mencionó, en las menciones directas sólo se contó con menciones hacia una política, por lo que el valor presentado en la Figura 5.1 sólo corresponde al promedio de la opinión hacia ella.
- **País**: En el caso de la distribución por países, la mayoría de estos tienden a tener en promedio positivo de polaridad salvo Uruguay y Guatemala, como se aprecia en la Figura 5.2.
- **País y género**: Por último se estableció una distribución de tanto país y género con respecto al promedio de la polaridad de sentimientos. Dicha distribución está ordenada de menor a mayor con respecto al promedio masculino y se puede apreciar en la Figura 5.3.

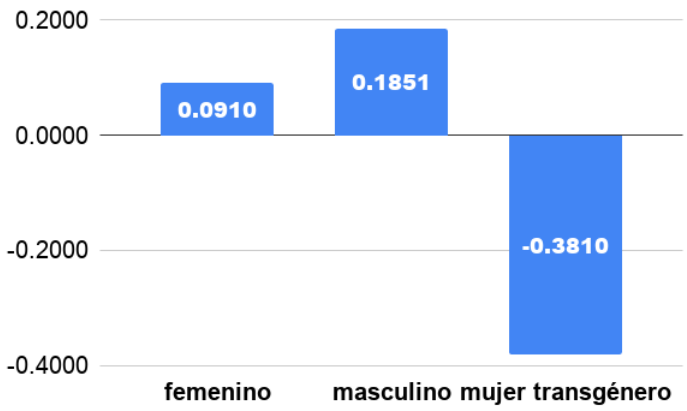


Figura 5.1: Promedio polaridad en menciones directas, diferenciadas por género

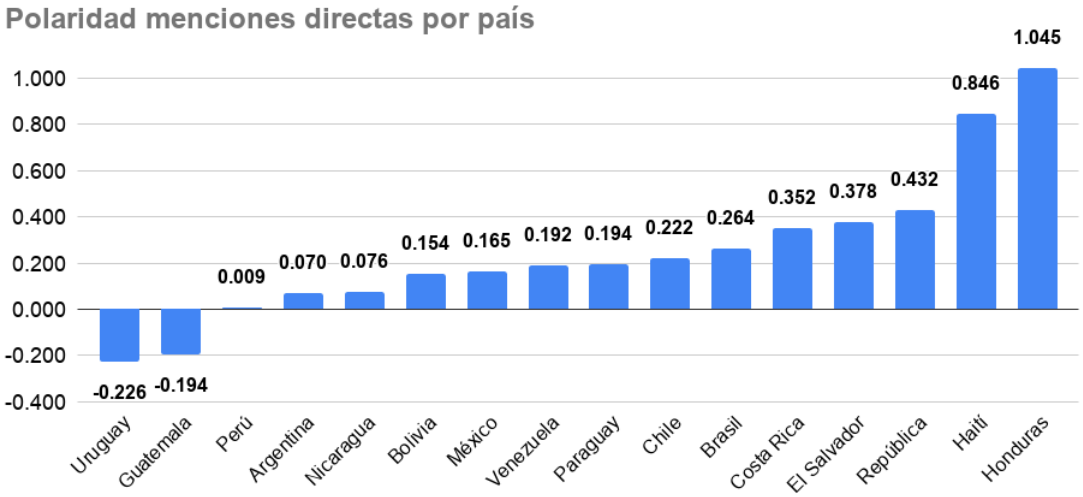
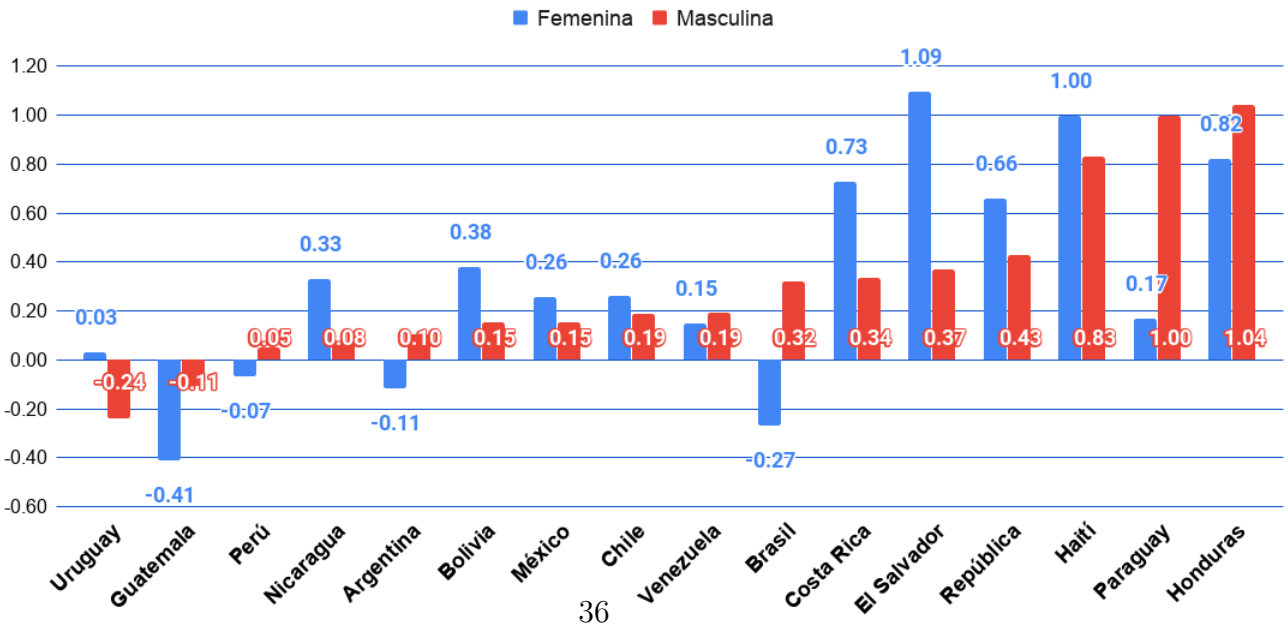


Figura 5.2: Promedio polaridad en menciones directas, diferenciadas por país



5.1.2. Sentiment analisis en menciones indirectas

Al igual que la contraparte directa se hicieron una serie de gráficos ilustrando tanto por país como género las distribuciones indirectas.

- **Género**: Al hacer el promedio de la polaridad separado netamente por género en las menciones indirectas, ambos resultados dan valores negativos tal como se aprecia en la Figura 5.4.

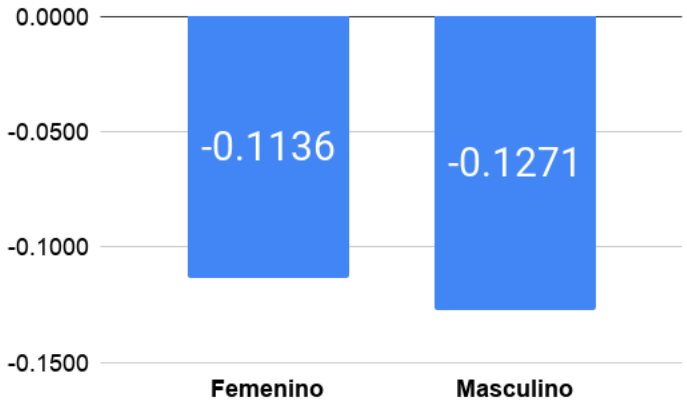


Figura 5.4: Promedio polaridad en menciones indirectas, agrupadas por género

- **País**: Por otra parte la representación del promedio de la polaridad de los políticos agrupados por países, entregó tanto resultados positivos como negativos. Dichos resultados se ordenaron de menor a mayor polaridad.

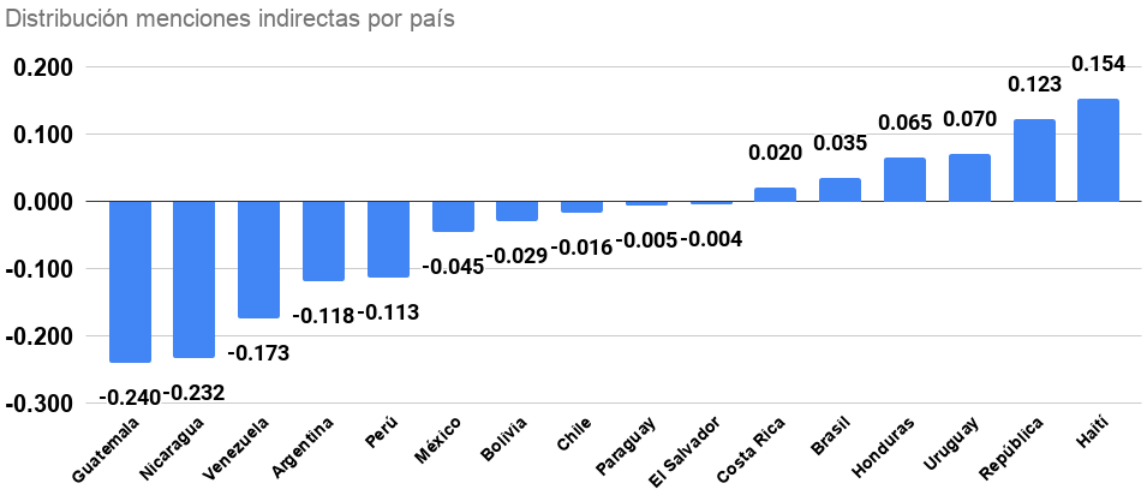


Figura 5.5: Promedio polaridad en menciones indirectas, agrupadas por país

- **Género y país**: Por último, al igual que en el caso anterior, se agruparon los resultados por país y género, para luego ser ordenados de menor a mayor con respecto a la promedio de polaridad masculino. El gráfico resultante corresponde a la Figura 5.5.

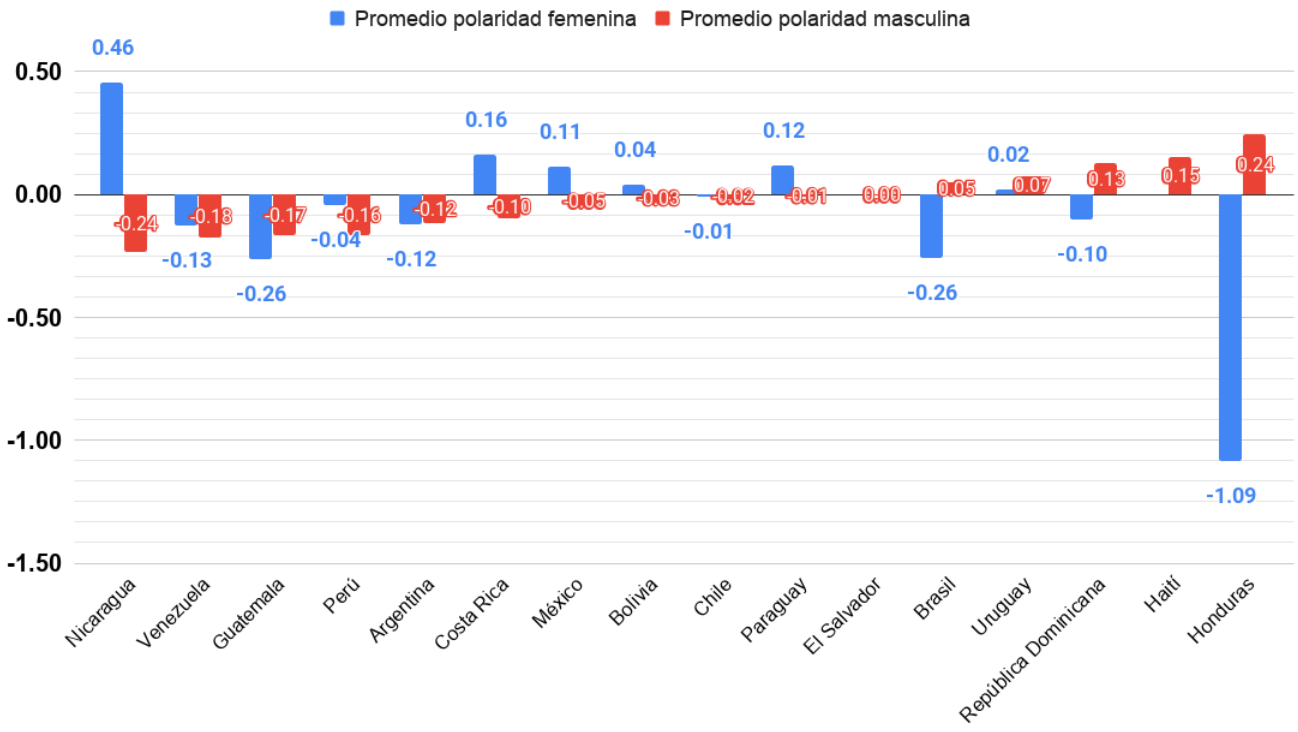


Figura 5.6: Promedio polaridad en menciones indirectas, agrupadas por país y género

Con respecto al promedio de análisis de sentimiento de la muestra, la división de esta en menciones directas e indirectas señalan que, en general, hay una polaridad promedio mayor en el caso de las menciones directas, siendo Honduras el país con polaridad promedio mayor en menciones directas, con un promedio 1,045 y Haití por parte de las menciones indirectas con un promedio de 0,154.

En el caso de Venezuela, el país más mencionado en ambos casos, el promedio de polaridad termina siendo 0,192 y $-0,173$ para los casos de menciones directas e indirectas.

5.2. Menciones entre políticos

Con la intención de mostrar las relaciones entre políticos, se decidió agrupar los tweets emitidos por políticos hacia los mismos.

5.2.1. Menciones agrupadas por países

Se agruparon los políticos emisores de los tweets por país, al igual que el político que hacían mención. De esta manera quedó una tabla que muestra de forma genérica la opinión de los políticos de un país por sobre otros. En la Tabla 5.1 se puede apreciar tanto la polaridad de las menciones directas como indirectas en conjunto con la suma de estas de aquellos países de la muestra en el que políticos hicieron referencia a otros políticos del mismo país.

País	Pol ind.	N.º ind.	Pol directa	N.º directo	Total	Polaridad
Argentina	-78	1331	150	392	1723	72
Bolivia	-2	30	27	26	56	25
Chile	-1	229	92	194	423	91
México	-1	138	258	261	399	257
Nicaragua	1	3			3	1
Perú	-143	655	5	286	941	-138
Rep. Dominicana	0	1	1	12	13	1
Venezuela	-150	1134	170	356	1490	20
El Salvador			2	1	1	2
Honduras			10	8	8	10
Uruguay			13	16	16	13

Tabla 5.1: Análisis de sentimiento entre políticos del mismo país

Impresión internacional

Además se generaron gráficos que muestran la opinión de los políticos sobre los países. En las Figuras 5.7 y 5.8 se pueden apreciar los mapas de calor de los países, considerando las menciones directas e indirectas respectivamente.

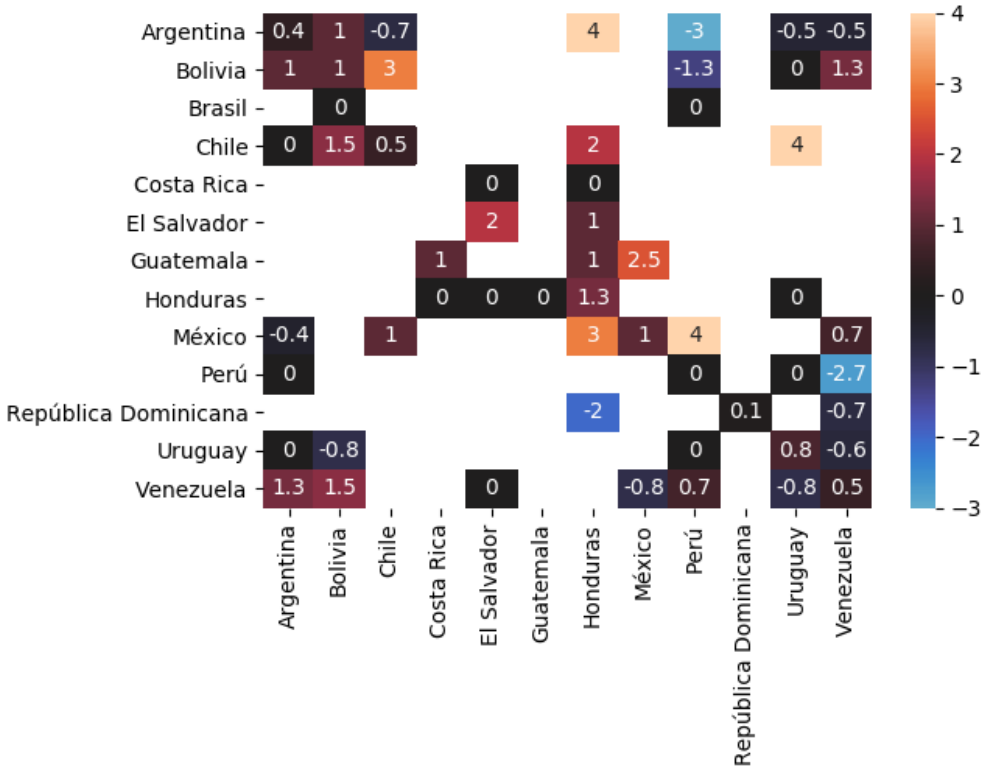


Figura 5.7: Mapa de calor de promedio de polaridad menciones directas

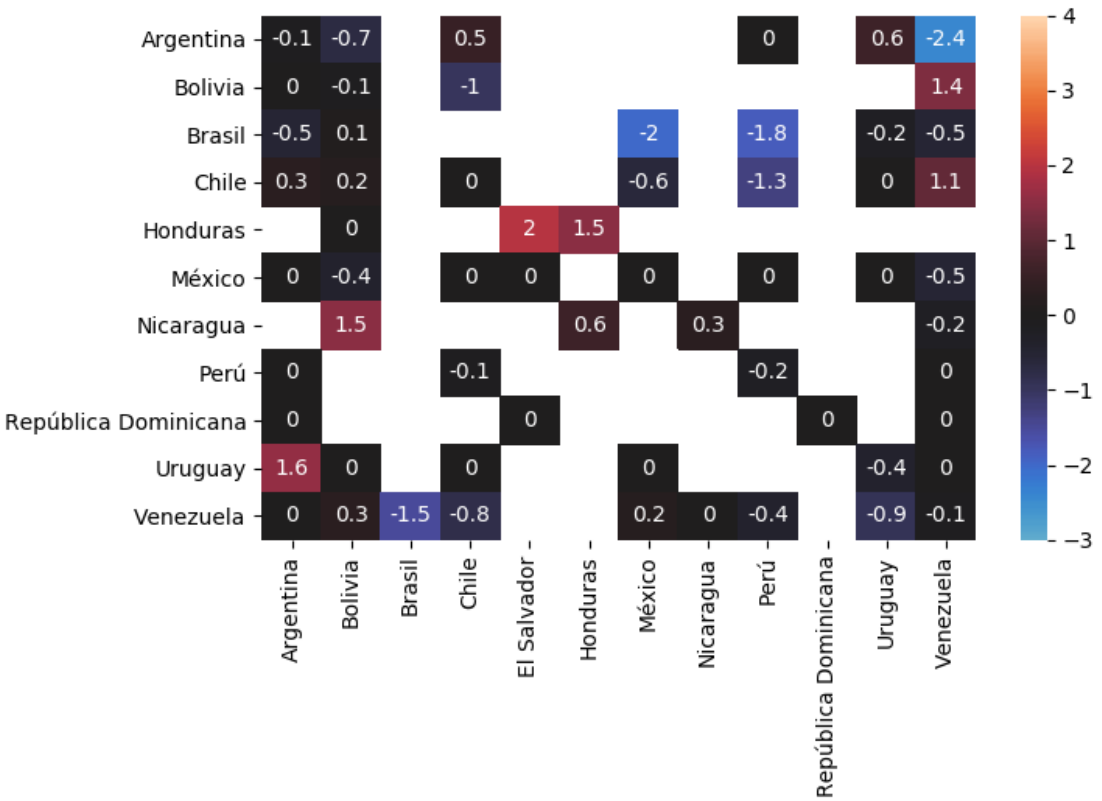


Figura 5.8: Mapa de calor de promedio de polaridad menciones indirectas

Mientras que la columna de los países representa el país emisor del tweet, la fila con los nombres de países corresponde al país mencionado. Esto quiere decir que; como se ve en la figura 5.8, en promedio, los tweets de políticos argentinos aludiendo a políticos venezolanos tienen una polaridad de -2.4, esto quiere decir que corresponden a tweets con mayor cantidad de sentimientos negativos que positivos.

5.3. Evolución de políticos a través del tiempo

Se separaron los tweets que hacían referencia a políticos por fecha. En conjunto con el análisis de sentimiento y las fechas tomadas se establecieron líneas de tiempo que mostraban la evolución promedio de los políticos con respecto a los sentimientos de los tweets.

Al ser miles los políticos de la muestra se decidió trabajar con dos casos locales, como también el top 10 de usuarios más mencionados en caso directos e indirectos.

Se seleccionaron los 10 políticos más mencionados de la muestra. El número de menciones se muestran en las Tablas 5.3 y 5.4, mientras que los gráficos son las Figuras 5.9 y 5.10.

Nombre	Tweets	Polaridad
Nicolás Maduro	829009	-108393
Mauricio Macri	766527	-94506
Sebastián Piñera	673705	-81580
Leopoldo López	475653	-58919
Enrique Peña Nieto	445513	-53805
Patricia Bullrich	399185	-49919
Daniel Ortega	383938	-48139
Amado Boudou	325158	-38351
Elisa Carrió	314075	-40443
Henrique Capriles	309164	-38392

Tabla 5.2: Top 10 menciones indirectas

Nombre	Tweets	Polaridad
Nicolás Maduro	156249	67056
Enrique Peña Nieto	47948	3262
Freddy Guevara	41488	-3734
Mauricio Macri	39951	7435
Henrique Capriles	35915	-7920
Juan Orlando Hernández	27374	30146
Diosdado Cabello	24446	2106
Andrés Manuel López Obrador	23970	1717
Michelle Bachelet	19304	3811
Luis Almagro	18585	-5393

Tabla 5.3: Top 10 menciones directas

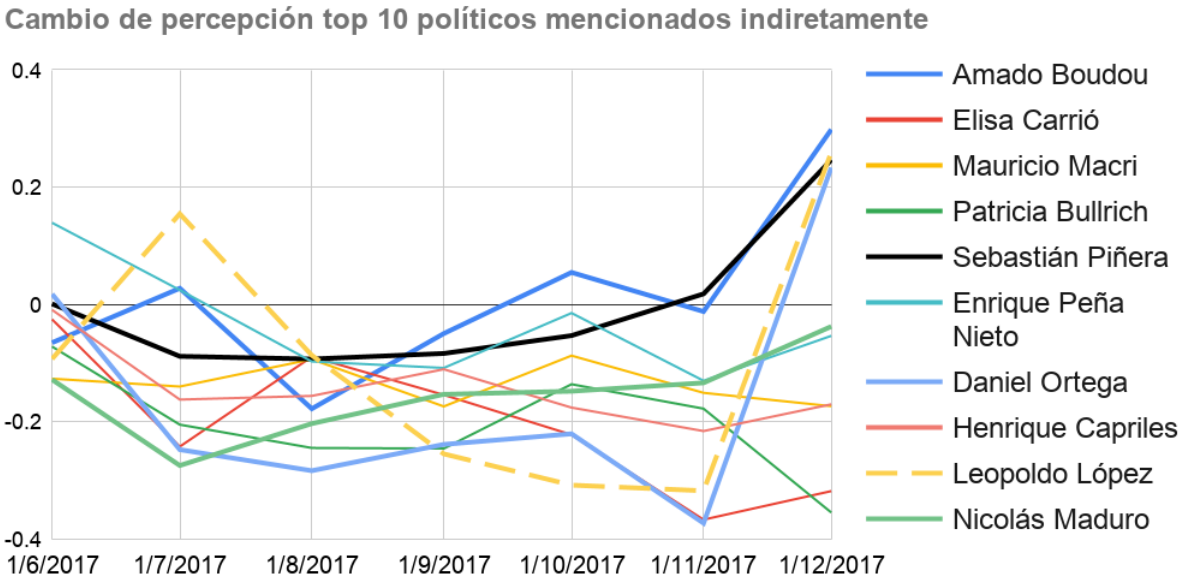


Figura 5.9: Sentiment analysis Top 10 políticos más mencionados indirectamente

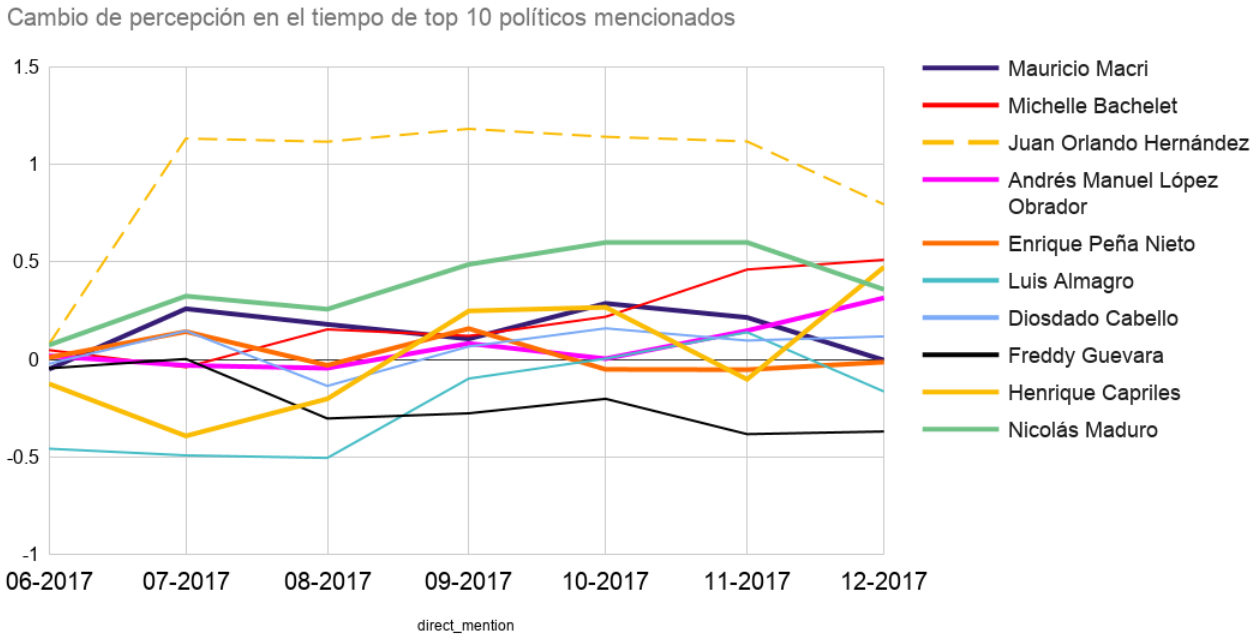


Figura 5.10: Sentiment analysis Top 10 políticos más mencionados directamente

5.4. Comparación de resultados obtenidos con líneas bases

Se decidió realizar una comparación de los datos sociopolíticos extraídos de este trabajo con datos que representasen una línea base. Para esta parte del proyecto se decidió trabajar en dos casos; la aprobación y reprobación de, entonces presidenta, Michelle Bachelet, y en segundo caso la primera vuelta de las elecciones presidenciales de Chile 2017, realizada el 11 de noviembre. Inicialmente se tenía pensado realizar una comparativa del índice de aprobación y reprobación de todas las cabezas de estado de los países latinoamericanos en 2017, pero debido a la falta de acceso a fuentes confiables, se prefirió reducir el rango del experimento a sólo un alcance nacional. De esta manera ambos casos se podrán comparar con la misma fuente (CADEM).

5.4.1. Índice de aprobación presidencial

Para este caso en particular, se extrajo de la página oficial de CADEM todos los índices de aprobación y reprobación de la ex presidenta de Chile Michelle Bachelet que se encontraron entre las fechas correspondientes a la extracción de los tweets (julio a diciembre del año 2017). Los datos se filtraron de tal manera que sólo se seleccionaran los tweets correspondientes a las menciones a Michelle Bachelet; luego de esto se agruparon por mes emitido y usuarios haciendo una suma general de la polaridad de sentimiento que tenían los tweets del usuario. Se consideró que un usuario aprobaba el trabajo de Bachelet si la suma de su polaridad en ese mes era mayor a 0, si es que era menor se consideró que el usuario rechazaba su trabajo. Por último se omitieron todos aquellos comentarios que tenían polaridad acumulada 0. Este

proceso se hizo tanto en menciones directas como indirectas y luego de esto se sumaron ambos resultados.

En la Tabla 5.4 se aprecia la cantidad de usuarios que tanto aprobaban como rechazaban el gobierno según Twitter. Por otra parte, en la Figura 5.11 se puede apreciar la diferencia entre los resultados del índice de aprobación de la encuesta CADEM durante el segundo semestre del 2017 con los resultados generados por el proceso previamente señalado.

	Rechazo	Aprobación	% Rechazo	% Aprobación
Julio	986	1243	44.24	55.76
Agosto	748	1178	38.84	61.16
Septiembre	1338	1967	40.48	59.52
Octubre	808	1353	37.39	62.61
Noviembre	1665	1827	47.68	52.32
Diciembre	238	531	30.95	69.05

Tabla 5.4: Número de usuarios que aprueban y rechazan la entonces presidenta de Chile según Twitter

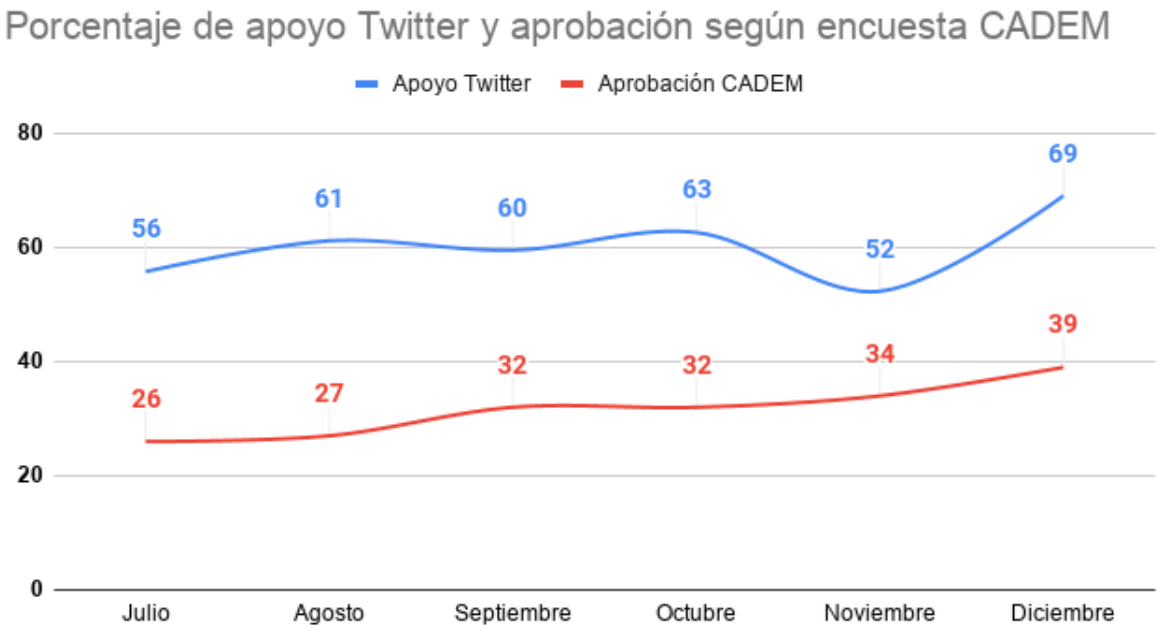


Figura 5.11: Porcentaje de apoyo Twitter y aprobación según encuesta CADEM

Se puede apreciar que ambas curvas se encuentran distanciadas en alrededor de 30 por ciento en la gran mayoría de los meses. Esta diferencia se encuentra disminuida solamente en el mes de noviembre, para luego continuar con el comportamiento en el mes siguiente.

5.4.2. Presencia de candidatos presidenciales en Twitter

En este experimento se seleccionaron sólo aquellos usuarios que tenían una polaridad positiva al referirse a alguno de los candidatos durante los meses de julio a octubre del año

2017. Se decidió utilizar este rango de meses debido a que la primera vuelta de las elecciones se realizó el día 11 de noviembre del año 2017. Cabe recalcar que bajo este método es posible que el mismo usuario de Twitter se considere como posible votante para más de algún candidato.

La distribución porcentual de los datos obtenidos se compararon por candidato, tanto por los resultados de la última encuesta CADEM antes de la elección, como por los resultados de la elección en sí, esto se puede apreciar en la Figura 5.12.

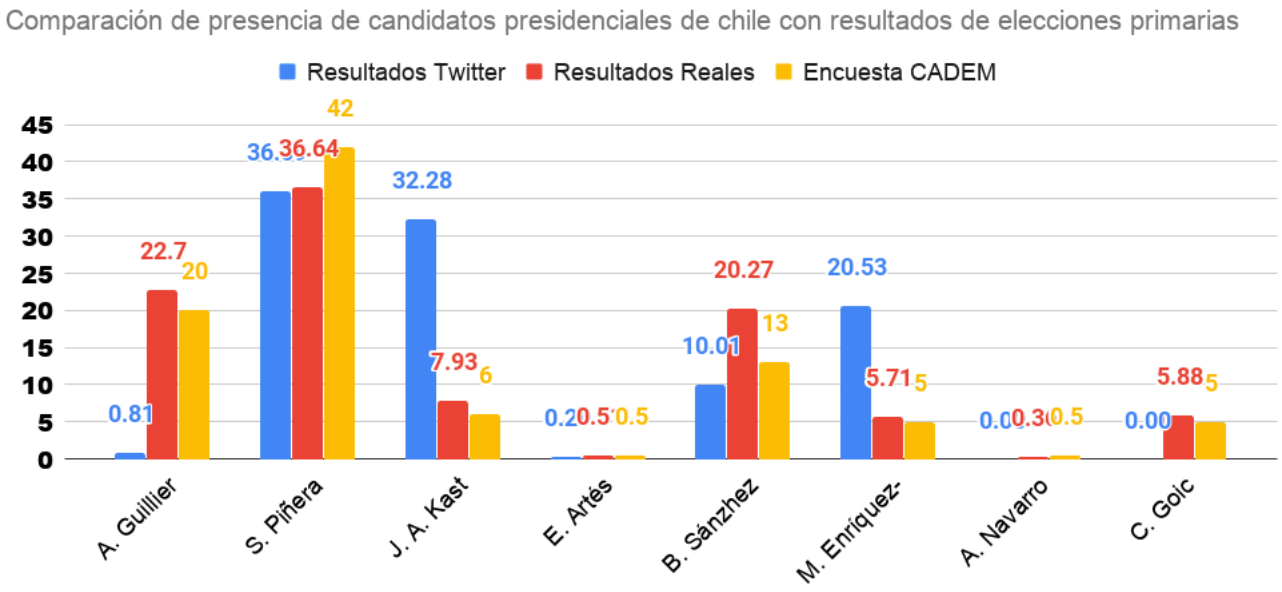


Figura 5.12: Porcentaje de presencia de usuarios con sentimientos positivos hacia candidatos presidenciales en Twitter

De estos resultados, se destaca la anomalía que presentan los candidatos Alejandro Guillier y José Antonio Kast, siendo estos candidatos los con mayor diferencia entre la presencia positiva en Twitter comparado con su resultado final en las elecciones. Cabe mencionar que algunas entidades de los candidatos de Wikidata no tenían usuarios de Twitter asociados, estos fueron; Alejandro Guillier, Alejandro Navarro, Eduardo Artés y Carolina Goic. Esto pudo repercutir en una falta de presencia en los resultados obtenidos.

Capítulo 6

Análisis de resultados y discusión

El siguiente capítulo corresponde a un análisis de todo lo anteriormente señalado, detallando en las particularidades de las carácteristicas y las consideraciones del trabajo que hay que tener en cuenta al ver los resultados.

6.1. Análisis

6.1.1. Muestra de políticos de Wikidata

Con respecto a la muestra señalada en Wikidata se puede notar una alta superioridad de la muestra de presencia de políticos de género masculino llegando a ser no más de 20 % de la muestra de políticos en general; dicha diferencia se reduce al considerar sólo aquellos políticos vivos que tengan usuario en Twitter (la muestra de políticos que se utilizó para la selección de menciones directas). Sin embargo la diferencia que prevalece entre la cantidad de políticos por femeninos con respecto a su contraparte en el total de la muestra corresponde a una relación 72 % masculino v/s 28 % femenino.

Se muestra una considerable diferencia entre mujeres y hombres en el mundo político, más aún si se considera que, según el banco mundial, en la mayoría de los países latinoamericanos existe una mayor presencia de mujeres en la población. Sin embargo, la distribución observada de los géneros en el mundo de la política corresponde a una proporción cercana a la actual (una diferencia no mayor al 5 %) si se habla netamente de políticos [27].

De la misma manera, al comparar el promedio de edad de los políticos por país y género, el resultado fue que en la mayoría de los casos el promedio de edad que poseen las mujeres es menor al de los hombres exceptuando el caso de República Dominicana. Por lo tanto los resultados obtenidos evidencian que el poder de decisión con respecto a lo político es principalmente decidido por personas de edad mayor a sesenta años, siendo la gran mayoría hombres de la tercera edad.

Sumado a esto el hecho de que las mujeres políticas existentes en Wikidata tengan un promedio de edad menor al de los hombres es un dato no menor. Si bien esto podría ser un indicio de que las mujeres más jóvenes están formando discurso político en cada país, la

diferencia que existe entre el promedio de edades de hombres y mujeres dentro de cada país nunca es mayor a 10 años.

Hay que recalcar también que a la fecha de este trabajo, la base de conocimientos al ser hecha la consulta no entregó ninguna referencia a políticos ecuatorianos ni colombianos, esto debido a que la información de Wikidata sobre los paises era incompleta (los países pertenecían a América del Sur e Hispanoamérica, mas no a Latinoamérica).

6.1.2. Características de la base de datos

De los 70 millones de datos que se tenía inicialmente, estos fueron reducidos a alrededor de 1.800.000 datos, sumando tanto menciones directas, indirectas y políticos de la muestra que hubiesen emitido algún tweet. Esto corresponde a un 2,57 % de la muestra principal; es importante señalar que la muestra no detecta tweets que hablen de política por sí mismos, sino sólo aquellos que mencionen de alguna forma a algún político en cuestión.

De las características de la muestra, en las menciones directas, se hace alusión en la gran mayoría a Venezuela, seguido por México y Argentina, los cuales corresponden a los países hispanohablantes más poblados de latinoamérica. Sin embargo las menciones de Venezuela llegan a ser más del triple que las de México y más de el doble que Argentina en las menciones indirectas. La mayoría de las menciones a Venezuela fueron referencias hacia el actual mandatario Nicolás Maduro, posicionándolo como el político más mencionado tanto directa como indirectamente.

Con respecto a la distribución porcentual de ambos tipos de menciones, se observa que en cada país prima un mayor porcentaje de menciones a políticos de género masculino, siendo la excepción Paraguay en las menciones directas y Guatemala con respecto a las menciones indirectas. En el caso de las menciones a políticas femeninas en Chile, si bien este fue mayor al 45 % en menciones directas, las menciones indirectas a políticas de género femenino es menor al 25 %, dato de interés ya que en el año al cual pertenecen los tweets Michelle Bachelet era la mandataria máxima del estado y además habían dos mujeres postulando a la presidencia (Carolina Goić y Beatriz Sánchez).

Para el caso de distribución de menciones diferenciadas por país y género (Figuras 4.16 y 4.17) los grados máximos y mínimos de polaridad se los lleva el género femenino. También se refleja una vez más lo antes señalado de que las menciones directas tienden a tener valores más positivos comparadas con las menciones indirectas. Esto puede darse debido a que las personas que critican políticos no quieren mencionarlos directamente o no simplemente no se saben el usuario de Twitter de este.

Curvas de Lorenz

En el caso de la curva de los tweets emitidos por los políticos (Figura 4.14) se puede apreciar que, si bien esta distribución no es tan drástica como las otras dos señaladas en la sección 4.3 (Figura 4.15 y Figura 4.16), señala una diferencia entre políticos activos e inactivos en Twitter, siendo algunos de estos parte del 1 % de los usuarios que más emite tweets refiriéndose a políticos tanto directa como indirectamente, como lo es, por ejemplo, el caso de Enrique Peña Nieto (usuario encontrado dentro del 1 % que más tweets emite,

descrito en la sección 4.4).

Menciones entre políticos

Con respecto a la impresión entre políticos del mismo país (Figuras 5.7 y 5.8), se tiene que en la mayoría de los casos los resultados directos dieron una polaridad promedio no negativa, a diferencia del caso con las menciones indirectas, de las cuales 22 de las 46 dieron una polaridad promedio negativa. Los políticos argentinos son los que más mencionas a otros políticos de su país, seguidos por los venezolanos, siendo más de mil menciones en cada grupo, luego de esto se encuentran los peruanos, mexicanos y chilenos, los cuales tienen una magnitud de cientos de tweets con los cuales se refieren a sus compatriotas.

Dentro de estos cinco países con más tweets en la muestra, México es el país con mejor polaridad promedio en la muestra indirecta y Chile en la muestra directa. No obstante Chile es el país con peor polaridad promedio en menciones indirectas de los países mencionados, y Argentina el peor en menciones directas. El caso de Chile pudo verse influenciado por la que la muestra coincide con el periodo campaña presidencial y parlamentaria del país.

Usuarios frecuentes en la muestra

Se pudo apreciar que la mayoría de los tweets se encuentran aglomerados en el 1 % de los usuarios más activos de la red social, siendo los usuarios que más mencionan a políticos latinoamericanos usuarios de Twitter que simbolizan medios de prensa. No obstante, existen usuarios pertenecientes a este 1 % que corresponden a personas naturales e incluso habían personas naturales que estaban dentro del 0,1 % que más emitió en la muestra.

Considerando que la muestra inicial correspondía al 1 % de los tweets emitidos en el mundo en el periodo de tiempo designado (similar a 6 meses), esto quiere decir que probablemente los usuarios que están dentro del 1 % que más tweets emitió refiriéndose a políticos latinoamericanos y que corresponden a personas naturales, emiten alrededor de 620 tweets al mes (considerando que la cota inferior para estos usuarios es emitir 37 tweets); esto se traduce en alrededor de 20 tweets al día sin contar todos aquellos tweets que no mencionan a políticos latinoamericanos.

Evolución de políticos a través del tiempo

En el caso de los políticos más mencionados a nivel general, se logra apreciar que el promedio de su polaridad durante los meses varia de 0.4 a $-0{,}4$ en el caso de las menciones indirectas y de 1.5 a $-0{,}6$ en el caso de las menciones directas. Importante señalar que de los políticos más mencionados sólo hubo una mujer en cada categoría, Michelle Bachelet en las menciones directas y Patricia Bullrich en las menciones indirectas.

6.1.3. Comparación de resultados obtenidos con líneas base

De la comparación de resultados obtenidos por la aprobación del gobierno Michelle Bachelet según CADEM comparado con el apoyo de distintos usuarios en Twitter se puede determinar que la red social presenta un mayor porcentaje de apoyo que la encuesta. Este porcentaje presenta una anomalía en su comportamiento durante el mes de noviembre; esta

anomalía podría estar condicionada con que durante ese mes se realizó la primera vuelta de las elecciones presidenciales en el país. Cabe señalar que, como muestra la Tabla 5.4 el número de usuarios que, según la métrica ocupada, apoyaban el gobierno, fue el segundo más alto de toda la muestra, sin embargo el número de usuarios que rechazaron durante ese mes fue el más alto, siendo más que el doble que el mes de octubre.

Con respecto a la presencia de candidatos de la elección presidencial de Chile, en la Figura 5.12 se puede apreciar que tanto en la encuesta CADEM como los resultados reales de la votación y la presencia de apoyo al candidato en Twitter presentan a Sebastián Piñera como el con mayor porcentaje. Sin embargo, cuatro de los candidatos presentan una diferencia mayor al 10 % entre el apoyo de usuarios en Twitter y el resultado de las elecciones. Esto puede deberse a múltiples factores, como lo son el sesgo que hay al sólo considerar usuarios de Twitter, el tipo de campaña que decidió tener cada uno de los candidatos , si es que estos se enfocan en promocionar su campaña a través de redes sociales y/o medios más tradicionales como lo son los diarios y los canales de televisión. Cabe recalcar que cuatro de los políticos no tenían cuenta de Twitter asociada a su entidad de Wikidata, por lo que sus números pueden verse disminuidos con relación a los demás.

6.2. Discusión

En esta parte del informe se mencionan la serie de aspectos a tener en cuenta para evaluar los resultados obtenidos.

6.2.1. Creación de la base de datos

- Dentro de la muestra extraída de la base de conocimientos de Wikidata, no todos los políticos se encontraban con una cuenta de Twitter asociada, por lo que las menciones directas sólo abarcan un espectro del total de políticos: 669 para ser exactos.
- Para la selección de políticos en menciones indirectas se alteró el tweet para que todas las palabras que no estuviesen dentro de un diccionario empezasen con mayúscula. Esta solución ayudó a la hora de detectar ciertos nombres y/o apellidos que no se encontraban en el diccionario, como por ejemplo; “Piñera” o “Bachelet”. No obstante esta solución no ayuda si los apellidos corresponden a palabras que están dentro del idioma español, como el caso de; “maduro”.
- Se determinó la precisión de DBpedia Spotlight al detectar políticos latinoamericanos mencionados indirectamente; esta precisión fue hecha de forma manual evaluando un número de casos y sólo contraponiendo los positivos con los falsos positivos (93 %). No se encontró forma de determinar los datos falsos negativos, esto quiere decir que no hay forma confiable de suponer la cantidad total de aquellos tweets que mencionaban políticos latinoamericanos de manera indirecta pero fueron omitidos por la herramienta, por lo que no se sabe cuantos datos se omitieron.
- DBpedia Spotlight al trabajar con múltiples entidades, le asigna peso a las mismas, por lo que existieron casos en que el tweet mencionaba el apellido de un político y la herramienta le asignaba la entidad con mayor relevancia, como lo es el caso de “José Piñera” y “Sebastián Piñera”. En este ejemplo los dos políticos eran parte de la muestra

sacada de Wikidata, pero al escribir solamente el apellido la herramienta identificaba al ex mandatario.

- La coloquialidad al hablar en los tweets también pudo haber afectado en la selección de los mismos, al no poder identificar de manera óptima un político en caso de que estos estuvieran siendo mencionado por apodos que la base de conocimiento desconoce.
- Debido a la decisión inicial de seleccionar sólo los tweets en español, la muestra sobre Brasil y Haití (Los países de la muestra con una idioma principal distinto al español) no es significativa.
- La falta de políticos originiarios de Colombia, Ecuador, Guayana, Surinam, entre otros genera que se omitan muchos tweets.

6.2.2. Caracterización de la base de datos y análisis de sentimiento

- Al agrupar los políticos por país, varios de estos contaban con menos de un décimo de los tweets que hacían referencia a los países con mayor volumen de menciones, por lo que no se podía hacer una comparativa justa.
- Los tweets retwitteados se contaban como tweets distintos, por lo que una noticia o tweet viral, influye mucho en la muestra.
- El análisis de sentimiento está entregado al tweet en general, por lo que si un tweet hace mención de manera positiva a un político pero menciona de manera negativa a otros dentro del mismo, estos tienen la misma polaridad asociada.
- Al ser la base de datos un filtro de tweets con respecto a mención de políticos latinoamericanos y estar la mayoría de estos acumulados por medios de prensa, la visión general de los tweets presenta un sesgo; esto puede derivar en aumentar o disminuir opiniones generales hacia los políticos. A esto se le debe sumar la posible existencia de usuarios creados por los mismos medios antes señalados, utilizados netamente para divulgar aun más sus noticias [28]

Capítulo 7

Conclusión

7.1. Contribución y relevancia

Gracias al trabajo realizado se ha podido crear una base de datos que enlaza las entidades de Wikidata con los tweets emitidos por múltiples personas, cumpliendo a cabalidad el objetivo presentado en este trabajo. El proceso de creación de la misma es replicable y extensible al conjunto de muestra que se desee, por lo que no está sujeto solamente al ámbito político.

La base de datos generada no tan sólo ha permitido hacer una caracterización del pensamiento general de los usuarios de Twitter hacia los políticos, incluyendo la presencia de estos en la red social, sino que también puede caracterizar a los propios usuarios; este proceso puede servir para extraer bases de datos específicas de ciertos países y así analizar en mayor detalle.

De forma general los usuarios tienden a emitir mensajes con sentimientos más negativos cuando se refieren indirectamente a las personas; esto puede deberse a varias razones como lo son; el desconocer el usuario de Twitter de la persona a la que se le hace alusión o que conscientemente se omita escribir el usuario para evitar que este responda los mensajes.

La distribución general de los políticos, muestra que no solo hay pocas mujeres en el mundo de la política, sino que a su vez esta diferencia aumenta considerablemente en el rol mediáico, aumentando más aún la presencia de políticos masculinos en las redes sociales, disminuyendo más aún la visibilidad de la presencia femenina en el mundo de la política.

Tratar de comprender el actuar de los líderes políticos que nos gobiernan es clave en el mundo de hoy y este trabajo entrega una herramienta más para favorecer ese entendimiento. La comprensión de que la mayoría de los mensajes de la muestra es abarcada por no más de 50 usuarios, tanto medios de prensa como personas naturales, se traduce en una aglomeración sobre la decisión de qué información se transmite de manera más globalizada en las redes. Los datos en cuestión deben conformar parte de la realidad, pero no deben interpretarse como reflejo total de la misma ya que estos acentúan ciertos ámbitos por sobre otros.

7.2. Limitaciones

En la siguiente sección se detallarán y enunciarán las principales limitantes que tuvo el desarrollo del trabajo.

- Al decidir ocupar una base de datos de tweets previamente extraída y trabajar con esto, el trabajo se limitó a considerar solamente el tiempo estipulado en esas bases de datos.
- Si bien existe una predominancia del español como el lenguaje mayormente hablado en latinoamérica, filtrar desde un comienzo los tweets sólo en español fue una limitante ya que no pudo entregar conclusiones claras de políticos de la muestra donde su país no tiene como idioma principal el español, como lo son Brasil y Haití.
- La muestra de políticos de Wikidata estuvo sujeta únicamente a la información que contenía la base de conocimientos en ese momento, esto se tradujo en que se omitieran políticos de varios países que también eran parte de latinoamerica, ya que el país no figuraba como parte de esta en Wikidata, siendo los países sin mencionar y con mayor población Ecuador y Colombia.
- El tiempo necesario para extraer los datos del IMFD llegó a ser de días, al tener esta limitante en el desarrollo del trabajo, se prefirió usar parte de las tablas a las que se tuvo acceso ya que no se tenía en el momento una idea clara de su posible aporte a los resultados.
- Si bien se tenía el acceso a los partidos políticos a los cuales habían militado las diferentes entidades de la muestra, no se tenía un valor asociado a su tendencia, o algún estimado (si el partido era más afín a la derecha o izquierda), por lo que no se pudo ahondar en el análisis en este ámbito.
- Al ser tratarse de un estudio de carácter exploratorio, para ver si es factible llegar a un conjunto de tweets o no, no se ha considerado la actualización del conjunto de datos con tweets nuevos.

7.3. Trabajo futuro

La metodología del trabajo es replicable en caso de desear contribuir con la generación de nuevas bases de datos específicas en diferentes áreas. Además se entrega como resultado una base de datos a disposición para contribuir en estudios multidisciplinarios del mundo de la política.

Por otra parte también se puede aumentar el número de datos con nuevas extracciones de los mismos o añadir otros posibles atributos al conjunto de propiedades de los políticos de Wikidata. Tener un esquema que represente la ideología política de cada partido ayudaría a enriquecer más los resultados extraídos por la base de datos.

Dados los resultados entregados por este trabajo. Se encuentra la posibilidad de replicar el mismo con tweets más actualizados, o con extracción que se actualice automáticamente. Dicha extracción o seguimiento de tweets se escapa de los marcos de este trabajo, por lo implica un posible desarrollo futuro.

Replicar trabajo con datos más actualizados o de un carácter local generaría resultados que ayudarían a comprender la situación actual de un país y la evolución del sentimiento de las personas. Si el trabajo se aplicase a datos que estuviesen constantemente actualizándose, se podría generar una trazabilidad en tiempo real del pensar ciudadano sobre alguna entidad pública, siendo esto una herramienta útil tanto para la creación de encuestas como para los consejeros políticos.

Bibliografía

[1] Twitter. http://twitter.com/.

[2] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World wide Web*, pages 591–600. AcM, 2010.

[3] Wolfgang Maier and Carlos Gómez-Rodríguez. Language variety identification in spanish tweets. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 25–35, 2014.

[4] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

[5] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[6] Wikidata:statistics. https://www.wikidata.org/wiki/Wikidata:Statistics#noSuchAnchor.

[7] Wikidata query service. https://www.mediawiki.org/wiki/Wikidata_Query_Service.

[8] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015.

[9] Wikidata query service. https://www.wikidata.org/wiki/Q306.

[10] Filip Ilievski, Giuseppe Rizzo, Marieke van Erp, Julien Plu, and Raphaël Troncy. Context-enhanced adaptive entity linking. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*, 2016.

[11] Julien Plu, Raphaël Troncy, and Giuseppe Rizzo. Adel@ oke 2017: a generic method for indexing knowledge bases for entity linking. In *Semantic Web Evaluation Challenge*, pages 49–55. Springer, 2017.

[12] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014.

[13] Sujan Perera, Pablo N Mendes, Adarsh Alex, Amit P Sheth, and Krishnaprasad Thirunarayan. Implicit entity linking in tweets. In *European Semantic Web Conference*, pages 118–132. Springer, 2016.

[14] Giuseppe Rizzo, Amparo Elizabeth Cano Basave, Bianca Pereira, Andrea Varga, Matthew Rowe, Milan Stankovic, and A Dadzie. Making sense of microposts (# microposts2015) named entity recognition and linking (neel) challenge. In *# MSM*, pages 44–53, 2015.

[15] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124. ACM, 2013.

[16] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8, 2011.

[17] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[18] Kavita Pabreja et al. Sentiment analysis on gst tweets. *International Journal of Advanced Research in Computer Science*, 9(2), 2018.

[19] Bing Liu et al. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666, 2010.

[20] Pablo Andrés Tapia Caro. Diseño e implementación de un sistema para la clasificación de tweets según su polaridad. 2014.

[21] David Vilares, Mike Thelwall, and Miguel A. Alonso. The megaphone of the people? spanish sentistrength for real-time analysis of political tweets. *J. Inf. Sci.*, 41(6):799–813, 2015.

[22] World Health Organization et al. Global health observatory data repository: Life expectancy data by country. *World Health Statistic*, 2015.

[23] Henry Rosales-Méndez, Barbara Poblete, and Aidan Hogan. Multilingual entity linking: comparing english and spanish. 2017.

[24] pyspotlight. `https://pypi.org/project/pyspotlight/`.

[25] Uruguay wikidata. `https://www.wikidata.org/wiki/Q77`.

[26] ZuXiang Wang, Yew-Kwang Ng, and Russell Smyth. A general method for creating lorenz curves. *Review of Income and Wealth*, 57(3):561–582, 2011.

[27] Women in national parliaments.

[28] Jennifer Golbeck and Derek Hansen. Computing political preference among twitter followers. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1105–1108, 2011.