2017-10

ANÁLISIS DE TENDENCIAS EN ELECCIONES PRESIDENCIALES EN CHILE BASADO EN REDES SOCIALES

COVARRUBIAS VICUÑA, JAIME ALEJANDRO

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE

"ANÁLISIS DE TENDENCIAS EN ELECCIONES PRESIDENCIALES EN CHILE BASADO EN REDES SOCIALES"

JAIME ALEJANDRO COVARRUBIAS VICUÑA

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO
CIVIL INFORMÁTICO

PROFESOR GUÍA: MARCELO MENDOZA ROCHA
PROFESOR CORREFERENTE: JOSE LUIS MARTI LARA

OCTUBRE — 2017

Resumen

Las redes sociales son una parte importante de la sociedad actual, permitiendo relacionar a las personas en múltiples aristas. Varias de estas redes sociales permiten obtener una gran cantidad de información, con la particularidad de permitir realizar estudios y análisis. Una buena manera de trabajar con esta información consiste en realizar análisis de redes, identificando en estos análisis patrones y tendencias. La finalidad de este estudio es exponer como los análisis de redes, a través de varias técnicas, pueden reflejar lo que está ocurriendo en la realidad. También se expondrá, mediante varias herramientas de visualización, una forma más práctica para poder interpretar la información obtenida, después de ser procesada. Por último, se definen conclusiones que pueden ayudar a futuros estudios similares a tener una visión más amplia en análisis relacionados.

Abstract

Social networks are an important part of today's society, making it possible to relate people on multiple edges. Several of these social networks allow to obtain a great amount of information, with the particularity of allowing to carry out studies and analysis. A good way to work with this information is to perform network analysis, identifying in these analyzes patterns and trends. The purpose of this study is to explain how network analysis, through various techniques, can reflect what is happening in reality. It will also be exposed, through various visualization tools, a more practical way to interpret the information obtained, after being processed. Finally, we define conclusions that may help future similar studies to have a broader view in related analyzes.

Agradecimientos

Mi mayor agradecimiento es para mi profesor guía, Marcelo Mendosa, quién tuvo la paciencia y dedicación en todo este arduo proceso.
También agradezco a Guillermo Marguart quién fue un pilar indispensable para el correcto desarrollo del documento.
Como no puede faltar agradezco a mi familia por insistir tanto que obtuviera mi título.
Agradezco a mi polola por su paciencia y apoyo, y por no dejar que me diera por vencido nunca.
Por último y no menos importante agradezco a todos mis amigos por su apoyo incondicional, ya que sin ellos no habría podido realizar un trabajo de estas magnitudes solo.

Glosario

- **Twitter**: Twitter es una red social que permite escribir mensajes de forma pública y seguir a las personas que te interesan (`https://twitter.com/`).
- **Tweet**:Tweet o tuit (en español) se le denomina a un mensaje enviado vía la red social de Twitter.
- **BI**: Corresponde a la tecnología Business Intelligence, es el conjunto de estrategias, aplicaciones, y arquitectura técnicas, para la administración y creación de conocimiento, a través del análisis de los datos existentes en una organización
- **OM**: Minería de Opiniones, esta es una técnica que se encarga de la problemática de determinar las opiniones y sentimientos expresados en un texto en redes sociales.
- **API**: API es una abreviación de Application Programming Interface, las cuales consisten en especificaciones para comunicar interfaces de distintas componente de software. En el caso de esta memoria, se utilizo una API de Twitter para extraer la información analizada
- **Followers**: Es la red de seguidores en la red de Twitter. Es importante remarcar que a diferencia de Facebook las relaciones en la red social de Twitter son unidireccionales, es decir que si una persona sigue a otra no necesariamente esa otra persona te sigue.
- **Followee**: Son las personas a las que siguen a otra cuenta en la red de twitter, también se les denomina friends o amigos.
- **SERVEL**: El SERVEL es el servicio electoral, el cual provee información respecto a las mesas de votación y resultados por mesa, se utilizará la información obtenida en su sitio para comparar los resultados (`http://www.servel.cl`).
- **Stemming**: Es un método para reducir una palabra a su raíz o a un stem o lema (facil, rápido; no contempla contexto). In computational linguistics, lemmatisation is the algorithmic process of determining the lemma for a given word (complejo; contempla contexto).
- **En semántica lingüística**: Se denomina hipónimo a la palabra que posee todos los rasgos semánticos, o semas, de otra más general -su hiperónimo- pero que en su definición añade otras características semánticas que la diferencian de ésta (del hiperónimo).1 Por ejemplo, los hipónimos de día son: lunes, martes, miércoles, etcétera. Es decir, es una palabra que posee todos los rasgos semánticos y añaden otras características para diferenciarlas de esta.
- **Se denomina merónimo**: A la palabra cuyo significado constituye una parte del significado total de otra palabra, denominada ésta holónimo.

- **Un token**: O también llamado componente léxico es una cadena de caracteres que tiene un significado coherente en cierto lenguaje de programación. Ejemplos de tokens podrían ser palabras clave (if, else, while, int, ...), identificadores, números, signos, o un operador de varios caracteres, (por ejemplo, :=).Término – token con significado según un corpus (por ejemplo diccionario)

- **Tokenización**:(análisis léxico) Un analizador léxico y/o analizador lexicográfico (en inglés scanner) es la primera fase de un compilador consistente en un programa que recibe como entrada el código fuente de otro programa (secuencia de caracteres) y produce una salida compuesta de tokens (componentes léxicos) o símbolos. Estos tokens sirven para una posterior etapa del proceso de traducción, siendo la entrada para el analizador sintáctico (en inglés parser).

- **Analizador sintáctico**: Un analizador sintáctico (o parser) es una de las partes de un compilador que transforma su entrada en un árbol de derivación. El análisis sintáctico convierte el texto de entrada en otras estructuras (comúnmente árboles), que son más útiles para el posterior análisis y capturan la jerarquía implícita de la entrada. Un analizador léxico crea tokens de una secuencia de caracteres de entrada y son estos tokens los que son procesados por el analizador sintáctico para construir la estructura de datos, por ejemplo un árbol de análisis o árboles de sintaxis abstracta.

- **Clase**: Define las propiedades y el comportamiento de un Objeto específico. Esta contiene **métodos** (funciones de la clase) y **atributos**. Además puede tener **herencia** de otra clase, lo que implica que hereda (métodos y atributos) de otra **clase**.

- **Objeto**: Un objeto es la instancia de una clase en Programación orientada a objetos.

- **Collocations**: Se denomina de esta forma a los conjuntos de dos o mas palabras en un texto determinado, cuando las palabras se usan regularmente juntas se crean reglas sobre su uso no por razones gramaticales sino por simple asociación, por ejemplo: bigramas y trigramas, que corresponde a un grupo de dos o tres palabras que son utilizadas en el análisis estadístico de texto.

- **Arreglo**: Corresponde a una estructura de datos con forma de lista, la que permite manipular datos de manera muy flexible, ejemplo: array([6, 1, 3, 9, 8]).

- **Caracter**: Corresponde a una letra, signo, número o signo de puntuación.

- **CSV**: : Un archivo CSV (de Valores Separados por Comas) es un tipo de documento que representa los datos de forma parecida a una tabla, es decir, organizando la información en filas y columnas.

- **CIRCOS**: Es un paquete de software para la visualización de datos e información.

Índice general

1 Introducción **1**
- 1.1 Explosión de datos . . . 1
- 1.2 Políticos en la web . . . 2
- 1.3 Técnicas Emergentes . . . 2
- 1.4 Procesamiento de texto . . . 2
- 1.5 Análisis descriptivo de vocabulario . . . 2
- 1.6 Análisis de sentimiento . . . 3
- 1.7 Métricas de redes sociales . . . 3

2 Definición del análisis **4**
- 2.1 Descripción del análisis . . . 4
- 2.2 Objetivos . . . 5
  - 2.2.1 Objetivo General . . . 5
  - 2.2.2 Objetivos Específicos . . . 5
- 2.3 Alcance de este análisis . . . 5
  - 2.3.1 Característica general de la Base de Datos . . . 5
  - 2.3.2 Características de políticos (candidatos) de la Base de Datos 5

3 Estado del arte **7**
- 3.1 Análisis político . . . 9
- 3.2 Estudios similares . . . 10
- 3.3 No transformar social media en otra encuesta . . . 10
  - 3.3.1 Algunas lecciones del caso . . . 11
- 3.4 Social Media, democracia y democratización . . . 13
- 3.5 Combinando Fortalezas, Emociones y Polaridades para Apoyar el Análisis de Sentimientos de Twitter . . . 15

4 Marco Teórico **17**
- 4.1 Análisis de texto . . . 17
- 4.2 Análisis de sentimiento . . . 18
  - 4.2.1 Como funciona el análisis de sentimiento . . . 18
- 4.3 Métricas en redes sociales . . . 21
  - 4.3.1 Twitter . . . 21
- 4.4 Segmentación . . . 23
- 4.5 Ley de Zipf . . . 23

5 Recursos disponibles **25**
- 5.1 Datos: Datos API Twitter . . . 25
- 5.2 Herramientas de análisis usadas . . . 25

5.3 Datos disponibles y modelo de datos . . . 27
5.3.1 Targets . . . 28
5.3.2 Tweets . . . 29
5.3.3 Followers . . . 29
5.3.4 Followees . . . 29
5.4 Preparación de los datos . . . 29
5.5 Lo que pasó en las elecciones . . . 30
5.5.1 Primarias . . . 31
5.5.2 Bajada de Pablo Longueira . . . 32
5.5.3 Primera vuelta . . . 32
5.5.4 Segunda vuelta . . . 33

6 Análisis Exploratorio **34**
6.1 Análisis Léxico . . . 34
6.1.1 Volumen, distribución de términos, ley de Zipf . . . 34
6.1.2 Análisis de sentimiento (polaridad, distribuciones por volumen de tweets), segmentación por candidato y periodo . . . 47
6.2 Análisis de la red . . . 52
6.2.1 Análisis de co-seguimiento . . . 52
6.2.2 Análisis de co-mención . . . 53
6.2.3 Análisis de línea de tiempo . . . 55
6.2.4 Análisis de las cuentas que más tweets realizaron a un candidato en específico . . . 55
6.3 Análisis comparado (elecciones versus datos). Muestra sesgada (análisis de sesgo) . . . 57

7 Conclusiones **59**
7.1 Análisis de los resultados obtenidos . . . 59
7.2 Futuros trabajos . . . 61

Bibliografía **62**

A Anexos de código **67**

B Anexos Imágenes **69**

C Anexos de Gráficos **70**
C.1 Co-menciones . . . 70
C.2 Co-seguimiento . . . 72
C.3 Línea de Tiempo . . . 74
C.4 Primer Periodo . . . 75
C.5 Segundo Periodo . . . 83
C.6 Tercer Periodo . . . 92
C.7 Zipf . . . 93

Índice de figuras

5.1 Modelo de Datos Inicial 27
5.2 Modelo de Datos Final 28
5.3 Candidatos Presidenciales 31
5.4 Resultados Primarias Nueva Mayoría [33] 31
5.5 Resultados Primarias Alianza [33] 32
5.6 Resultados Primera Vuelta [33] 33
5.7 Resultados Segunda Vuelta [33] 33

6.1 Importación de Librerías 34
6.2 Conexión Base de Datos 35
6.3 Consulta Cantidad de Registros 35
6.4 Consulta por cada Iteración 36
6.5 Concatenación 36
6.6 Tokenización 37
6.7 Texto a minúsculas 37
6.8 Eliminación de Tokens que no son palabras 38
6.9 Eliminación de caracteres en Tokens 38
6.10 Definición de Stopwords 38
6.11 Eliminación de Stopwords 38
6.12 Frecuencia Distribuida de Tokens 39
6.13 Frecuencia Distribuida de Bigramas y Trigramas 39
6.14 Vocabulario 39
6.15 Consolidación de Términos 39
6.16 Generación de CSV 40
6.17 Frecuencia de Palabras, Mes de Junio 43
6.18 Frecuencia de Palabras, Mes de Noviembre al 4 de Diciembre 44
6.19 Frecuencia de Palabras, Base de Datos completa 46
6.20 Tweet Franco Parisi 47
6.21 Menciones en Twitter, Primer Periodo, Michelle Bachelet 48
6.22 Menciones en Twitter, Primer Periodo, Evelyn Matthei 49
6.23 Menciones en Twitter, Segundo Periodo, Franco Parisi 50
6.24 Menciones en Twitter, Segundo Periodo, Marco Henriquez-Ominami 50
6.25 Menciones en Twitter, Tercer Periodo 51
6.26 Co-seguimiento, Bachelet y Matthei 52
6.27 Co-mención, Bachelet y Matthei 54
6.28 Línea de Tiempo 55
6.29 Tweets bajada de Longueira, subida de Matthei 55
6.30 Usuarios Twitter por grupor etarios [37] 57
6.31 Votantes Elecciones Municipales por grupos etarios [38] 58

A.1 Función de eliminación de tildes 67
A.2 Lista de Stopwords 68

B.1 Estructura de un Tweet 69

C.1 Co-menciones Bachelet Matthei 70
C.2 Co-menciones todos los candidatos 71
C.3 Co-seguimiento Bachelet Matthei 72
C.4 Co-seguimiento todos los candidatos 73
C.5 Línea de Tiempo 74
C.6 Menciones en Twitter Primer Periodo, Michelle Bachelet 75
C.7 Menciones en Twitter Primer Periodo, Marcel Claude 76
C.8 Menciones en Twitter Primer Periodo, Marco Enriquez-Ominami . . 77
C.9 Menciones en Twitter Primer Periodo, Tomás Jocelyn-Holt 78
C.10 Menciones en Twitter Primer Periodo, Evelyn Matthei 79
C.11 Menciones en Twitter Primer Periodo, Roxana Miranda 80
C.12 Menciones en Twitter Primer Periodo, Franco Parisi 81
C.13 Menciones en Twitter Primer Periodo, Alfredo Sfeir 82
C.14 Menciones en Twitter Segundo Periodo, Michelle Bachelet 83
C.15 Menciones en Twitter Segundo Periodo, Marcel Claude 84
C.16 Menciones en Twitter Segundo Periodo, Marco Enriquez-Ominami . 85
C.17 Menciones en Twitter Segundo Periodo, Marco Ricardo Israel . . . 86
C.18 Menciones en Twitter Segundo Periodo, Tomás Jocelyn-Holt 87
C.19 Menciones en Twitter Segundo Período, Evelyn Matthei 88
C.20 Menciones en Twitter Segundo Período, Roxana Miranda 89
C.21 Menciones en Twitter Segundo Período, Franco Parisi 90
C.22 Menciones en Twitter Segundo Período, Alfredo Sfeir 91
C.23 Menciones en Twitter Tercer Período 92
C.24 Frecuencias de Palabras,Bases de Datos Completa 93
C.25 Frecuencias de Palabras, Mes de Mayo 94
C.26 Frecuencias de Palabras, Mes de Junio 95
C.27 Frecuencias de Palabras, Mes de Julio 96
C.28 Frecuencias de Palabras, Mes de Agosto 97
C.29 Frecuencias de Palabras, Mes de Septiembre 98
C.30 Frecuencias de Palabras, Mes de Octubre 99
C.31 Frecuencias de Palabras, Mes de Noviembre al 4 de Diciembre . . . 100

Índice de cuadros

2.1 Cantidad de referencias por político 6
2.2 Cantidad de seguidores por político 6

6.1 Tweets y Vocabulario por Mes 41
6.2 20 términos más frecuentes de junio 42
6.3 20 términos más frecuentes de noviembre 44
6.4 20 términos más frecuentes, Base de Datos completa 46
6.5 Co-seguimiento, Bachelet y Matthei 52
6.6 Co-mención, Bachelet y Matthei 53
6.7 Cuentas de Twitter que realizaron más Tweets 56

Capítulo 1

Introducción

Las redes sociales han sido parte importante los últimos años, muy en especial en la sociedad chilena, cada día con más opciones para relacionar a las personas, en diferentes ámbitos y tendencias.

A través de éstas poder acceder a gran cantidad de información, exponer opiniones o noticias entre otros y hasta poder compartir la vida personal. Con esto podemos obtener una gran cantidad de información relacionada con cualquier tema, pudiendo ser de interés, actualidad, tendencias, gustos, etc. Con lo anterior se puede realizar análisis a gran cantidad de datos en la web.

1.1 Explosión de datos

En todas las redes sociales se invita a los usuarios a participar realizando preguntas como, ¿en que estas pensando? (Twitter y Facebook) o "Comparte Algo" como invita LinkedIn o la frase "¿Tienes algo nuevo que contar?" que figura en Google+. Todas estas llamadas a la acción están pensadas para que los usuarios de las diferentes redes generen contenido e interactúen entre si.

Como resultado de lo anterior las redes sociales están generando muchísima información, día tras día los usuarios generan contenido, comentarios y cambios de estado, según cifras reveladas por Twitter indican que se están contabilizando 50 millones de tweets por día lo que equivale a 600 tweets por segundo. [1]

Algunos datos interesantes para acentuar lo comentado anteriormente referente la inmensa cantidad de datos que se está generando en redes sociales e Internet los aporta el sitio pingdom.com donde indica por ejemplo que en el año 2012 existían 2,2 billones de usuarios de correos electrónicos en el mundo, 634 millones de sitios web, 1,2 billones de búsquedas en el buscador Google, 7 petabytes de fotografías que se añaden a Facebook cada mes. [2]

Lo anterior conlleva un gran desafío, como poder lograr analizar esta gran cantidad de información y obtener de ella resultados que ayuden a la toma de decisiones por ejemplo en el ámbito de la economía o la política.

1.2 Políticos en la web

El Marketing digital es una tendencia ya masiva en nuestros tiempos, las grandes empresas se han desarrollado ampliamente en este ámbito, la política no ha querido quedar fuera y se ha unido al uso de estas estrategias para promover sus ideas y campañas. La red social Twitter tiene características que la hacen un campo idóneo para publicar y promover ideas y debate ya que los perfiles de los usuarios son públicos y todas las conversaciones las puede ver cualquiera y lo más importante nada se elimina, todo se queda allí.

El presidente **Barack Obama** usó las redes sociales para su campaña y sin ellas no hubiera sido electo como presidente de los Estados Unidos. **The Washington Post** llamo a Barack Obama como el rey de las redes sociales el año 2008 en plena campaña presidencial. **The New York Time** realizó un análisis de las publicaciones que Obama hizo en redes sociales, resultando un total de 14 millones de horas a costo cero, esto hubiera costado 47 millones de dólares realizarlo en medios tradicionales como la Televisión. [3]

1.3 Técnicas Emergentes

Estamos en presencia de un nuevo fenómeno que permite analizar los contenidos de las redes sociales denominado "Minería de Opiniones"(OM), esta técnica aborda la problemática de determinar las opiniones y sentimientos expresados en un texto, por ejemplo las opiniones que un grupo de personas tiene sobre un producto determinado del mercado o en el caso de esta memoria cual es la opinión de un grupo de personas sobre un candidato a presidente. [4]

1.4 Procesamiento de texto

El procesamiento de texto o análisis de palabras por medio de técnicas permite obtener estadísticas de las palabras, como por ejemplo con que frecuencia se usan dichas palabras y permite encontrar las denominadas "palabras claves", palabras que ayudan a interpretar el estudio que se realiza y como veremos más adelante nos ayuda al análisis de sentimiento.

1.5 Análisis descriptivo de vocabulario

Esta técnica permite encontrar las palabras más usadas, también permite determinar la frecuencia de uso y con esto logrando determinar las palabras principales (claves) de un tema en particular.
Claramente con este análisis se pueden elaborar estadísticas las que se pueden ver en diferentes perspectivas para poder entender el comportamiento de las personas al hablar de un determinado tema.

1.6 Análisis de sentimiento

Las palabras en los mensajes pueden ser analizadas como objetivas o subjetivas, de esta forma se puede determinar si la palabra es positiva o negativa.

A esta técnica se le denomina análisis de sentimiento, puntualmente en los datos obtenidos desde Twitter se usó la técnica SentiWordnet que permitirá hacer un análisis de la información.

Es importante medir los sentimientos en Twitter dado que se puede determinar si los comentarios tienen una influencia positiva o negativa referente a un determinado tema.

En esta memoria el análisis de sentimiento se usará para analizar los comentarios de la audiencia seleccionada referente a los candidatos a la presidencia de las elecciones del año 2013.

1.7 Métricas de redes sociales

En redes sociales podemos determinar desde el numero de seguidores hasta el alcance de los comentarios emitidos, a esto se le denominan métricas de redes sociales. Algunas redes sociales muestra las métricas más usadas como el número de seguidores.

En esta memoria se analizarán las métricas de redes sociales con el análisis de sentimientos que se obtengan de los datos recolectados.

Capítulo 2

Definición del análisis

Las redes sociales han sido parte importante los últimos años, muy en especial en la sociedad chilena, cada día con más opciones para relacionar a las personas, en diferentes ámbitos y tendencias, a través de éstas poder acceder a gran cantidad de información, exponer opiniones o noticias entre otros y hasta poder compartir tu vida personal. Con esto podemos obtener una gran cantidad de información relacionada con cualquier tema, pudiendo ser de interés, actualidad, tendencias, gustos, etc. Gracias a esto se pueden realizar análisis a gran cantidad de datos en la web.

Las redes sociales han sido fieles testigos de la participación de este grupo de personas que opinan, se informan y adquieren conciencia de lo que ocurre en torno a la política chilena, más aún en un periodo de elecciones presidenciales.

Una forma muy efectiva de conocer el mercado y un público especifico con respecto a un tema, es a través de la minería de datos. Este proceso es fundamental para obtener información a partir de cualquier tipo de datos, y encontrar un comportamiento o patrones los cuales tengan un estructura o visibilidad comprensible para un uso posterior. Por ejemplo, visualizar la inclinación política durante las elecciones presidenciales, dando la posibilidad de realizar fuertes campañas hacia ese segmento de público, enfatizando lo que opinan y logrando cautivar a dicho público objetivo.

2.1 Descripción del análisis

El propósito de esta investigación es evidenciar cómo a través de una plataforma de red social, como Twitter, se puede aplicar un proceso de minería de datos, y así obtener una gran cantidad de información recopilada, analizada y estructurada de distintas maneras, con la finalidad de comprender y estudiar en específico el proceso de las elecciones presidenciales en Chile durante el año 2013.

2.2 Objetivos

2.2.1 Objetivo General

Exponer las tendencias y relaciones que existen entre los candidatos presidenciales en Chile, por parte de los usuarios de Twitter durante el periodo de 2013.

2.2.2 Objetivos Específicos

- Conocer y estudiar la red social Twitter y comprender la información que entrega cada tweet.
- Analizar una base de datos de Twitter para obtener las tendencias y patrones existentes relacionados con los candidatos presidenciales.
- Seleccionar y aplicar técnicas de minería de datos adecuadas para el estudio en conjunto con las herramientas de visualización correspondientes.

2.3 Alcance de este análisis

El análisis de la presente memoria fue realizado gracias una base datos extraída desde Twitter, la cual fue ordenada en un modelo de datos con las características requeridas para hacer los diferentes estudios usando los métodos de análisis seleccionados.

2.3.1 Característica general de la Base de Datos

- **Forma de captura de los datos**: API de Twitter
- **Espacio en disco de la BD**: 3.3 GB.
- **Rango de fechas de datos utilizados**: 01-05-2013 al 04-12-2013.
- **Cantidad de Tweets**: 5.800.484 registros de Tweets.
- **Cantidad de cuentas de usuarios**: 176.501 cuentas registradas.

2.3.2 Características de políticos (candidatos) de la Base de Datos

- **Políticos estudiados**: 9 candidatos presidenciales, y Pablo Longueira
- **Numero de cuentas de políticos**: 18 cuentas asociadas a los políticos analizados (9 candidatos y Pablo Longueira).

Cantidad de referencias por político:

Político	**N.º Tweets referenciado**
Evelyn Matthei	1.084.975
Michelle Bachelet	1.640.647
Marco Enríquez-Ominami	439.063
Alfredo Sfeir	101.629
Franco Parisi	1.963.178
Tomás Joselyn-Holt	79.028
Ricardo Israel	24.211
Marcel Claude	468.108
Roxana Miranda	188.956
Pablo Longueira	130.197

Cuadro 2.1: Cantidad de referencias por político

Cantidad de seguidores por político:

Político	**N.º Tweets referenciado**
Evelyn Matthei	142.508
Michelle Bachelet	211.275
Marco Enríquez-Ominami	65.832
Alfredo Sfeir	28.843
Franco Parisi	609.557
Tomás Joselyn-Holt	28.435
Ricardo Israel	11.506
Marcel Claude	52.337
Roxana Miranda	53.638
Pablo Longueira	37.714

Cuadro 2.2: Cantidad de seguidores por político

Capítulo 3

Estado del arte

Internet está cambiando la forma de las comunicaciones, pero el mayor impacto está en los medios de comunicación masivos como diarios, radios y televisión. Estos medios están con una curva descendente en su demanda, ya muchas personas dejaron de ver TV y ahora usan Netflix para ver sus series favoritas cuando ellos quieren y no cuando la emisora de TV lo programa, además muchas personas cuentan con Smart TV lo que conlleva a que las personas tienen contenido gratis on line como videos en Youtube, noticias en línea sin la necesidad de esperar al día siguiente cuando el periódico llega a casa.

En los Estados Unidos, este fenómeno se conoce como "cord cutting". Consiste en cortar el cable. Durante los últimos cinco años, las suscripciones a la TV paga en ese país han disminuido un 10 % y a fines de 2015, el periódico financiero The Wall Street Journal publicó un estudio que estimaba que para 2019, un 23 % de los hogares no estaría pagando ningún tipo de servicio de cable, favoreciendo plataformas de streaming. [5]

Manuel Castells describe esta actividad de comunicación continua online como "Yo++ es vivir siempre comunicado, es transportable, es personalizado.". Esto implica que dado un determinado perfil de usuario la información que este ve está mucho más relacionada con sus intereses y no consume la información que no desea ver. [6]

El concepto del Yo++ es muy interesante ya que juega un poco con el ego natural de cada ser humano, dado que las tecnologías hoy permiten este acceso a internet personalizado y una cualidad muy importante es que está en todos lados, en todos los dispositivos, Smart TV, computadores, tablets, teléfonos inteligentes, relojes inteligentes y actualmente en el IoT (internet de las cosas)

En el último tiempo hemos sido testigos como se dan a conocer a la luz escándalos originados por hechos de corrupción que a nadie dejan incólume, las redes sociales difunden detalles y sirve de vía de escape para que muchas personas den su opinión, generando con esto debates sociales.

Claramente las redes sociales han permitido que se divulgue cada vez más y más rápido los hechos noticiosos en todo el mundo. Muy conocidas son las publicaciones de wikileaks en donde se desclasifican documentos secretos que implican

fundamentalmente a países y sus políticas internas y externas.

Con el uso que las personas le dan hoy a las redes sociales la política y la economía ya no serán las mismas. Los ciudadanos pueden influir en cambios de gobiernos y llevar al poder a sus candidatos, con el uso de redes sociales los consumidores pueden destruir una industria y potenciar otra.

Hoy vivimos una crisis de confianza en las instituciones que se suponía que eran sólidas y que sostienen la sociedad como la conocemos; la iglesia, la política, la economía y el libre mercado. Todo hoy en día está en cuestionamiento, cada uno de estos pilares de nuestra sociedad están débiles y gracias a las redes sociales se han desenmascarado y conocidos los detalles de la crisis que los afecta.

3.1 Análisis político

Si hablamos netamente de los temas políticos, y queremos evaluar el impacto que tiene internet en estos temas, primeramente es menester mencionar que hoy se realizan varias encuestas en los medios convencionales, descritos anteriormente (TV, radio, etc.).

En primer lugar, la mas relevante de las encuestas dado que tiene un mayor nivel de cobertura o se considera dentro de las mas validas por los expertos es la **Encuesta de Caracterización Socioeconómica Nacional (CASEN)** del Ministerio de Desarrollo Social, que según su definición es una encuesta a hogares de carácter multipropósito, es decir, que abarca múltiples y diversos temas como educación, trabajo, ingresos, salud, entre otros; además es una encuesta transversal, por lo tanto, incluye a todo el espectro de la población del país. [7]

Otras encuestas que intentan recabar información de la opinión publica son las encuestas del **Centro de Estudios Públicos (CEP)** que según indica su sitio web (www.cepchile.cl), es una fundación privada, sin fines de lucro, de carácter académico y dedicada a los temas públicos. Su finalidad es el estudio y difusión de los valores, principios e instituciones que sirven de base a una sociedad libre. [8]

Por otro lado existe **GFK Adimark**, empresa investigadora de mercado y opinión pública. Esta empresa tiene en Chile más de 40 años de experiencia escuchando la opinión de los chilenos, sobre su vínculo con el consumo, el mercado, las empresas, la economía y las instituciones políticas y sociales. [9]

Por último existe otra empresa de investigación de opinión pública relevante en Chile llamada MORI (Market & Opinion Research International). Esta realiza investigaciones desarrollando nuevos productos por medio de encuestas cuantitativas, cualitativas, focus groups, entrevistas en profundidad, entre otros. [10]

Pese a que estas encuestas y estudios cuentan con un excelente grado de validez, pueden ser cuestionables por el diseño de sus instrumentos utilizados.

Por lo anterior se torna muy relevante estudiar los patrones de conversaciones en redes sociales, entre las cuales destacan Twitter, Facebook, Instagram. El hecho de poder hacer análisis de la información que circula en internet y puntualmente sobre la plataforma Twitter en la que se basa este estudio, la gran cantidad de datos almacenados en esta plataforma nos puede ayudar a entender que están opinando las personas sobre un tema en particular en un determinado momento y esa gran cantidad de información analizada puede ser útil para la toma de decisiones.

Con las nuevas tecnologías, la opinión publica puede ser estudiada en línea sin solicitar a las personas que respondan una encuesta, todo esto gracias a la "minería de opinionesz una serie de métodos que permiten analizar las emociones de las personas frente a un determinado evento.

3.2 Estudios similares

Existen algunos estudios similares a la presente memoria como por ejemplo Nepotistic Relationships in Twitter and their Impact on Rank Prestige Algorithms. Este trabajo ofrece una lista de algoritmos factibles para clasificar a los usuarios en las redes sociales, examina sus vulnerabilidades para vincular la mala práctica en dichas redes y sugiere un criterio objetivo para compararlos dichos algoritmos y un primer paso hacia la "desensibilización"de estos y de su prestigio contra el engaño por los spammers y otros usuarios abusivos.

3.3 No transformar social media en otra encuesta

Todo contenido publicado en sistemas de micro-blogging como Twitter se piensa que es factible para la extracción de datos y "tomar el pulso"de la sociedad. Recientemente, se han publicado varios estudios positivos sobre la práctica de enfoques de muestreo, minería de opinión y análisis de sentimientos. El presente documento intenta hacer del abogado del diablo, detallando un estudio en el cual tales acercamientos sobrestimaron en gran parte la victoria de Obama en las elecciones presidenciales 2008 de los EEUU.

Los usuarios de Twitter no sólo proporcionan información sobre sí mismos, sino que también publican en tiempo real. Por lo tanto, Twitter esta recibiendo una fuente constante de información sobre eventos actuales, en tiempo real de millones de usuarios que están reaccionando a esos eventos.

El objetivo de esta publicación no es proporcionar una encuesta exhaustiva sobre este tema, sino más bien centrarse en: La predicción de eventos actuales y futuros mediante el uso de datos de Twitter. Tal aplicación parece bastante natural a la luz de los excelentes resultados obtenidos mediante registros de consultas de minería.

A partir de diciembre de 2008, el 11 % de los adultos estadounidenses en línea estaban utilizando Twitter o servicios análogos. Si bien esa es una cantidad importante, la realidad muestra que la mayoría de los usuarios de Internet, por no hablar de la gente en general, no usan Twitter. Así, los usuarios de Twitter son sólo una muestra y probablemente una muy sesgada. Por este motivo no se puede predecir las elecciones.

Además, otro tipo de sesgo favorece la investigación: la tendencia de los investigadores a contar más de una vez resultados positivos, al mismo tiempo que se suprimen los negativos. Esto se denomina "drawer", y puede tener una influencia perjudicial si las personas comprenden claramente que las conclusiones de unas pocas experiencias positivas seleccionadas pueden aplicarse directamente a cualquier otro escenario concebible.

Para los propósitos del estudio, se inició una colección de tweets poco después de las elecciones presidenciales de los Estados Unidos de 2008 para comprobar la

factibilidad de usar Twitter para predecir los resultados futuros de las elecciones. Se usó el Search API de Twitter, empleando una consulta para cada candidatura. Se usó un parámetro de API para indicar un área geográfica para considerar solamente los tweets publicados por los residentes de EE.UU. Al uso de estas consultas "Geolocalizadas", también se empleó otro parámetro para indicar un intervalo temporal para la consulta. Por lo tanto, mediante la emisión de consultas limitadas por razones geográficas y temporales, era posible obtener 100 tweets por cada candidato, cada día. Para cada condado habría implicado el envío de un gran número de solicitudes HTTP a los servidores de Twitter.

El uso de la API de esta manera fue posible recolectar datos a septiembre de 2008. Para obtener tweets desde principios de junio, se rastreo cada usuario en los datos ya recogidos, guardando los tweets mencionando en una de las candidaturas. Esto significó que la colección comprendía 250.000 tweets, publicados por 20.000 usuarios entre el 1 de junio de 2008 y el 11 de noviembre de 2008. Lo primero que se debía comprobar era si el conjunto de datos podía considerarse como una muestra representativa. Por lo tanto, el número de tweets y usuarios únicos en cada estado se compararon con sus poblaciones. Además, los errores de muestreo se calcularon basándose en el supuesto de que la recolección se aproximaba a una muestra de escala. [13]

A pesar de la extensa bibliografía sobre el análisis automático del sentimiento, prácticamente todas las investigaciones actuales sobre el análisis de microblogs se basan en métodos bastantes simples. Para el propósito de este estudio se aplicaron cuatro métodos diferentes. Uno se basó en los recuentos de mención, dos se basaron en léxicos de polaridad, y el último se basó en la orientación temática. La idea subyacente del primer método era sencillo: contar el numero de solicitudes de un candidato en los tweets del usuario asumiendo que el que se mencionara con mayor frecuencia sería el que el usuario votaría. Esta heurística es suficiente, pero, curiosamente, parecía funcionar para predecir el resultado de las elecciones en Alemania [14]. Lo que ha llevado a Tumasjan a comentar: "El mero número de tweets refleja las preferencias de los votantes y se acerca a las comicios tradicionales".

El segundo método se basó en el léxico compilado por Wilson [15], que consistía en una lista de Términos de laboratorio positivos y negativos. Por lo tanto, el Tweet fue tratado positivamente si contenía más términos positivos que negativos y viceversa. Debido a que cada tweet en la colección trató con sólo un candidato, fue posible contar, para cada usuario, el número de tweets positivos y negativos para cada candidato. Por lo tanto, se supone que un usuario votaría por el candidato con la puntuación más alta. Un método similar fue empleado por O'Connoret al. [16] Con resultados mixtos; Quien afirmó: "Una alta tasa de error simplemente implica que el detector de sentimientos es un instrumento de medición ruidosa. Con un número bastante grande de mediciones, estos errores se anularán en relación con la cantidad que nos interesa en la estimación de la opinión pública agregada".

3.3.1 Algunas lecciones del caso

En resumen, las Elecciones Presidenciales de los Estados Unidos de 2008 no pudieron haber sido certeras aplicando las metodologías actuales más comunes. Esto

es consistente ya que no se hizo ninguna correlación sustancial entre un análisis de sentimiento de tweets y varias encuestas preelectorales llevadas a cabo durante la campaña. Además, los posibles sesgos en los datos son coherentes con los resultados de [17], [18]. Por lo tanto, el problema de predecir los resultados de estas elecciones no estaba en la recopilación de datos. En su lugar, el problema se ha producido al minimizar el impacto del sesgo en los datos de los Medios Sociales y al ignorar cómo estos datos difieren de la población. Se pueden extraer varias lecciones de esto:

- **La gran caída de los datos**: Los Medios Sociales son muy útiles porque los investigadores pueden obtener grandes colecciones de datos para extraerlas. Sin embargo, el hecho de que sean acotados no hace que tales colecciones sean estadísticamente representativas de la población en su conjunto.

- **Véase el sesgo demográfico** : En la misma línea que la primera lección: Los usuarios medianos tienden a ser relativamente jóvenes y dependiendo del grupo de interés, esto puede introducir un sesgo importante. Para mejorar los resultados es necesario conocer la edad del usuario y tratar de corregir el sesgo en los datos.

- **Análisis del sentimiento ingenuo**: Es posible que algunas aplicaciones pueden alcanzar resultados razonables. Sin embargo, como se muestra en este documento, se deben evitar las metodologías que promueban resultados ruidosos y los investigadores deben siempre revisar cuidadosamente si están o no usando un clasificador aleatorio. Además, los textos de naturaleza política son especialmente difíciles de tratar [19].

- **La solidez dice mucho**: Si la falta de información es mayoritariamente de un solo grupo, los resultados pueden diferir considerablemente de la realidad. No es necesario decir que estimar el grado y la falta de respuesta es muy difícil si no es completamente imposible. Y por lo tanto los investigadores deben ser cautelosos de los peligros que implica.

- **Unos pocos resultados positivos del pasado no garantizan generalización**: Los investigadores deben ser siempre conscientes del .[ef]ecto cajónz debe evaluar cuidadosamente los representantes positivos antes de asumir que los métodos representados pueden aplicarse mejor a cualquier escenario similar con resultados idénticos. Esto es particularmente importante si hay contraejemplos, como el detallado en este estudio.

En resumen, hasta que los Medios Sociales lleguen a ser utilizados regularmente por la gran mayoría de personas, sus usuarios no pueden ser considerados una muestra representativa y por lo tanto, las previsiones de tales datos serán de valor cuestionable e incorrectas en muchos casos. Hasta entonces, si se utilizan estos datos, es necesario identificar los diferentes tipos de usuarios basados en la edad, el ingreso, el género, la raza, etc. con el fin de ponderar sus opiniones de acuerdo con la proporción de cada estrato en la población

3.4 Social Media, democracia y democratización

En 2015, con alrededor de 2,6 millones de tweets, se pintó un cuadro detallado de la reacción Twittersphere a los temas cubiertos por Obama. Esto es sólo uno de muchos ejemplos de la mezcla de los medios de comunicación social y la política [20].

Este y otros casos se utilizan para apoyar el supuesto potencial de los medios de comunicación social para empoderar a los ciudadanos y fomentar la democracia y para ilustrar la viabilidad de medir el pulso de la opinión pública de los medios de comunicación social. Sin embargo, también existen contraejemplos preocupantes donde los medios sociales son utilizados como herramientas de represión. Sobre esta luz, uno podría pensar que los medios sociales no son buenos ni malos. Desafortunadamente, no son neutrales; Las redes sociales son productos del capitalismo comunicativo y su objetivo no es impulsar la acción política, sino mercantilizar la comunicación individual y monetizarla. Ciertamente, los operadores de medios sociales cuidan los intereses de sus usuarios -incluyendo la libertad de expresión, pero sólo en la medida en que no afectan a sus inversores las leyes bajo las cuales operan, no importa que sean justas o no. Facebook, por ejemplo, puede reclamar "Je suis Charlie" en los EE.UU. en un día y prohibir las páginas en Turquía sobre la base de la blasfemia del otro día.

¿Son los medios sociales un ámbito para la deliberación democrática? La deliberación es crucial en la democracia moderna. Sin embargo, aunque "el diálogo es preferible a la violencia, y un buen diálogo es preferible a un diálogo pobre", no todas las conversaciones califican para la deliberación democrática. La deliberación democrática apropiada asume que los ciudadanos son participantes iguales, los puntos de vista opuestos no sólo son aceptados, sino alentados, y el objetivo principal es lograr "un consenso racionalmente motivado". Desafortunadamente, existen fuertes argumentos en contra de que las discusiones de los medios sociales sean deliberaciones de este tipo.

Para empezar, no todos los usuarios de redes sociales son iguales. De hecho, las élites políticas, empresariales y de medios han "colonizado" los medios sociales. Los usuarios son los actores centrales en las redes sociales políticas.

Respecto a la diversidad ideológica, los usuarios de redes sociales no están aislados en cámaras de eco. De hecho, tienen cierto grado de exposición a las ideas transversales, incluso los usuarios que son claramente partidarios y la interacción entre los usuarios con ideas opuestas no es infrecuente. Sin embargo, la mayoría de los usuarios de redes sociales prefieren evitar tales discusiones, y cuando se encuentran con argumentos conflictivos, no los propagan dentro de su red. Por otra parte, la homofilia política es una fuerza que teje y desteje redes sociales.

Finalmente, cuando ocurren discusiones políticas, no son deliberaciones racionales y democráticas por una serie de razones:

- La información política en los medios sociales carece generalmente de la calidad y de argumentos fuertes, es generalmente incoherente y altamente de opinión.

- Los usuarios de las redes sociales tienen una propensión al humor y la ridiculez en los puntos centrales de la discusión política.

- Los debates en vivo, son cada vez más comunes, pero no son verdaderas deliberaciones. Son discusiones guiadas indirectamente impulsadas por la agenda establecida por los medios de comunicación y los actores políticos. Por otra parte, no es infrecuente en esas situaciones que los usuarios, tales como cuentas oficiales de partidos o de candidatos, intentan refutar las críticas de los usuarios regulares.

- Por último, los medios sociales son los más adecuados para la "deliberación intermitente", en la que los usuarios no están hablando entre sí, sino publicando mensajes para que todos los demás lean. Eso no es "deliberación pública sino deliberación en público".

Medios sociales y opinión pública

Debemos distinguir la opinión pública como el resultado colectivo de la deliberación racional sobre temas de interés común y la opinión pública como los resultados agregados de encuestas administradas a una muestra de una población dada. Como se mencionó anteriormente, la primera variedad no existe actualmente en las redes sociales. Sin embargo, en los medios sociales rebosan mensajes de opinión y ese material se destila en la segunda variedad de la opinión pública. Desafortunadamente, esa clase de opinión pública sufre de algunas debilidades.

- Cada mensaje social se considera igualmente válido independientemente de su procedencia (es decir, usuario regular o de élite, spammer, cuenta automatizada, etc.), o la certeza de que una cantidad sustancial de ellos es engañosa, incluso manipuladora.

- Sólo se pueden procesar los datos observables y, por lo tanto, no se ponderan las abstenciones. Por lo tanto, al estudiar la opinión pública en los medios sociales estamos observando la opinión de una minoría muy extrovertida. Esto, unido al efecto "espiral del silencio", debería ser una preocupación importante al extraer la opinión de los medios de comunicación social.

- Los usuarios de redes sociales no son monolíticos, incluso los grupos ideológicos supuestamente homogéneos reaccionan de manera diferente a diferentes temas. La opinión de los medios sociales sale de la opinión pública y lo hace de manera diferente dependiendo del tema: a veces, la opinión de los medios sociales es muy liberal, mientras que otros son más conservadores.

- Además, los usuarios de redes sociales no son una muestra aleatoria de la población: los hombres, los jóvenes y las personas urbanas están sobrerepresentados. Dado que todas esas características son importantes con respecto a las elecciones políticas, tal no aleatoriedad de los medios de comunicación social es otro problema.

3.5 Combinando Fortalezas, Emociones y Polaridades para Apoyar el Análisis de Sentimientos de Twitter

El análisis del sentimiento de Twitter o la tarea de recuperar automáticamente los comentarios de los tweets ha recibido un creciente interés de la comunidad que estudia la minería de datos. Esto se debe a su importancia en una amplia gama de campos como los negocios y la política.

La gente expresa sentimientos sobre temas específicos o entidades con diferentes intensidades, donde estos sentimientos están fuertemente relacionados con sus sentimientos y emociones personales.

Se han propuesto una serie de métodos y recursos léxicos para analizar el sentimiento a partir de textos de lenguaje natural, abordando diferentes dimensiones de opinión.

En este documento se propone un enfoque para impulsar la clasificación de sentimientos de Twitter utilizando diferentes dimensiones de sentimiento como características de alto nivel.

Se combinan aspectos como la fuerza de opinión, la emoción y los indicadores de polaridad, generados por los métodos y recursos de análisis de sentimientos existentes.

Esta investigación muestra que la combinación de las dimensiones del sentimiento proporciona una mejora significativa en las tareas de clasificación del sentimiento de Twitter como la polaridad y la subjetividad.

Las herramientas de análisis de sentimiento se centran en diferentes ámbitos dentro de las opiniones. Aunque estos alcances son muy difíciles de categorizar explícitamente, este estudio propone las siguientes categorías:

- **Polaridad**: Estos métodos y recursos apuntan a extraer la información de polaridad de un pasaje. Los métodos orientados a la polaridad normalmente devuelven una variable categórica cuyos posibles valores son positivos, negativos y neutros. Por otro lado, los recursos léxicos orientados a la polaridad están compuestos por listas de palabras positivas y negativas.

- **Emoción**: Métodos y recursos centrados en la extracción de emociones o estados de ánimo de un pasaje de texto. Un método orientado a la emoción debe clasificar el mensaje a una categoría emocional como la tristeza, la alegría, la sorpresa, entre otros. Los recursos léxicos orientados a la emoción deben proporcionar una lista de palabras o expresiones marcadas de acuerdo a diferentes estados emocionales.

- **Fuerza**: Estos métodos y recursos proporcionan niveles de intensidad de acuerdo con una dimensión de sentimiento determinado que puede tener una polaridad o un alcance emocional. Los métodos orientados a la fuerza devuelven diferentes puntuaciones numéricas que indican la intensidad o la fuerza de una dimensión de opinión expresada en un pasaje de texto. Por ejemplo, las puntuaciones numéricas que indican el nivel de positividad, negatividad u otra dimensión emocional. Los recursos léxicos orientados a la fuerza proporcionan una lista de palabras de opinión junto con puntuaciones de intensidad con respecto a una dimensión de opinión. [21]

Capítulo 4

Marco Teórico

La presente memoria se pudo realizar dado que se obtuvo un conjunto de información desde Twitter usando la API que este sistema facilita de manera abierta y gratuita para que cualquier persona extraiga información libremente.

Esta es una práctica frecuente en las diferentes redes sociales, dando de esta forma la oportunidad a sus usuarios de beneficiarse de los datos almacenados y así realizar análisis en diferentes ámbitos.

En la presente memoria se extrajo información relevante para hacer análisis relacionado con las elecciones presidenciales del año 2013 en nuestro país.

4.1 Análisis de texto

El análisis de texto consiste en la utilización de diversos métodos algunos semánticos y otros de sintaxis, en el primero se analizan palabras solas y en el segundo conjunto de palabras y su orden.

El término semántica se refiere a los aspectos del significado, sentido o interpretación de signos lingüísticos como símbolos, palabras, expresiones o representaciones formales. En principio las expresiones del lenguaje formal o de una lengua natural admiten algún tipo de correspondencia con situaciones o conjuntos de cosas que se encuentran en el mundo físico o abstracto que puede ser descrito por dicho medio de expresión.

El análisis sintáctico es el análisis de las funciones sintácticas o relaciones de concordancia y jerarquía que guardan las palabras cuando se agrupan entre sí en forma de sintagmas, oraciones simples y oraciones compuestas de proposiciones. Como no está muchas veces claro el límite entre la sintaxis y la morfología a estos respectos, especialmente según el tipo de lengua de que se trate, también se suele denominar análisis morfosintáctico, aunque esta denominación se suele reservar también para un análisis más profundo y detenido.

Su estudio es importante, ya que de un correcto análisis sintáctico depende a menudo la interpretación y comprensión de los textos, especialmente de los documentos problemáticos en legislación, política o tecnología (el llamado procesamiento

de lenguajes naturales). Diversas corrientes de la lingüística han propuesto a su vez diversos métodos de análisis; el que se enseña en las escuelas es el de la gramática tradicional, algo influido por el Estructuralismo.

4.2 Análisis de sentimiento

Se puede determinar el estado de ánimo de una persona analizando su información emitida en forma verbal, corporal y escrita, a este método se denomina análisis de sentimiento.

La presente memoria realiza análisis de sentimiento a los textos generados en Twitter por un grupo de usuarios determinados y busca determinar la emoción que se genera al realizar un determinado comentario en Twitter.

En el contexto de esta memoria se busca determinar frente a la política chilena la polaridad de los mensajes de Twitter y con esto saber de los candidato presidenciales si están siendo bien recibidos políticamente o no.

4.2.1 Como funciona el análisis de sentimiento

Un algoritmo de clasificación de datos conocido como Support Vector Machine (SVM) es usado para detectar emociones, basándose en "Thumbs up" en información de reseñas de películas, esto implica que los usuarios con una simple acción de "pulgares arriba" determinan si la película es de su agrado o no.

Lo interesante es que las técnicas avanzan rápidamente y ya no es necesario determinar el sentimiento basado en una acción del usuario como "Pulgares arriba" sino que solo basta analizar sus comentarios en las redes sociales, así un comentario se clasifica en un hecho o una opinión.

Al hablar de hechos nos referimos a expresiones objetivas que no son emociones. Las emociones que son subjetivas (emociones y juicios de valor). Esto se puede plantear en forma positiva o negativa, la clasificación subjetiva o objetiva se define como polaridad.

La minería de opiniones usa varias técnicas, la mayoría consisten en detectar palabras positivas y contarlas, al sumarlas se puede determinar si el texto es positivo o negativo.

La herramienta Sentiwordnet consiste en una base de 4 palabras clasificadas (positivo y negativo PN; subjetivo y objetivo SO), cada palabra tiene 3 puntajes; positivo, negativo u objetivo. [22]

Luego del análisis anterior también se podría buscar determinar si la persona esta enojada, triste o alegre usando otras técnicas. [23]

4.2.1.1 Métricas de polaridad y objetividad

El análisis de algoritmos en el mundo de las métricas de sentimientos pueden ser usados en toda magnitud de información, ya sea de unos pocos datos recabados de bases de información como emoticones donde se analizan sus características como positivas o negativas dependiendo el caso de cada uno asignándole un puntaje positivo o negativo hasta grandes volúmenes de datos en bases léxicas como es Sentinet. La forma como trabaja Sentinet es relacionando cada palabra con su definición y uso las que han sido extraídas de Wordnet, a esto se le asocia un puntaje positivo o negativo desde 0 a 1, por ejemplo si usamos la palabra "correcto" esta tendría asociado un puntaje de 1 punto positivo y un puntaje de 0 puntos negativos, por otro lado si usamos la palabra "incorrecto" tiene asociado un puntaje de 0 puntos positivos y 1 punto negativo. Existen otras palabras que para el hecho de este análisis no corresponde calificarlas con ningún puntaje ya que son palabras que no expresa emocionalidad como por ejemplo las palabra "casa" o "automóvil".

Opinion Finder Lexicon (OPF)

Es un recurso léxico basado en la polaridad creado por Wilson [25]. Es una extensión del conjunto de datos de preguntas y respuestas multi-perspectiva (MPQA), que incluye frases y sentencias subjetivas. Un grupo de anotadores humanos etiquetó cada oración de acuerdo a las clases de polaridad: positivo, negativo, neutro. Luego, se llevó a cabo una fase de poda sobre el conjunto de datos para eliminar las etiquetas con un bajo acuerdo. Así, se consolidó una lista de frases y palabras únicas, con sus etiquetas de polaridad. En este estudio se consideraron las palabras simples (unigramas) etiquetadas como positivas o negativas, que corresponden a una lista de 6.884 palabras en inglés. Se extrae de cada tweet dos características relacionadas con el Opinion Finder lexicon, Opinion Finder Positive Words (OPW) y Opinion Finder Negative Words (ONW), que son el número de palabras positivas y negativas del tweet que coincide con el léxico del Buscador de Opinión, respectivamente.

AFINN Lexicon

Este léxico está basado en las Normas Afectivas de Palabras en Inglés (ANEW). ANEW proporciona calificaciones emocionales para un gran número de palabras en inglés. Los análisisson calculados según la reacción psicológica de la persona a una palabra específica, siendo "valencia" el valor más útil para el análisis del sentimiento. "Valence" se extiende en la escala agradable-desagradable. ANEW fue lanzado antes de la subida de microblogging y por lo tanto, muchas palabras de argot usadas comúnmente en medios sociales no fueron incluidos. Teniendo en cuenta que hay evidencia empírica sobre las diferencias significativas entre las palabras de microblogging y el lenguaje utilizado en otros dominios una nueva versión de ANEW se requiere. Inspirado en ANEW, Nielsen creó el léxico AFINN, que se centra más en el lenguaje utilizado en las plataformas de microblogging. La lista de palabras incluye argot y palabras obscenas como sinónimos y jerga web. Las palabras positivas se califican de 1 a 5 y las palabras negativas de -1 a -5, por lo que este léxico es útil para estimar la fuerza. El léxico incluye 2,477 palabras en inglés. En modo de ejemplo, se extraen de cada tweet dos características relacionadas con el léxico

AFINN, AFINN Positivity (APO) y AFINN Negativity (ANE), que son la suma de las calificaciones de palabras positivas y negativas del tweet que coincide con el léxico AFINN, respectivamente.

4.2.1.2 Métricas de análisis de sentimiento

Para entender que tan positivo, neutro o negativo es un mensaje existen algunas métricas basadas en el análisis de sentimiento. Estas métricas son independiente del método de análisis en el que fue calculado el puntaje asociado a cada palabra y su finalidad es conocer la polaridad y objetividad final del mensaje.

$$N_t + N_p + N_o + N_n + N_? = 0$$

donde:

- N_t: Número total de palabras
- N_p : Número de palabras positivos
- N_o : Número de palabras neutrales
- N_n : Número de palabras negativos
- $N_?$: Número de palabras no clasificada

Las palabra no clasificada puede contar como objetiva o simplemente existir en una categoría específica para las palabras no clasificadas. Como se realizaron los experimentos de esta Memoria, solo se enfoco en la polaridad de las palabras, para el análisis de sentimiento de un Tweet sin considerar la emoción o polaridad de éstas.

4.2.1.3 Ratios de sentimiento

Con los conteos anteriores se pueden establecer diferentes ratios, que nos permiten determinar el grado de polaridad u objetividad, e incluso pueden ser utilizada para medir el grado de popularidad dependiendo de la muestra de datos.

- Ratio de popularidad positiva

$$r_{pp} = \frac{N_p + N_o}{N_p} = PP$$

- Ratio de popularidad negativa

$$r_{ppn} = \frac{N_t + N_o}{N_t}$$

- Diferencia de sentimientos

$$Sent_{diff} = \frac{N_p - N_n}{totales}$$

- Diferencia de sentimientos porcentual y normalizada

$$I_{sent} = (Sent_{diff}/2 + 0,5) * 100$$

4.3 Métricas en redes sociales

Existen muchos tipos de métricas sociales, las más sencillas son asociadas a la red del individuo, a esta métrica se le denomina network por su nombre en inglés. Sin embargo previo a explicar estas métricas sociales corresponde explicar la red social de Twitter

4.3.1 Twitter

Según wikipedia actualmente Twitter cuenta con más de 500 millones de usuarios, generando 65 millones de tweets al día y maneja más de 800.000 peticiones de búsqueda diarias.

Twitter es una red social de microblogging donde los usuarios registrados divulgan públicamente mensajes. La red permite enviar mensajes de texto plano de corta longitud, con un máximo de 140 caracteres, llamados tweets.

Las etiquetas (# o hashtags) originalmente no formaba parte de la estructura de Twitter. La propia necesidad de los usuarios los llevó a utilizar etiquetas para clasificar la temática de los tweets.

Para poder seguir con mayor facilidad un tema, conversación o discusión existen las etiquetas, que puede crear o utilizar cualquier usuario. Sólo hay que incluir el símbolo # delante de palabras, expresiones o temas. Si alguien busca ese término en el cuadro de búsqueda, los tweets que contengan esa etiqueta aparecerán en los resultados.

4.3.1.1 Características de un mensaje en Twitter

Un tweet contiene un autor, un mensaje, y un nivel de propagación. También se asume que trata de al menos un tópico, debido a que ha sido escrito con un propósito. Como se dijo anteriormente el mensaje puede tener solo hasta 140 caracteres, lo que desafía la creatividad de los usuarios para expresar sus opiniones con tan poco espacio.

El desarrollo de twitter y de las redes sociales a permitido también en el desarrollo de la industria del marketing digital la generación de nuevos puestos de trabajo en las empresas como es el denominado "Community Manager" quien sería un encargado de gestionar las redes sociales de la organización con fines muy diversos desde el hacerse cargo de la difusión de comunicados y noticias hasta la gestión de quejas y reclamos de los usuarios.

4.3.1.2 Network

Twitter es una red unidireccional, esto quiere decir que si tu sigues a una persona, ella no necesariamente te seguirá a ti. En cambio en otras redes como Facebook el ser amigo es bidireccional ya que al solicitar la amistad de alguien de inmediato las dos personas quedan conectados y pueden ver sus comentarios.

La métrica que determina con cuantas personas se vincula un individuo es muy importante en redes sociales. En twitter esta métrica se divide en dos, los usuarios a quienes yo sigo (followees) y los que me siguen (followers), básicamente podemos deducir que tan influenciador es la persona solo viendo la cantidad de followers que tiene.

Como se comento anteriormente en esta memoria, la información de twitter es de libre acceso ya sea por medio de la cuenta de la persona o por medio de la API que la plataforma facilita para extraer los datos.

4.3.1.3 Crecimiento

El crecimiento es el aumento de followers en una ventana de tiempo determinado, se calcula de la siguiente manera:

$$Crecimiento = \frac{followers}{tiempo}$$

4.3.1.4 Reach

Klout es un Servicio Web que mediante un índice llamado Klout Score mide el grado de influencia de una persona o una marca en las Redes Sociales. Para determinar el Klout Score de una persona el Servicio Web analiza más de 400 parámetros distintos de las 7 Redes Sociales más importantes y se asigna una puntuación entre 1 y 100 a los usuarios.

El promedio de los usuarios de Klout es de 40 y se considera como un Influenciador a aquellas personas con un índice alto por encima de la media, por ejemplo, los que pertenecen al grupo del 5 % de usuarios con un valor superior a 60 están considerados como los más influyentes.

En Facebook cuando un usuario realiza un comentario y sus amigos le dan "me gusta" o recibe un post en su muro, ese mensaje es visto por sus amigos.

En Twitter cuando un mensaje es de un usuario es retwiteado lo ve toda la red de la persona que lo reenvió (retwiteo). Reach es el número de personas que son alcanzadas por una acción.

4.4 Segmentación

Para llevar a cabo los experimentos fue necesario agrupar los datos. Se segmentó por:

- **Candidato**: Esta segmentación estaba en la Base de Datos.
- **Periodos de tiempo**: Esta segmentación se aplicó en los análisis que se realizaron en los eventos de la línea de tiempo.
- **Polaridad del Tweet**: Esta segmentación se realizó previamente a los análisis.

4.5 Ley de Zipf

Fue formulada en la década de los cuarenta por el lingüista de Harvard George Kingsley Zipf, y afirma que un pequeño número de palabras son utilizadas con mucha frecuencia, mientras que frecuentemente ocurre que un gran número de palabras son poco empleadas. Como una rama más de la hermenéutica, la Ley de Zipf sirve, básicamente, para contar palabras.

George Kingsley Zipf (1902-1950) fue un lingüista y filólogo estadounidense que aplicó el análisis estadístico al estudio de diferentes lenguas. Sus descubrimientos sobre el lenguaje, le han llevado también a ser uno de los autores más citados que se han importado al área de la Información y Documentación científica, especialmente en el área de Recuperación de información y la indización automática.

Para usar esta ley, debemos tomar un texto con más de 5.000 palabras y, entonces, se calcula cuántas veces aparece una palabra concreta. Se ordena la tabla de palabras de más a menos frecuente. El orden en que aparece cada palabra en esta lista ordenada se llama "rango".

En el idioma español, por ejemplo, las palabras que encabezan este rango siempre son artículos y preposiciones. En un texto en inglés, la palabra que casi siempre estará en el primer lugar será "the". Si "La" tiene rango 1, palabras como "oxímoron" o "escible" tendrán un rango altísimo (sobre todo "escible", que según la RAE es una palabra ya en desuso).

En términos matemáticos, la ley se expresa del siguiente modo: el número Y de veces que aparece una palabra es inversamente proporcional a su rango X de forma que:

- $P_n = (\frac{1}{n})^a$

Otra forma de calcular la Ley de Zipf es contar cuantas veces aparece una palabra y dividirla entre el número total de palabras del texto.

Pero esto no ha hecho más que empezar. Gracias a la digitalización de contenidos, en pocos años es más que posible que dispongamos en Internet de todos los libros y textos que aparezcan en el mercado, como una gigantesca y multiforme Biblioteca de Alejandría formada de ceros y unos.

Cuando esto suceda, se podrá aplicar la ley de Zipf de manera global, idioma por idioma, y generar toda clase de estadísticas. Y más aún: se podrán formular nuevas leyes. Por ejemplo, una ley para calcular la frecuencia de repetición de metáforas, frases, expresiones y hasta, por qué no, grado de creatividad.

Capítulo 5

Recursos disponibles

5.1 Datos: Datos API Twitter

Para la captura de datos, Twitter posee una interfaz de programación de aplicaciones a la que se realizan llamadas para obtener los datos necesarios.

La extracción de datos se realizó sobre la red social de Twitter, por lo que se debe establecer una conexión directa con su API. Hay que reseñar que para la conexión a la API de Twitter será necesaria la configuración de un protocolo de conexión segura. Si la API de Twitter se actualiza y cambia también lo hace la forma de acceder a ella, se tendría que modificar la aplicación que extrae los datos.

Para la conexión con la API de Twitter es necesario un paso previo de autentificación para poder interactuar con ella. Una vez realizada esta conexión, se podrá usar su servicio para recuperar los datos. La extracción de los datos usando la API de Twitter debe considerara la búsqueda en la base de datos a partir de un rango de fechas.

Se deben aceptar y aplicar todas las políticas de desarrollo del uso de la API de Twitter, las infracciones de políticas también se consideran violaciones del Acuerdo de Desarrollador.

Para hacer uso de este framework será necesario crear el objeto de TwitterAPI con las key generadas por la API de Twitter. Para esto será necesario registrar una aplicación (ésta se conectará al servicio API de Twitter) dentro de Twitter, para que a partir de estas claves puedas identificarte unívocamente, y así poder realizar consultas a su API. Una vez generado el objeto, podremos hacer llamadas a las distintas funciones que nos ofrece la API de Twitter. [28]

5.2 Herramientas de análisis usadas

Python

Python Software Foundation (PSF) es una corporación sin fines de lucro que tiene los derechos de propiedad intelectual detrás del lenguaje de programación Python. Para este proyecto se utilizó la version de Python 2.7, incluyendo las siguientes librerías: [29]

- **NLTK**: Esta librería se usa para trabajar con los análisis semánticos, proporciona interfaces para recursos léxicos, como WordNet, con un conjunto de bibliotecas de procesamientos de texto para clasificación, tokenización, derivado, etiquetado, análisis y de razonamiento semántico.

- **MySQLdb**: Esta librería permite la conexión a la Base de Datos MySql desde la aplicación Python.

- **CSV**: Esta librería permite la extracción de datos desde la Base de Datos MySql a archivos de extención CSV. Un archivo CSV (de Valores Separados por Comas) es un tipo de documento que representa los datos de forma parecida a una tabla, es decir, organizando la información en filas y columnas. En un archivo CSV los datos de las diferentes columnas son separados, habitualmente, por un signo de puntuación (una coma, un punto y coma, etc.) u otro carácter que actúe como separador. Sin embargo, las diferentes filas suelen separarse por un salto de línea. Además, muchas veces los datos pueden ir precedidos por un encabezado con los nombres de campos o identificadores de columnas

- **RE**: Esta Librería de Python permite realizar una de las operaciones más comunes que es la búsqueda de una subcadena; ya sea para obtener su posición en el texto o simplemente para comprobar si está presente. Al buscar direcciones de correo electrónico, números de teléfono, validar campos de entrada, o una letra mayúscula seguida de dos minúsculas y de 5 dígitos entre 1 y 3; es necesario recurrir a las Expresiones Regulares, también conocidas como Patrones.

Circos

Es un paquete de software para la visualización de datos e información. Visualiza los datos en una disposición circular. Esto hace a Circos ideal para explorar las relaciones entre los objetos o posiciones. Hay otras razones por las que una disposición circular es ventajosa, no menos importante es el hecho de que es atractivo.

Circos es ideal para crear infografía con calidad de publicación y las ilustraciones con una alta proporción de datos. Las capas y simetrías de los gráficos son muy agradables. Usted tiene el control fino de cada elemento de la figura para adaptar sus puntos de enfoque y los detalles para su óptima exposición. [30]

LIBREOFFICE CALC

Es la hoja de cálculo de LibreOffice, quienes la usa por primera vez les resulta intuitiva y fácil de aprender. Especialistas en minería de datos, profesionales y contadores apreciarán la amplia gama de funciones avanzadas. CALC es una hoja de calculo similar a Microsoft Excel.

En esta memoria CALC fue usada para agrupar, clasificar, contabilizar y desplegar la información. [32]

5.3 Datos disponibles y modelo de datos

MySql

Es la base de datos de código abierto más popular del mundo. Con su rendimiento, confiabilidad y facilidad de uso comprobados, MySQL se ha convertido en la principal opción de base de datos para aplicaciones basadas en la Web, utilizada por propiedades web de alto perfil como Facebook, Twitter, YouTube, y los cinco principales sitios web*. Además, es una alternativa extremadamente popular como base de datos integrada, distribuida por miles de ISV y OEM.

En esta memoria se realizan los análisis almacenando la información en una Base de Datos MySql. [31]

El modelo de la Base de Datos usado en este análisis fue creado en un proceso anterior a esta memoria, por lo que veremos el modelo como proveedor de información para el análisis realizado y no se profundizará en los detalles de la generación del modelo de datos.

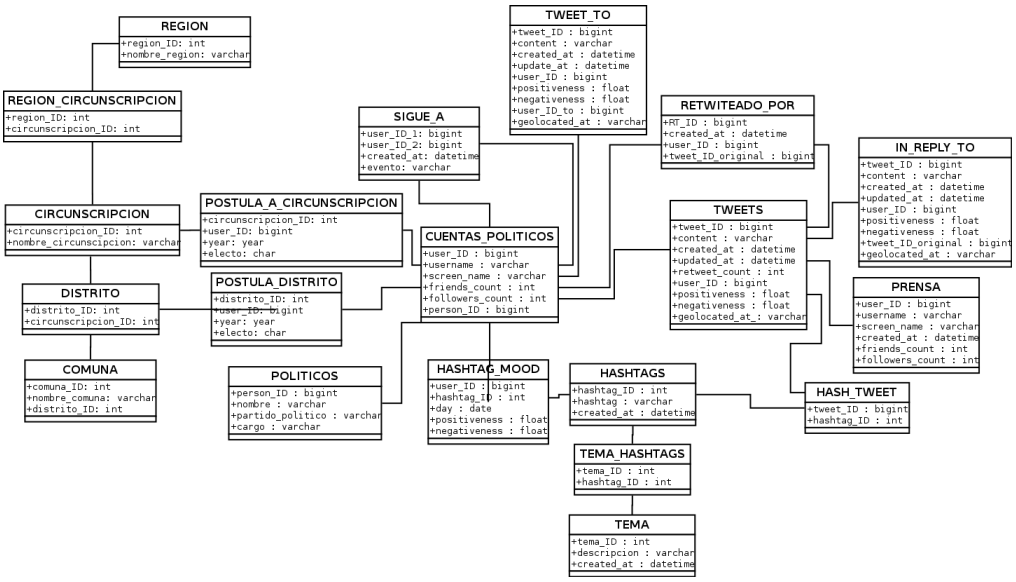

Figura 5.1: Modelo de Datos Inicial

En un principio la captura y modelado de los datos se realizó con varias finalidades, como por ejemplo observar la realidad política completa del país. Como prueba piloto, la base de datos se alimentó con datos de todos los políticos que participaran en el congreso y esto se abortó devido a que el volumen de información que generaba no era significativo. La inteción de esto era poder tomar todo dato relevante que pudiera entregar información, lamentablemente la cantidad de datos no fue suficiente. Una característica que se quizo capturar fue la segmentación de los datos por geolocalización, pero no todos estos la contenían (más de un 60 % de los datos sin geolocalización), por lo tanto, también se abortó.

Devido a lo descrito anteriormente se transformó la base de datos en otro modelo llamado Observatorio, donde se caracterizó y organizó la información, con esto se facilitó el trabajo para el estudio que se realizó en esta memoria. Respecto al modelo usado, podemos indicar que se usaron tablas para almacenar información relevante para este análisis:

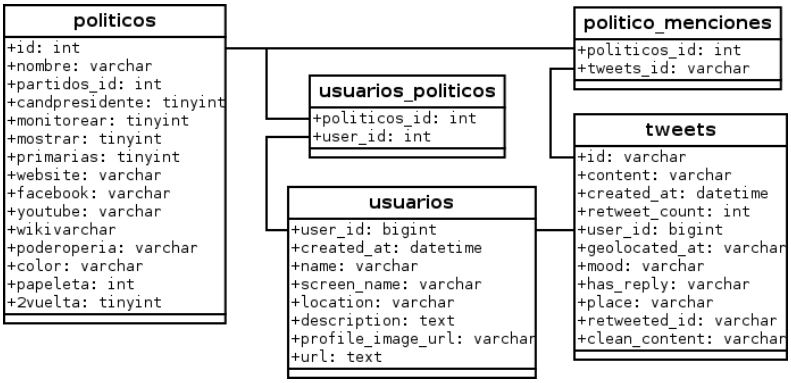

Figura 5.2: Modelo de Datos Final

1. **Tweets**: Contiene la información de todos los Tweets.
2. **Politico_ menciones**: Tabla de relación entre políticos y Tweets
3. **Users**: Cuentas de usuarios de Twitter.
4. **Politicos**: Candidatos presidenciales de Chile.

Este modelo se generó relacionando manualmente las cuentas de Twitter con el candidato presidencial correspondiente, asignando estas cuentas como oficiales del político. Por otro lado, los seguidores se asignaron tomando las cuentas de usuarios que seguían a alguna cuenta oficial de los candidatos.

5.3.1 Targets

Para esta memoria se obtienen datos de candidatos presidenciales por medio de la API de Twitter, para esto se debe identificar el usuario de Twitter que inicia con el carácter @, este es el identificador único del usuario, este nombre está compuesto por caracteres.

La API de Twitter necesita el nombre de usuario para poder obtener los tweets. Se extrajeron todos los tweets en los que estaban referenciados los candidatos a través de sus cuentas oficiales.

Para el uso de la API de Twitter se debe tener en cuenta la documentación que nos proporciona a la hora de hacer búsquedas:
`https://dev.twitter.com/docs/api/1.1/get/search/tweets`

En esta documentación, podemos ver que si queremos obtener información de Twitter tenemos que usar la siguiente dirección URL:

```
https://api.twitter.com/1.1/search/tweets.json
```

Se debe llamar la URL con el parámetro requerido "q", del inglés "query" cuyo valor va a ser lo que queramos buscar. También tenemos unos parámetros opcionales donde podemos hacer la búsqueda más restrictiva. Una vez que sabemos como obtener los datos de la Api, se construye una función que devuelva una estructura JSON con los datos de la API de Twitter en función de una búsqueda y el número de resultados que queremos obtener

5.3.2 Tweets

Los tweets se almacenan en una BD luego que son extraídos desde la API de Twitter, Los tweets cuentan con una estructura bien definida que nos sirve para filtrar, ordenar y manipular la información. Rafik Krikorian, VP, Engineering – Twitter entre julio de 2009 hasta julio de 2014 , San Francisco, California [36], provee una imagen grafica de la estructura de un Tweet describiendo cada componente.

En Anexo Imagenes se muestra el detalle de la estructura de un Tweet.

5.3.3 Followers

Es importante para algunas tablas y gráficos de este análisis determinar los seguidores de los candidatos a la presidencia en las elecciones del año 2013. Por medio de la API de Twitter es factible obtener la información de los seguidores pero con ciertas limitaciones y restricciones que da la plataforma con el objetivo de no saturar la infraestructura, por esta razón la extracción de los seguidores es lenta

5.3.4 Followees

La lógica existente detrás del modelo de followees es idéntica a la de los followers . En este modelo un target tiene varios followees y un follower le pertenece a un target.

5.4 Preparación de los datos

Otro aspecto importante fué la preparación de los datos. Primero se segmentó toda la base de datos en las fechas que se consideraron relevantes para el estudio y se eliminaron todos los registros que estaban fuera del rango de fechas a estudiar.

También se limpia la BD para extraer el "ruido" y dejar solo la información relevante, por ejemplo en el mismo periodo del estudio se estrenó una película en los EEUU, donde se capturaron datos relacionados a ésta y por error se incluyeron en el Observatorio, por lo que los datos se vieron alterados justo en la fecha del estreno de dicha película. El \ruido" se extrajo generando un script donde se buscaron todos los Tweets que contenían datos relacionados con actores, y se eliminaron definitivamente para no tener información errada.

Los siguientes serian los pasos a seguir para realizar un procesamiento de texto aplicando la LEY de ZIPF.

1. **Document**
   Se trata del texto o la data rescatada su posterior análisis o procesamiento.

2. **Estructure recognition**
   Corresponde al reconocimiento de la estructura del texto y la limpieza y ordenamiento de este.

3. **Accent, Spacing, etc.**
   En esta etapa se realiza la limpieza de los caracteres que generan ruido en el análisis del texto.

4. **Stopwords**
   Corresponde a la limpieza de las palabra que no generan información relevante para el análisis.

5. **Noun groups**
   Un grupo de palabras basado en un sustantivo o pronombre, eje: En la frase: «Él puso la botella de vino en la mesa de la cocina», «Él», «la botella de vino» y «la mesa de la cocina» son todos Noun groups.

6. **Stemming**
   Es un método para reducir una palabra a su raíz.

7. **Automating or manual indexing**
   Corresponde a una indexación del texto en general pero ordenado en base a las necesidades del análisis, luego de haber realizado todos los pasos anteriores.

5.5 Lo que pasó en las elecciones

Para efectos de esta memoria, nos centraremos en el estudio de los datos obtenidos desde Twitter para las elecciones presidenciales de Chile del año 2013, dicha elección presidencial para el período 2014-2018, se realizó el 17 de noviembre de 2013, esta elección no pudo ser resuelta en una primera vuelta dado que los porcentajes obtenidos por la candidata que obtuvo mayoría no fueron suficientes según la ley para adjudicarle la presidencia. La segunda vuelta electoral tuvo lugar el 15 de diciembre, y dio como vencedora a Michelle Bachelet.

Como antecedente adicional se debe indicar que esta elección se desarrolló bajo el régimen de inscripción automática y voto voluntario para los votantes. Además los partidos políticos pudieron someterse al sistema de primarias voluntarias, dichas primarias fueron organizadas por el Servicio Electoral (SERVEL).

A esta elección presidencial se presentaron nueve candidatos una cifra poco usual en la historia electoral de nuestro país. [24]

Figura 5.3: Candidatos Presidenciales

En la primera vuelta, Michelle Bachelet de la Nueva Mayoría obtuvo el 46,70 % de los votos válidamente emitidos y se enfrentó en segunda vuelta a la candidata de la Alianza, Evelyn Matthei, quien llegó al 25,03 %. En segunda vuelta, Michelle Bachelet alcanzó el 62,16 %, mientras Evelyn Matthei logró el 37,83 %, con una participación del 41,98 % de los electores. Además fue la primera vez que dos mujeres se enfrentaron en este tipo de elecciones. Con este resultado, Michelle Bachelet se convirtió en la primera mujer reelecta en la historia de Chile.

5.5.1 Primarias

En este proceso electoral se desarrollaron primarias tanto en la Nueva Mayoría como en la Alianza.

La Concertación crea un nuevo movimiento de centro e izquierda para su proceso de primarias denominado "Nueva Mayoría". En marzo de 2013, Michelle Bachelet anunció su retorno al país. Al proceso de primarias de la centroizquierda se sumó su ex ministro, el independiente Andrés Velasco, mientras el Partido Demócrata Cristiano levantó la candidatura de Claudio Orrego y el PRSD la de su presidente, José Antonio Gómez. Otros anunciaron participar directamente en la primera vuelta, ese fue el caso de Marco Enríquez-Ominami, Marcel Claude, Alfredo Sfeir y Roxana Miranda. Por ultimo existieron algunos candidatos que desearon participar como independientes destacando Tomás Jocelyn-Holt y Franco Parisi.

Cuadro primarias Nueva Mayoría:

#?	Candidato	Partido		Apoyo político	Votos	% Pacto	% Total
A1	Michelle Bachelet Jeria	PS	PS	PS-PPD-MAS-PCCh-IC	1 565 269	73,07 %	53,06 %
A2	José Antonio Gómez Urrutia	PRSD	PRSD	Partido Radical Socialdemócrata	108 365	5,06 %	3,67 %
A3	Claudio Orrego Larraín	PDC	PDC	Partido Demócrata Cristiano	189 752	8,86 %	6,43 %
A4	Andrés Velasco Brañes	Ind	Ind	Independiente	278 684	13,01 %	9,45 %

Figura 5.4: Resultados Primarias Nueva Mayoría [33]

La derecha (La Alianza) también organizo sus primarias entre el RN Andrés Allamand y Laurence Golborne, que era apoyado por la UDI. Una serie de errores llevaron a que, a pocos días de la inscripción de la primaria, la UDI quitara el apoyo a Golborne y decidiera apostar por un militante histórico del partido, Pablo Longueira.

Cuadro primarias Alianza:

B5		**Pablo Longueira Montes**		**UDI**	**Unión Demócrata Independiente**	**415 087**	**51,37 %**	**14,07 %**
B6		Andrés Allamand Zavala		RN	Renovación Nacional	392 915	48,63 %	13,32 %

Figura 5.5: Resultados Primarias Alianza [33]

5.5.2 Bajada de Pablo Longueira

En una sorpresiva decisión que se mantuvo hermética hasta el momento de su anuncio, el candidato Pablo Longueira renunció a la campaña presidencial, afectado por una depresión de la que no se tenía noticia ni en su entorno político más cercano. De este modo, quedó en suspenso la definición del candidato presidencial de gobierno, la cual concluyó pocos días después, cuando la coalición gobernante, luego de varios desencuentros, optará por Evelyn Matthei como su carta presidencial. La recientemente aprobada Ley de Primarias dejó a los partidos del oficialismo en libertad de acción para llevar un candidato unitario o más de uno.

5.5.3 Primera vuelta

Tras su victoria en las primarias, Michelle Bachelet se posicionó como la candidatura con más probabilidades de ganar las elecciones, apareciendo como victoriosa en la totalidad de las encuestas publicadas y centró su campaña en tres ejes principales: reforma tributaria, reforma educacional y nueva Constitución Política. La candidatura de la Nueva Mayoría potenció la imagen carismática de la ex presidenta, pero también dio un foco ciudadano a su campaña, titulada "Chile de todos". Esto generó críticas respecto a una candidatura "silenciosa" en que Bachelet se habría mantenido lejos de controversias y definiciones claras sobre su programa de gobierno.

La candidata oficialista Evelyn Matthei tuvo una difícil entrada en la campaña, luego de la renuncia de Pablo Longueira. Tras asegurar el apoyo de Renovación Nacional, Matthei lanzó su campaña con el lema "Ganemos juntos". La idea apuntaba a reavivar las posibilidades de victoria en una carrera, pero varios especialistas criticaron la falta de contenido en la idea y una sensación de desesperación. Tras el sorteo de las posiciones de votación, el lema fue cambiado a "Un 7 para Chile", haciendo un juego de palabras entre el número asignado a la candidata y la nota máxima de calificación en Chile, aunque el cambio también fue criticado por algunos especialistas.

Candidato	Partido		Coalición/Partido	Votos	%
Franco Parisi Fernández	Ind		Independiente	666 015	10,11 %
Marcel Claude Reyes	PH		Todos a La Moneda	185 072	2,81 %
Ricardo Israel Zipper	PRI		Partido Regionalista de los Independientes	37 744	0,57 %
Marco Enríquez-Ominami Gumucio	PRO		Si tú quieres, Chile cambia	723 542	10,98 %
Roxana Miranda Meneses	IGUAL		Partido Igualdad	81 873	1,24 %
Michelle Bachelet Jeria	**PS**		**Nueva Mayoría**	**3 075 839**	46,70 %
Evelyn Matthei Fornet	**UDI**		**Alianza**	**1 648 481**	25,03 %
Alfredo Sfeir Younis	ECOV		Partido Ecologista Verde Partido Ecologista Verde del Norte	154 648	2,34 %
Tomás Jocelyn-Holt Letelier	Ind		Independiente	12 594	0,19 %
Total de votos válidos				6 585 808	98,31 %
Votos nulos				66 935	0,99 %
Votos en blanco				46 268	0,69 %
Total de sufragios emitidos				**6 699 011**	**100 %**
Total de inscritos				13 573 088	Abstención: 50,64 %
100 % de las mesas escrutadas (TRICEL)					

Figura 5.6: Resultados Primera Vuelta [33]

5.5.4 Segunda vuelta

En la segunda vuelta presidencial efectuada el 15 de diciembre de 2013 Michelle Bachelet se impuso con un 62,17 % de los votos mientras que Evelyn Matthei logró un 37,83 % de los votos. Así Bachelet, la primera presidenta de la República cumple un nuevo hito histórico; se convirtió en la primera mujer reelecta en la historia de Chile. Asimismo, el alto porcentaje recibido por Bachelet en segunda vuelta (62,17 %) la coloca como la cuarta mayoría electoral más alta de la historia de las elecciones presidenciales chilenas, desde que existe el sufragio universal.

A pesar de esta alta mayoría electoral, cabe constatar que Michelle Bachelet obtiene menos votos que en su primera elección en 2006, cuando en la segunda vuelta frente a Sebastián Piñera obtuvo el 53,50 % de los votos (es decir, casi 9 puntos porcentuales menos) pero con aproximadamente 250 mil votos más.

Cuadro resumen segunda vuelta:

Candidato	Partido		Coalición/Partido	Votos	%
Michelle Bachelet Jeria	**PS**		**Nueva Mayoría**	**3 470 055**	62,17 %
Evelyn Matthei Fornet	UDI		Alianza	2 111 830	37,83 %
Total de votos válidos				5 581 885	97,97 %
Votos nulos				83 000	1,45 %
Votos en blanco				32 639	0,57 %
Total de sufragios emitidos				**5 697 524**	**100 %**
Total de inscritos				13 573 143	Abstención: 58,02 %
100 % de las mesas escrutadas (TRICEL)					

Figura 5.7: Resultados Segunda Vuelta [33]

Capítulo 6

Análisis Exploratorio

6.1 Análisis Léxico

6.1.1 Volumen, distribución de términos, ley de Zipf

Método Zipf Frecuencia de palabras: Para el análisis de palabras se utilizo el método Zipf en la rutina de nombre tweetsloop.py

Los siguientes son los pasos realizados en la rutina para la implementación usando el lenguaje de programación Python y la librería NLTK.

Pasos para el desarrollo del análisis 1:

1. **Importación de librerías**

   En el primer paso se importaron todas las librerías necesarias para utilizar las herramientas que estas proveen y realizar el procesamiento.

```
1  #!/usr/bin/python
2  import nltk
3  import re
4  import csv
5  import MySQLdb
6  from nltk.corpus.reader.plaintext import PlaintextCorpusReader
7  from nltk.tokenize import sent_tokenize, word_tokenize, TweetTokenizer
8  from nltk.tokenize import RegexpTokenizer
9  from nltk.stem.snowball import SnowballStemmer
10 from nltk.collocations import *
```

Figura 6.1: Importación de Librerías

2. **Conexión a la Base de Datos**

En esta etapa se realiza la conexión a la BD, especificando los datos para conectar al motor MySQL y tener acceso al "Observatorio" identificando el servidor, el nombre de la BD, usuario y contraseña.

```
75  db = MySQLdb.connect(host="localhost",      # your host, usually localhost
76                       user="root",           # your username
77                       passwd="**********",   # your password
78                       db="observatorio")     # name of the data base
```

Figura 6.2: Conexión Base de Datos

Como resultado se obtiene un objeto del tipo MySQLdb.connect(), el cual se utiliza para realizar consultas a la base de datos.

3. **Consulta a la Base de Datos**

Para realizar consultas a la base de datos, lo primero que se debe hacer es crear un objeto del tipo cursor(), el cual tiene las funcionalidades necesarias para poder realizar una consulta con la estructura gramatical definida, a un motor de base de datos determinado. La función execute() es la que se encargará de consultar en la base de datos y retornar los datos requeridos, para su posterior uso.

(a) Consulta de cantidad de registros totales

Se debe tener en cuenta que para poder realizar este análisis se tuvo que segmentar el procesamiento de los datos, debido a la cantidad de información y el tamaño que esta constituía. El proceso se realizó con consultas segmentadas, y se iteró tantas veces como el total de registros a analizar, dividido por 500.000 (cantidad de registros por iteración). Primero se debe realizar una consulta para contar la cantidad total de registros que serán procesados y así realizar este proceso por cada iteración.

```
95  cur = db.cursor()
96  cur.execute("SELECT COUNT(*) FROM tweets, politico_menciones WHERE tweets.
        created_at >= '2013-05-01' AND tweets.created_at < '2013-12-11' AND
        politico_menciones.tweets_id = tweets.id AND politico_menciones.
        politicos_id IN ( 19, 115, 142, 152, 185, 190, 191, 192, 194, 197 )
        ORDER BY tweets.id")
```

Figura 6.3: Consulta Cantidad de Registros

(b) Consulta de datos por iteración

Luego se realiza un loop para procesar cada iteración.

```
115     loop += 1
116     partir = (loop - 1) *  500000
117     consulta = "SELECT tweets.id, tweets.content FROM tweets,
            politico_menciones WHERE tweets.created_at >= '2013-05-01' AND tweets
            .created_at < '2013-12-11' AND politico_menciones.tweets_id = tweets.
            id AND politico_menciones.politicos_id IN ( 19, 115, 142, 152, 185,
            190, 191, 192, 194, 197 ) ORDER BY tweets.id LIMIT " + str(partir) +
            ", " + str(500000)
118     cur.execute(consulta)
```

Figura 6.4: Consulta por cada Iteración

Por cada iteración se realiza una nueva consulta que toma el contenido de los tweets, el cuál posteriormente se usará en el proceso de este análisis.

4. **Concatenación**

Para comenzar con el procesamiento se toma el contenido de cada Tweet y se crea un bloque de texto concatenado. Se detectó que al realizar esta concatenación se generaban "collocations" (bigramas y trigramas), juntando el último término de un Tweet con el primero del Tweet siguiente, los cuales no correspondían a términos reales del estudio. Para resolver esta situación se agrego entre cada contenido un separador con el objetivo de poder descartar el ruido generado.

```
106         for row in cur.fetchall():
107             texto_entero += " xx " + row[1]
```

Figura 6.5: Concatenación

La instrucción "fetchall"genera un arreglo de la consulta que contiene el contenido de los Tweets, el cual se iteró con la finalidad de concatenar el contenido de cada Tweet incluyendo entre cada contenido de los Tweets el separador "xx "(el espacio antes y después de xx es relevante para que los contenidos de cada Tweet no queden unidos).

5. **Reemplazo de caracteres tildados**

Para evitar conflictos con los caracteres tildados o caracteres que corresponden a otros idiomas, estos se reemplazan por el que corresponda, pero sin ningún tipo de tilde. Por ejemplo, si en el texto aparece "á", este se reemplaza por ä ·:

(código función eliminarAcentos() en Anexo Código)

6. **Tokenización**

   Ahora que ya se tiene el texto sin caracteres tildados, se debe tokenizar. Tokenizar consiste en separar el texto por palabras. Este proceso crea un arreglo y cada palabra del texto será un nuevo elemento en éste. La librería de NLTK nos provee una sencilla función para poder realizar este procesamiento en el texto.

```
116     tokens=tokenizer.tokenize(texto_entero)
```

Figura 6.6: Tokenización

   Finalmente se obtiene un arreglo que tendrá un largo igual a la cantidad de palabras (tokens) que se encuentren en el texto.

   Desde aquí en adelante se sabe que nuestro texto está contenido en un arreglo, lo que lleva a tener un estructura de procesamiento distinta a lo que se ha visto anteriormente. Para cada uno de los siguientes pasos, se procesarán los datos iterando por cada elemento del arreglo.

7. **Texto a minúsculas**

   Esta parte del proceso es bastante sencilla, ya que python provee herramientas y funcionalidades para trabajar con strings. En este caso se utilizó la función lower(), ésta toma una cadena de caracteres y reemplaza cuando sea necesario los caracteres en mayúsculas por la misma letra en minúscula.

```
125     tokens_minusculas=[w.lower() for w in tokens]
```

Figura 6.7: Texto a minúsculas

8. **Se eliminan tokens con números o símbolos**

   Se recorre la lista de tokens, y utilizando expresiones regulares se buscan números o símbolos a eliminar descartándolos de la lista.

```
129     tokens_limpio1 = [re.sub(r'\w*\d\w*', '', w).strip() for w in tokens_minusculas]
```

Figura 6.8: Eliminación de Tokens que no son palabras

9. **Se eliminan todo caracter que no sea una letra**

   En este paso se recorre la lista de tokens obtenida anteriormente, se identifica si contienen caracteres que no sean letras y se eliminan de la lista.

```
133     nonPunct = re.compile('.*[A-Za-z].*')
134     tokens_limpio2 = [w for w in tokens_limpio1 if nonPunct.match(w)]
```

Figura 6.9: Eliminación de caracteres en Tokens

10. **Se definen y eliminan Stopwords**

    A continuación se define una lista con los Stopwords que provee la librería NLTK, más algunos Stopwords definidos manualmente.

```
160     stopwords={}
161     nltk_stopwords = nltk.corpus.stopwords.words('spanish') + [
162         'a',    'b',    'c',    'd',
163         'e',    'f',    'g',    'h',
164         'i',    'j',    'k',    'l',
165         'm',    'n',    'o',    'p',
166         'q',    'r',    's',    't',
167         'u',    'v',    'w',    'x',
168         'y',    'z',    'http', 'co',
169         'si',   'dice', 'ser',  'https',
170         'rt'
171         ]
172     stopwords = set(nltk_stopwords)
```

Figura 6.10: Definición de Stopwords

Agregado en anexo stopwords de librería NLTK.
Luego se revisa cada elemento de la lista del texto completo y se eliminan de ésta.

```
180     words_incomplete = [token for token in words if token not in stopwords]
```

Figura 6.11: Eliminación de Stopwords

11. **Se obtiene la frecuencia distribuida de los datos**

    La librería NLTK provee la función FreqDist(), que toma la lista y calcula la frecuencia distribuida de todos los tokens.

```
191         fd += nltk.FreqDist(words_incomplete)
```

Figura 6.12: Frecuencia Distribuida de Tokens

12. **Se obtienen los bigramas y los trigramas (collocations)**
    Utilizando la misma librería NLTK, se obtienen los bigramas (términos de 2 tokens) y los trigramas (términos de 3 tokens), con las funciones bigrams() y trigrams() respectivamente.

```
193         bg_incomplete = nltk.bigrams(words_incomplete)
194         fd_bg += nltk.FreqDist(bg_incomplete)
195
196         tg_incomplete = nltk.trigrams(words_incomplete)
197         fd_tg += nltk.FreqDist(tg_incomplete)
```

Figura 6.13: Frecuencia Distribuida de Bigramas y Trigramas

En esta parte del proceso se vuelve al paso 3.2 hasta que se consulten y procesen todos los registros contados en el paso 3.1

13. **Generación y conteo de vocabulario**

    La función set() garantiza que cada elemento será único, para luego contar la cantidad de vocabulario mediante la función len().

```
209     vocab = set(fd)
210     q_vocab = len(vocab)
```

Figura 6.14: Vocabulario

14. **Consolidación de los términos, obtención de los más comunes y generación de archivo CSV**

    Por último se crea un archivo CSV de salida, en el cual se guardarán los términos más comunes (incluyendo tokens, bigramas y trigramas) para su visualización.

```
216     tabla = fd + fd_bg + fd_tg
```

Figura 6.15: Consolidación de Términos

```
f = open("tabla_comunes.txt", "wb")
w = csv.writer(f)

for key, val in tabla.most_common(100):

    if isinstance(key, str):
        matching = [s for s in sentscompleto if ' ' + str(key) + ' ' in s]
        w.writerow([str(key), val, len(matching)])
    else:
        matching = [s for s in sentscompleto if ' ' + ' '.join(str(i) for i in key) + ' ' in s]
        w.writerow([' '.join(str(i) for i in key), val, len(matching)])

w.writerow(["vocabulario", str(len(vocab))])

f.close()
db.close()
```

Figura 6.16: Generación de CSV

Resultados

Este análisis se realizó para cada mes, desde mayo a noviembre (noviembre incluye los 4 días de datos de diciembre). Además se realizó un procesamiento general de todos los Tweets contenidos entre las fechas analizadas.
Se quizo también obtener los términos más comunes, se generó una tabla considerando la cantidad de registros (Tweets) y el vocabulario total registrado para cada análisis.

Mes	**Tweets**	**Vocabulario**
noviembre, diciembre (4 días)	2096972	277743
octubre	1757489	220050
septiembre	488259	112902
agosto	669256	134726
julio	504416	112211
junio	206855	65150
mayo	77237	37939
completo	5800484	670723

Cuadro 6.1: Tweets y Vocabulario por Mes

Se puede apreciar que la cantidad de Tweets para el análisis completo corresponde a la sumatoria de todos los meses. En el vocabulario completo en cambio, no corresponde a la sumatoria, ya que existe vocabulario repetido para cada mes. El vocabulario registrado corresponde a todos los términos encontrados en el análisis, independiente si son palabras, también están considerados como vocabulario todos los modismos, abreviaciones o cualquier término aún sin significado alguno.

Por otro lado se puede ver en la tabla también que la cantidad de registros va aumentando a medida que se acerca la fecha de elecciones, aún así existe una baja en la cantidad de registros para el mes de Septiembre. La cantidad de Tweets registrados, cercano a las fechas de las elecciones (octubre y noviembre), es bastante más alta que en los meses anteriores.

Como resultados se obtuvieron tablas, a través de archivos CSV, las cuales se utilizaron para graficar los datos. A continuación se expondrán algunos resultados más relevantes:

Mes de Junio

Tabla con los 20 términos más frecuentes

Términos	**Frecuencia**	**Cantidad de Tweets**
pablo	55180	45483
parisi	50063	43842
longueira	49974	46140
fr	46373	45336
fr parisi	46308	42610
pablo longueira	45298	42754
marcelclaude	38555	35268
marcoporchile	33677	31750
comandomichelle	25604	23892
bachelet	20605	19641
mas	16999	15950
chile	16858	15477
tjholt	11859	11437
hoy	11777	11484
allamand	11220	10593
comando	10717	10217
comando pablo	9129	7815
candidato	8545	8332
debate	7336	7129
orrego	7245	6905

Cuadro 6.2: 20 términos más frecuentes de junio

Como se puede apreciar en la tabla la frecuencia indica la cantidad de veces que el término se encontró en los datos, mientras que la cantidad de Tweets corresponde al total de Tweets donde fue encontrado el término, sin considerar si este aparece más de una vez en un Tweet. Luego de generar la tabla se graficaron los 50 términos más comunes.

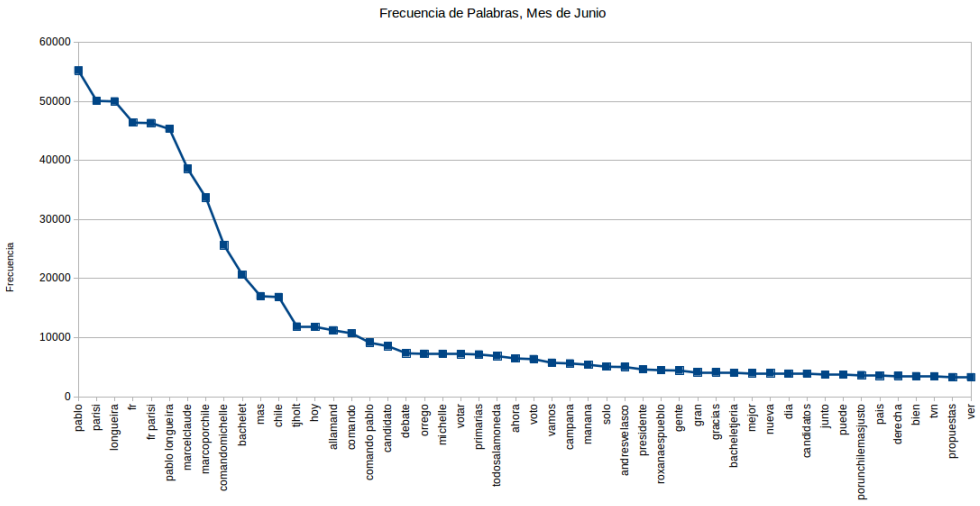

Figura 6.17: Frecuencia de Palabras, Mes de Junio

En el gráfico se puede apreciar que hay términos bastante más comunes que otros, considerando el primero y el último, con una diferencia de más de 10 veces mencionada.

Mes de noviembre

Tabla con los 20 términos más frecuentes

Términos	Frecuencia	Cantidad de Tweets
bachelet	673627	644929
matthei	547292	522718
parisi	494589	429570
fr	407339	398575
fr parisi	404550	370974
marcoporchile	228118	215410
marcelclaude	193991	168393
evelyn	160767	156614
chile	145578	133374
mas	138894	129253
evelyn matthei	138870	128894
longueira	131645	123336
michelle	93313	91947
pablo	89535	77682
comandomichelle	86523	78643
claude	84114	81406
marcel	74126	72463
pablo longueira	73851	69262
michelle bachelet	71217	65452
hoy	70834	69431

Cuadro 6.3: 20 términos más frecuentes de noviembre

A medida que se acerca el proceso de eleccion presidencial, el impacto de Tweets, referente a las elecciones, en la red aumenta considerablemente (10 veces más menciones).

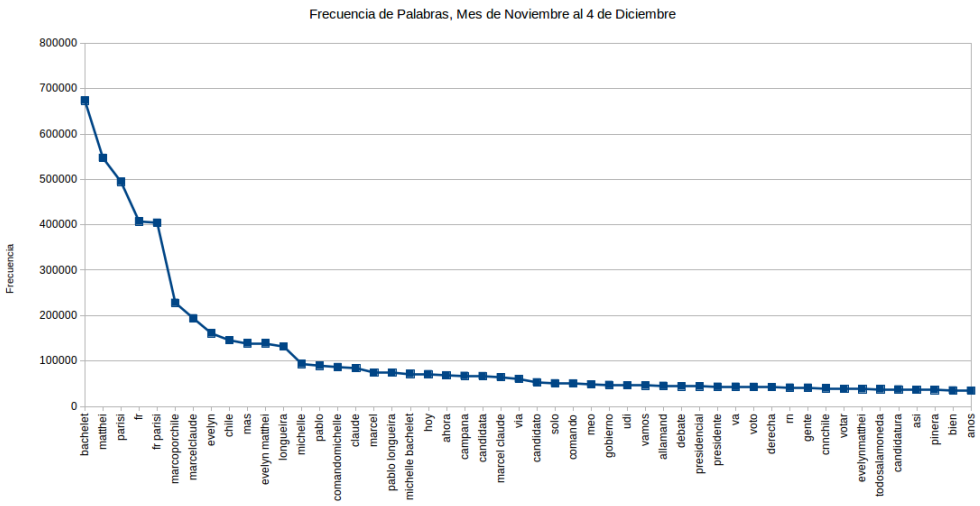

Figura 6.18: Frecuencia de Palabras, Mes de Noviembre al 4 de Diciembre

Se puede apreciar que la tendencia en la red social durante el mes de noviembre, posiciona los términos Bachelet y Matthei como los mas mencionados. Esto es un

reflejo del resultado de la primera vuelta.

Base de datos completa

Tabla con los 20 términos más frecuentes

Términos	Frecuencia	Cantidad de Tweets
bachelet	1864596	1783704
parisi	1553982	1364170
matthei	1346930	1286749
fr	984949	968385
fr parisi	976900	893613
franco	617625	580462
marcoporchile	486965	452425
marcelclaude	446980	346336
chile	388850	355042
mas	371054	343102
evelyn	300601	293110
evelyn matthei	238109	221478
claude	235007	227024
michelle	231236	227217
michelle bachelet	189676	175061
marcel	182866	178271
ahora	178375	174560
vuelta	177238	167480
hoy	4174359	170323
meo	172831	162651

Cuadro 6.4: 20 términos más frecuentes, Base de Datos completa

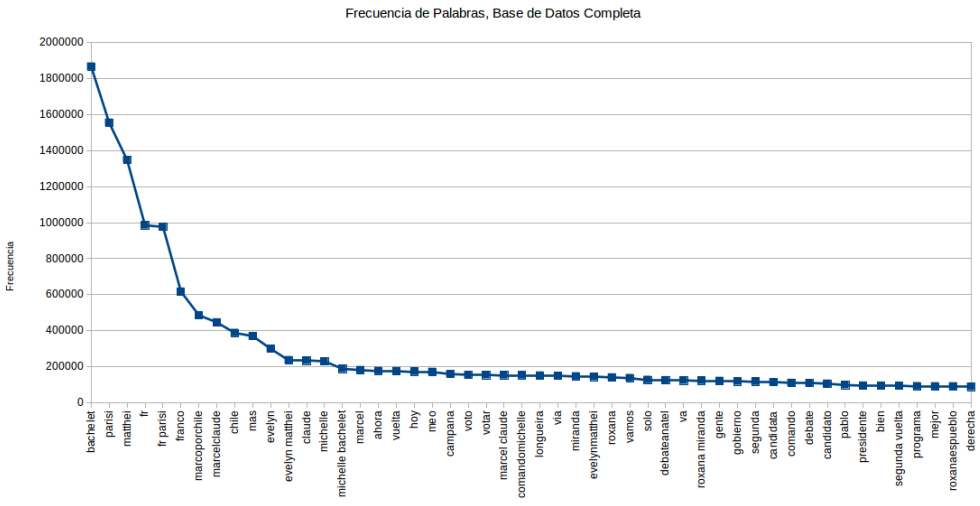

Figura 6.19: Frecuencia de Palabras, Base de Datos completa

Se puede apreciar claramente que la mayor cantidad de menciones hace refe-

rencia a la candidata Michelle Bachelet, lo que se vió reflejado en los resultados de las elecciones presidenciales en Chile del 2013.

Por otro lado, otro gran influyente es el candidato Franco Parisi. Este fue acusado de utilizar "bots", para aumentar su presencia y marcar tendencias en las redes sociales. No existen pruebas de que así fuera, pero la sola acusación y el carácter bromista de sus respuestas, generó mayor interés en las personas.

Figura 6.20: Tweet Franco Parisi

6.1.2 Análisis de sentimiento (polaridad, distribuciones por volumen de tweets), segmentación por candidato y periodo

Para el análisis de sentimiento se realizaron los siguientes pasos:

- Para la obtención de resultados se segmentaron los datos en 3 periodos:

  1. Primer periodo: Antes de las primarias Desde el 01-05-2013 hasta el 30-06-2013 (fecha de las primarias). Veinte semanas antes de las elecciones.
  2. Segundo periodo: Antes de la primera vuelta desde el 30-06-2013 (fecha de las elecciones primarias) hasta el 17-11-2013 (fecha de la primera vuelta). Longueira se baja el 17 de julio.
  3. Tercer periodo: desde el 17-11-2013 hasta el 04-12-2013 (fecha últimos datos del observatorio).

- Se realiza una consulta que cuenta la cantidad de menciones por candidato, considerando la polaridad de cada tweet, para cada periodo definido anteriormente.

  Para este análisis se crearon gráficos de línea, donde se representa la información obtenida por cada consulta realizada. Cada gráfico muestra la cantidad de Tweets en los que se referenció un candidato durante un periodo. Cada línea representa una cantidad diferente de referencia, correspondiendo a la

cantidad total de Tweets (línea azul), Tweets positivos (línea verde) y Tweets negativos (línea roja). La línea azul no corresponde a la suma de los tweets negativos más los positivos, ya que muchos Tweets no tienen una polaridad definida y se consideran neutrales o sin polaridad. En general los candidatos tienen más Tweets positivos que negativos, exceptuando en algunas fechas puntuales, donde sobrepasan los Tweets negativos a los positivos.

Considerando que no todos los candidatos tienen la misma influencia en Twitter, se expondrán los resultados que se consideran más relevantes.

Resultados: Primer Periodo

- **Michelle Bachelet**

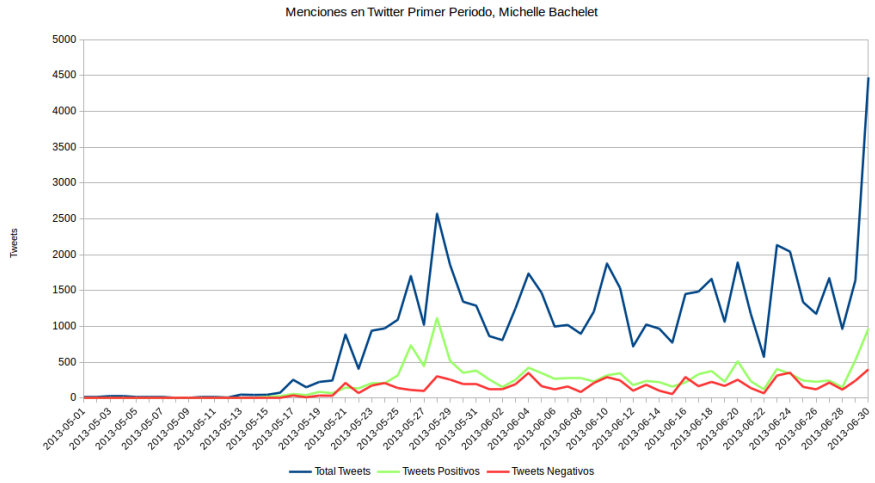

Figura 6.21: Menciones en Twitter, Primer Periodo, Michelle Bachelet

Como ya se explicó anteriormente, el gráfico muestra la cantidad de Tweets en cierto periodo. Este gráfico representa diariamente, durante el primer periodo, los tweets en los que se referenció Michelle Bachelet, y se puede apreciar que en general los Tweets por día son más positivos que negativos. La cantidad de Tweets en el primer periodo, en general es baja comparado con los gráficos durante los periodos más cercanos a la primera y segunda vuelta.

- **Evelyn Matthei**

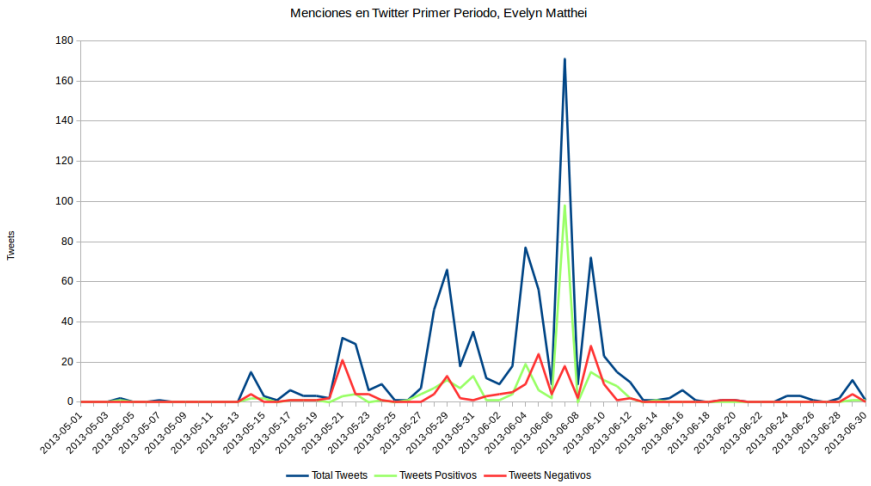

Figura 6.22: Menciones en Twitter, Primer Periodo, Evelyn Matthei

En el caso de Evelyn Matthei, la cantidad de Tweets es mucho más baja que los Tweets referenciados a Michelle Bachelet, y esto se debe a que aún no se sabía que Matthei sería candidata presidencial hasta que ocurre la bajada de Longueira.

Segundo Periodo

- **Franco Parisi**

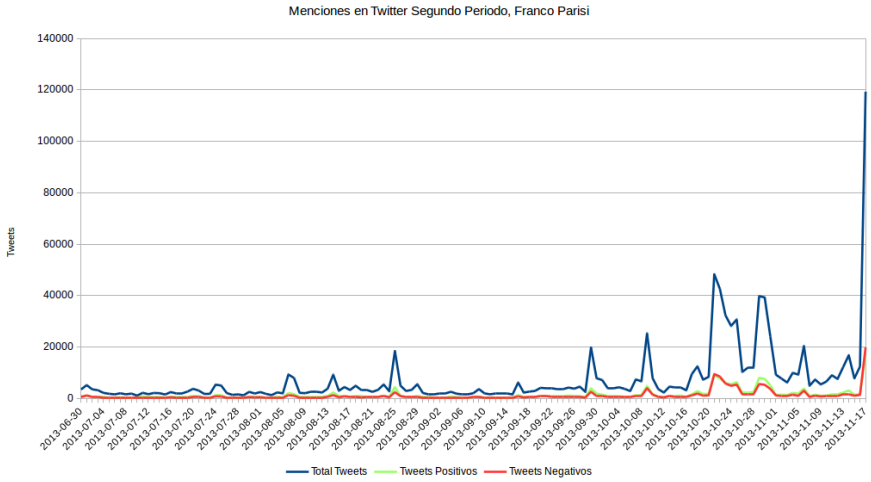

Figura 6.23: Menciones en Twitter, Segundo Periodo, Franco Parisi

Pasando al segundo periodo se puede apreciar que la cantidad de Tweets es muchísima mayor que en el periodo anterior y esto confirma que los usuarios están más interesados y activos en Twitter con respecto al proceso electoral. En el caso de Franco Parisi, su penetración en la red de Twitter es muy alta teniendo mucha presencia y menciones. Las mencione que tiene Bachelet son muy similares a las de Parisi, pero la penetración en Twitter no es representativa ni menos predictiva al momento de revisar los datos reales de las votaciones en las elecciones.

- **Marco Enriquez-Ominami**

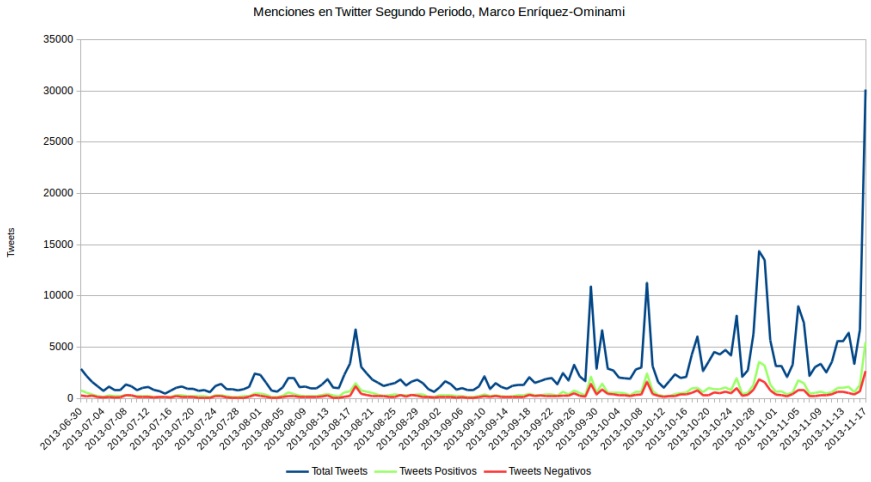

Figura 6.24: Menciones en Twitter, Segundo Periodo, Marco Henriquez-Ominami

Por otro lado, tenemos al candidato Marco Enriquez-Ominami, que en las elecciones presidenciales pasadas (2009) tuvo una penetración

muy interesante, considerando que estuvo más presente en Twitter que Eduardo Frei y Sebastián Piñera [39], candidatos que disputaron la segunda vuelta en las elecciones de ese año. En el caso de las elecciones que se estudiaron en esta memoria, MEO pierde ese prestigio en Twitter y Parisi toma un rol similar al de MEO.

Tercer Periodo

Comparación menciones entre Michelle Bachelet y Evelyn Matthei:

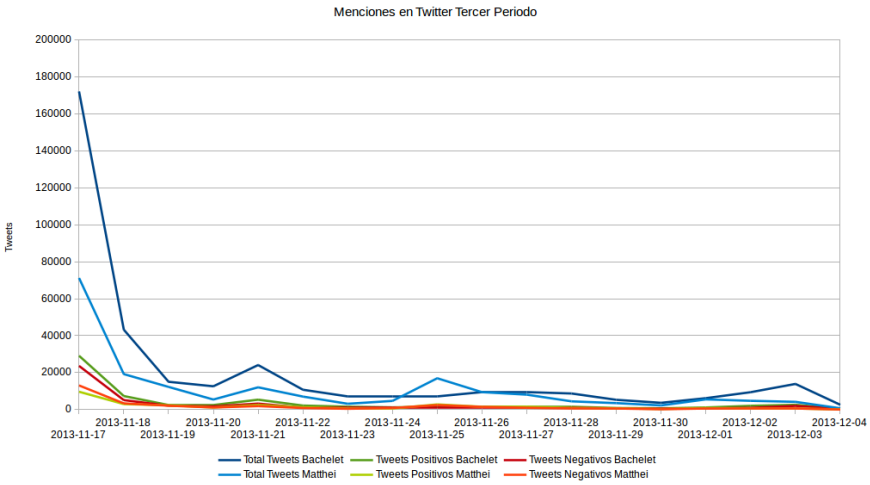

Figura 6.25: Menciones en Twitter, Tercer Periodo

En el último periodo solo se graficaron las menciones de Bachelet y Matthei, ya que ellas pasaron a la segunda vuelta. En el gráfico se puede ver que gran parte del periodo lidera con mayor cantidad de menciones Bachelet por sobre Matthei exceptuando el día 25 de noviembre, prácticamente duplicando la cantidad de menciones esta vez Matthei por sobre Bachelet. Ese día Matthei realizó candidatura en Talca y revolucionó a muchos con su discurso diciendo "Los 'politiqueros' tienen a la gente harta" [30].

6.2 Análisis de la red

6.2.1 Análisis de co-seguimiento

Análisis de seguimiento

La relación entre los seguidores de los candidatos.

- Se importan las librerías
- Se conecta la BD
- Se consultan los seguidores
- Se realiza un conteo de los seguidores para cada par de candidatos y como resultado se obtiene una matriz:

Candidato	Claude	Enriquez-Ominami	Jocelyn-Holt	Parisi	Miranda	Sfeir	Longueira	Israel
Bachelet	39616	42447	24607	74427	43112	22060	29971	10418
Matthei	35798	37700	22669	64257	39795	20333	29049	9975

Cuadro 6.5: Co-seguimiento, Bachelet y Matthei

- A partir de la tabla anterior, utilizando la herramienta de visualización CIRCOS, se obtiene el siguiente gráfico:

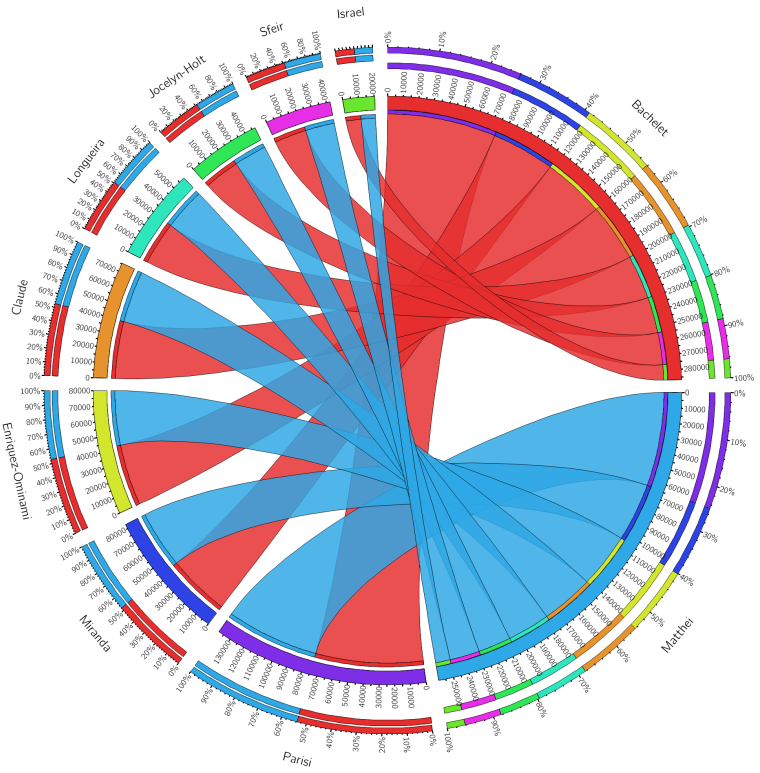

Figura 6.26: Co-seguimiento, Bachelet y Matthei

En este gráfico cada cinta en el centro representa la cantidad de seguidores en común entre Bachelet (cintas rojas) y Matthei (cintas azules) con los demás candidatos. Se puede inferir que los usuarios generalmente siguen a más de un candidato, y lo que refleja el gráfico es que existe una proporción similar entre los usuarios que siguen a Bachelet y a Matthei.

6.2.2 Análisis de co-mención

Análisis de co-menciones

La relación entre las menciones de los candidatos.

- Se importan las librerías
- Se conecta la BD
- Se consultan las menciones
- Se realiza un conteo de las menciones para cada par de candidatos y como resultado se obtiene una matriz

Candidato	Claude	Bachelet	Enriquez Ominami	Jocelyn Holt	Parisi	Miranda Sfeir	Longueira	Matthei	Israel
Bachelet	18401	-	20135	7731	110571	12737 1781	6911	205664	1310
Matthei	6142	205664	10342	1839	175756	8465 1048	10679	-	1527

Cuadro 6.6: Co-mención, Bachelet y Matthei

- A partir de la tabla anterior se obtiene el siguiente gráfico:

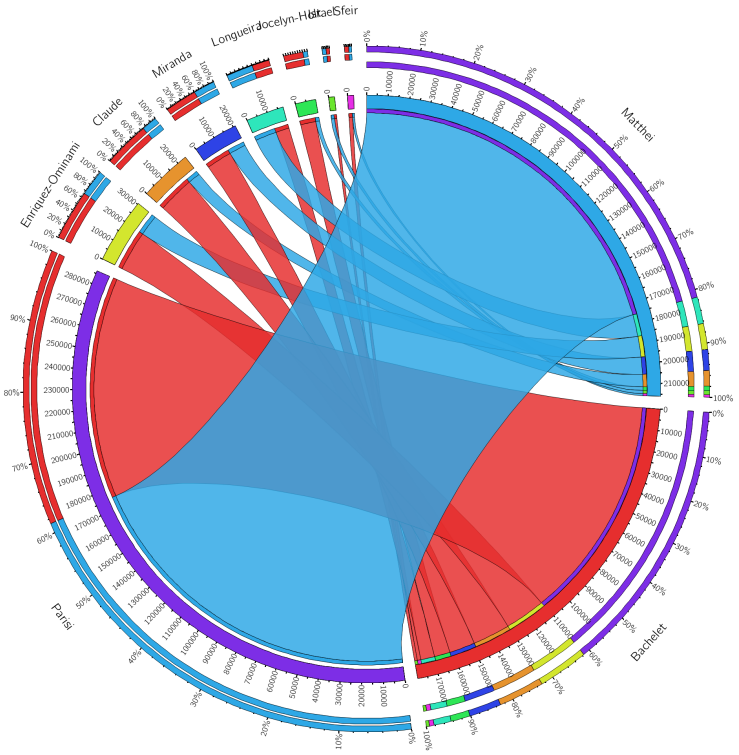

Figura 6.27: Co-mención, Bachelet y Matthei

En este gráfico cada cinta en el centro representa la cantidad de menciones en común entre Bachelet (cintas rojas) y Matthei (cintas azules) con los demás candidatos. A diferencia del analisis anterior, no existe una proporcionalidad en las co-referencias de Bachelet y Mattheil, y se puede apreciar que existe una fuerte relación entre Matthei y Parisi en más de un 60 % de las co-menciones. Indagando más en el tema, se encontraron noticias donde se relacionan. En el diario la Tercera, Matthei acusa a Parisi, de tener deudas con trabajadores [34], mientras que en 24horas.cl Franco Parisi acusa de "matonaje político" a Matthei [35].

6.2.3 Análisis de línea de tiempo

Para este análisis se realiza una consulta que cuenta la cantidad de Tweets al día y se obtiene el siguiente gráfico.

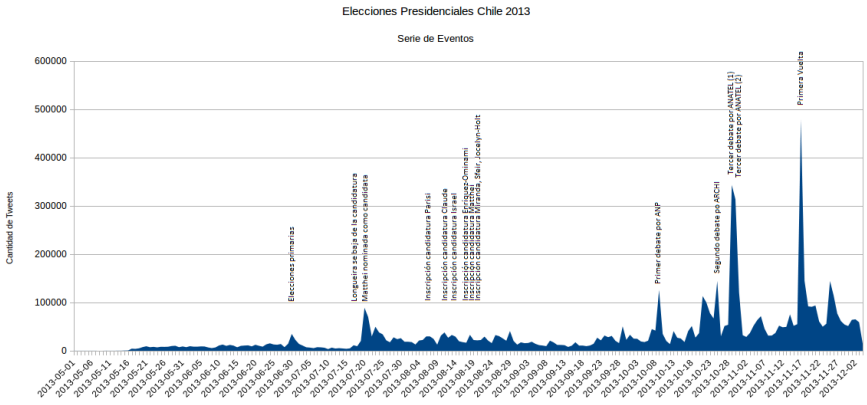

Figura 6.28: Línea de Tiempo

En el gráfico se pueden ver varios "peaks" de Tweets y cada uno de estos representa acontecimientos relevantes a las elecciones presidenciales de 2013 en Chile.

Bajada de Pablo Longueira, subida de Evelyn Matthei

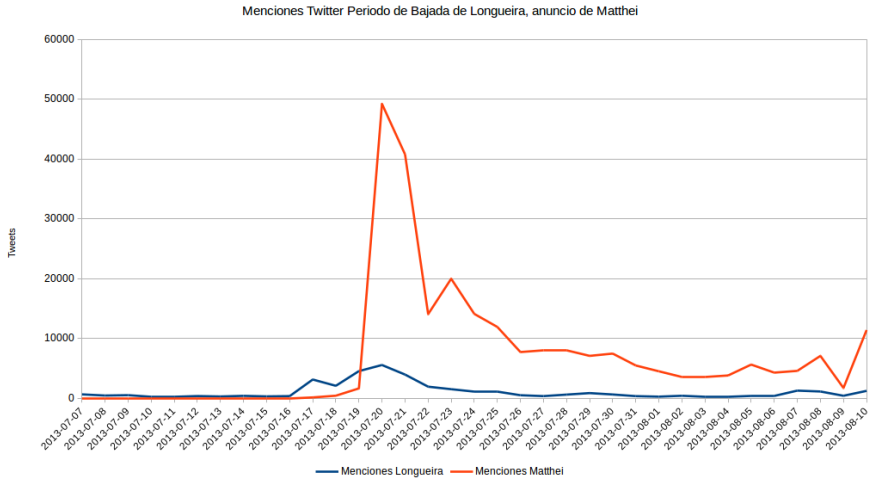

Figura 6.29: Tweets bajada de Longueira, subida de Matthei

Menciones en Twitter de Matthei y Longueira durante el periodo de la bajada de Longueira.

6.2.4 Análisis de las cuentas que más tweets realizaron a un candidato en específico

Corresponde a las cuentas de usuarios que mencionaron más veces a un determinado candidato. El objetivo de este análisis es determinar que tan significativo es el

aporte de una cuenta de usuario determinada para ser considerado como un influenciador en Twitter. La siguiente tabla muestra las 20 cuentas que más twittearon, considerando al candidato que referenciaron:

Cuenta de Twitter	**Nombre de la Cuenta**	**Candidato Referenciado**	**Cantidad de Tweets**
Fr_parisi	Franco Parisi	Franco Aldo Parisi	11313
RafaelWal	Rafael Alejandro Wal	Franco Aldo Parisi	9628
mquirozt	MARIO QUIROZ TREGUER	Franco Aldo Parisi	9260
YoProclamoMarco	YoProclamo#Marco2014	Marco Enríquez Ominami	9178
FcoPlazaG	Francisco Plaza	Franco Aldo Parisi	9051
J_Carol78	Carol#Parisi2014	Franco Aldo Parisi	8342
julietvillac	Julieta V.	Franco Aldo Parisi	7811
MacriEmy	Macris	Franco Aldo Parisi	7189
MigD2010	Miguel Diaz	Marco Enríquez Ominami	6646
sandroriquelme	ParisiPresidente2014	Franco Aldo Parisi	6554
DosMasxLaUC	Víctor Norambuena	Franco Aldo Parisi	6131
candidatosinde	CandiIndependientes	Franco Aldo Parisi	5747
eduardogonzalo	EDUARDO OLMEDO	Marcel Claude	5582
R_Montero_R	Ricardo Montero	Marcel Claude	5226
marcoporchile	MarcoEnríquezOminami	Marco Enríquez Ominami	5220
Ex7reme	David Mesias	Franco Aldo Parisi	5108
KarendTV	Karen Doggenweiler	Marco Enríquez Ominami	5005
RodrigoGBvensee	Rodrigo G. Bevensee	Franco Aldo Parisi	4805
alvarezpff	Patricio Alvarez	Franco Aldo Parisi	4628
RoxanaEsPueblo	Roxana Miranda 2014	Roxana Miranda	4542

Cuadro 6.7: Cuentas de Twitter que realizaron más Tweets

Este análisis se realizó con la finalidad de exponer como la presencia de algunos candidatos en Twitter se ve reflejada por la mera función de utilizar varias cuantas. Con esto se ve aumentada la frecuancia de sus Tweets debido a retweets en masa que condicionan los trending topics. Claramente se puede justificar la alta presencia de Parisi en todo el análisis rtealizado debido a esta técnica.

6.3 Análisis comparado (elecciones versus datos). Muestra sesgada (análisis de sesgo)

Se puede deducir del análisis realizado en esta Memoria que existe una diferencia entre el tipo de personas que utiliza Twitter v/s las personas que votan en las elecciones. Se puede determinar en base a las estadísticas que el tipo de personas que usa Twitter son mayoritariamente adulto joven, proclives a la tecnología, con interés en el debate público y muy al tanto de la contingencia, el mayor porcentaje de usuarios se encuentra en el rango de la edad promedio entre 25 y 44 años.

Cantidad de usuarios por grupos etarios, estadísticas Twitter 2015

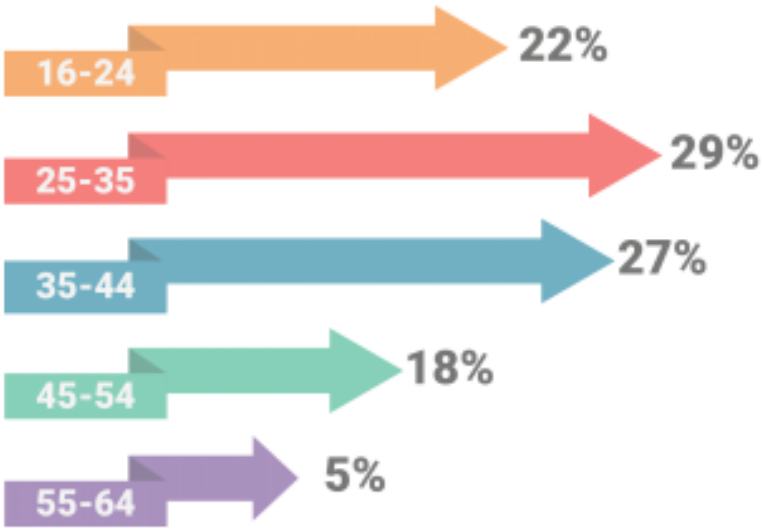

Figura 6.30: Usuarios Twitter por grupor etarios [37]

Por otro lado podemos deducir también basados en datos estadísticos que las personas que votan en las elecciones son mayoritariamente personas adultas, con tradición republicana y sentido de la responsabilidad en edad madura, los que según los datos recabados muestran una mayor votación en las últimas elecciones con promedio de edad entre 50 y 59 años.

Cantidad de votantes por grupos etarios, Elecciones Municipales 2016

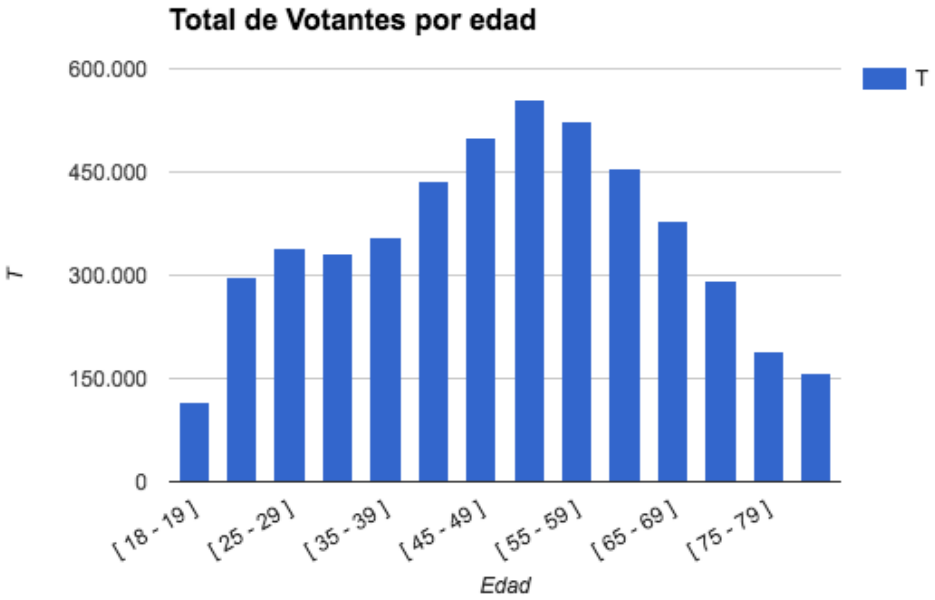

Figura 6.31: Votantes Elecciones Municipales por grupos etarios [38]

Capítulo 7

Conclusiones

7.1 Análisis de los resultados obtenidos

Gracias a la evolución de la tecnología, como se pudo ver en todo el estudio realizado, particularmente en la red social de Twitter, se puede llegar a determinar que no solo es una herramienta de comunicación, sino más bien es un sistema bastante completo, del cual se puede obtener toda la información que expresan las personas en esta. Esto además de permitir influenciar o potenciar las opiniones o comentarios de los usuarios, se puede utilizar para realizar estudios.

Además de la gran cantidad de información que se puede obtener a través de estos medios, existen muchas herramientas para trabajar con esta, que pueden ser gratuitas (open source) o de pago. Esta de más decir que no todas las personas tienen acceso a los mismos medios de comunicación, algunas a través de prensa escrita, televisión, radio y redes sociales. Esto implica directamente en la rapidez con la cual las personas pueden obtener información actualizada.

Una de las características más relevantes que demuestra el estudio es que gracias al correcto análisis de los datos se puede ver una directa relación entre la realidad y los cambios que se provocan en las redes sociales.

Por otro lado, en general los usuarios prefieren comentar o interactuar con publicaciones en las que están de acuerdo o se sienten identificados. Los casos en que los usuarios comienzan disputas o intercambios negativos de opiniones son menores en relación a lo explicado anteriormente.

Se puede afirmar que las redes sociales funcionan más como una caja de resonancia de las ideas propias de los usuarios, que como un medio de comunicación o información, donde los usuarios están leyendo desde sus posiciones de opinión, agregando diferentes matices, pero haciendo eco de sus propias opiniones.

Por otro lado, las redes sociales entregan a los usuarios una sensación de democracia en la opinión y en la información, llegando a pensar que las tendencias o afirmaciones más populares representan perfectamente al común denominador de la sociedad, sin embargo, la evidencia de los resultados obtenidos en el presente estudio ha mostrado que, a pesar de muchas publicaciones o usuarios interactuando con ellas, se trataría de una sobre representación de la opinión de un grupo en particular, y que no necesariamente representa la opinión de la comunidad en general.

Este estudio se realizó basado en las últimas elecciones presidenciales en Chile, con lo que como supuesto de entrada se puede considerar que los políticos ya utilizan redes sociales para realizar campañas políticas o difusión de sus ideas. En este punto surgen varias preguntas como, qué tan certero puede ser un pronóstico eleccionario basado en las publicaciones e interacciones de usuarios, o hasta qué punto es costo eficiente para un candidado realizar esfuerzos digitales de campaña.

Basados en los resultados se puede contestar la primera interrogante afirmando que si bien se logra encontrar una tendencia entre las interacciones de un candidato por sobre otro, no necesariamente representan con exactitud las proporciones en los resultados reales. Esto puede deberse a la penetración de las tecnologías digitales en ciertos sectores de la sociedad, que no necesariamente se ha permeado de manera homogénea. Este acceso a redes sociales parece no tener que ver con la capacidad adquisitiva del individuo, sino más bien con el segmento etário en el que se encuentra.

Si bien hoy, el esfuerzo de propaganda no debería estar centrado en las redes sociales debido a que no alcanzan a toda la sociedad, se puede inferir que en el futuro se podría llegar a un sector perfectamente representativo de la sociedad en general, donde para un candidato podría ser absolutamente costo eficiente basar sus principales esfuerzos en la comunicación de sus ideas a través de redes sociales.

Para finalizar se puede afirmar, con altos niveles de certeza, que lo reflejado en los resultados del estudio aquí realizado, y en otros que hayan utilizado las metodologías aquí descritas, son un reflejo de lo que ocurre en la realidad, representando con ciertos matices lo que la sociedad en generar piensa y opina. Esta afirmación está sustentada en la inexistencia de un estudio que haya logrado demostrar lo contrario.

7.2 Futuros trabajos

Sería un gran aporte al estado del arte la realización de un nuevo estudio en el que se pueda estimar el plazo necesario para que la penetración de las tecnologías digitales hayan alcanzado a toda la sociedad, para así observar la correlación con los resultados electorales reales, y así, poder validar o refutar esta inferencia.

Otro estudio que se podría realizar, considerando que Twitter refleja lo que ocurre en la realidad, sería poder indagar en la cantidad de personas que twitean vs la cantidad real de votantes, considerando rangos etários, estratos sociales, circunscripciones, etc.

Bibliografía

[1] *Matt McGee, By The Numbers: Twitter Vs. Facebook Vs. Google Buzz, [en linea]*

<http://searchengineland.com/by-the-numbers-twitter-vs-facebook-vs-google-buzz-36709> [26 de febrero del 2017]

[2] *Pingdom, Internet 2010 in numbers*, [en linea]

<http://searchengineland.com/by-the-numbers-twitter-vs-facebook-vs-google-buzz-36709> [26 de febrero del 2017]

[3] *Artículo The New York Times*, [en linea]

<https://bits.blogs.nytimes.com/2008/11/07/how-obamas-internet-campaign-changed-politics/?_r=0> [28 de febrero del 2017]

[4] *SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining (2006)* , [en línea]

<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.7217> [28 de febrero del 2017]

[5] *Articulo de Wall Street Journal*, [en linea]

<https://www.wsj.com/articles/cord-cutting-is-accelerating-1449745201> [2 de marzo del 2017]

[6] *Internet amplifica el poder de las personas*,[en linea]

- <https://www.youtube.com/watch?v=qpkENiSUcJM> (Minuto 8:40)
  CASTELLS, Manuel. La Galaxia Internet. Barcelona: Areté, 2001. 316 p.
- <http://www.ub.edu/geocrit/b3w-374.htm>
  Castells, M. (2001): Internet y la Sociedad Red
- <http://tecnologiaedu.us.es/cuestionario/bibliovir/106.pdf>
  [2 de marzo del 2017]

[7] *E*ncuesta CASEN, [en línea]
<http://www.encuestacasen.cl/> [2 de marzo del 2017]

[8] *E*ncuestas CEP, [en línea]
<https://www.cepchile.cl/quienes-somos/cep/2016-01-28/085754.html>
[2 de marzo del 2017]

[9] *G*FK Adimark, [en línea]
<http://www.adimark.cl/es/empresa.asp> [2 de marzo del 2017]

[10] *MORI (Market & Opinion Research International)*, [en línea]
<http://morichile.cl/> [2 de marzo del 2017]

[11] *Daniel Gayo-Avello. Information Processing & Management Volume 49, Issue 6, November 2013, Pages 1250-1280*

[12] *K.Bharat, and M.Henzinger. -Improved algorithms for topic distillation in hyper -linked environments, Proceedings* of the 21st International ACM SIGIR Conference, 1998, pp. 104-111.

[13] *Daniel Gayo-Avello, Don't Turn Social Media Into Another 'Literary Digest' Poll. Magazine Communications of the ACM, Volume 54 Issue 10, October 2011, Pages 121-128*

[14] *Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsup ervised Classication of Reviews. In Pro ceedings of the 40th Annual Meeting of the Asso ciation for Computational Linguistics*, 2002, pp. 417-424

[15] *Theresa Wilson, Janyce Wieb e, and Paul Homann. Recog-nizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Pro ceedings of HLT-EMNLP 2005, pp.* 347-354.

[16] *Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Rout-ledge, and Noah A. Smith, 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Pro ceedings of the 4th International AAAI Conference on Weblogs and Social Media.*

[17] *Amanda Lenhart, Susannah Fox, 2009. Twitter and status up dating. Pew Internet and American Life. Available at:*

<http://www.pewinternet.org/Reports/2009/Twitter-and-status-updating.aspx>

[18] *Aaron Smith, and Lee Rainie, 2008. The Internet and the 2008 election. Pew Internet and American Life. Available at:*

<http://www.pewinternet.org/Reports/2008/The-Internet-and-the-2008-Election.aspx>

[19] *Bei Yu, Stefan Kaufmann, and Daniel Diermeier, 2008. Exploring the characteristics of opinion expressions for political opinion clasfiication. In Pro ceedings of the 2008 international conference on Digital government research*, pp. 8291

[20] *Daniel Gayo - Avello, "Social Media, Democracy, and Democratization", IEEE MultiMedia, vol.22, no. 2, Apr. - June 2015, pp.* 10 - 16

[21] *Felipe Bravo-Marquez, Marcelo Mendoza, Barbara Poblete. Combining strengths, emotions and polarities for boosting Twitter sentiment analysis, August 11 - 11, 2013*

[22] *Andrea Esuli, Fabrizio Sebastiani. SENTIWORDNET: A Publicly Available Lexi- cal Resource for Opinion Mining, Proceedings of LREC, May 2006*

[23] *Sentiment analysis, Wikipedia*, [en línea]
<http://en.wikipedia.org/wiki/Sentiment_analysis> [09 de marzo del 2017]

[24] *GRS Social Networking*, [en línea]

<http://grssocial.com/presidenciales-2013-analisis-y-estadisticas-de-candidatos-en-redes-sociales-durante-septiembre/> [10 de Marzo 2017]

[25] *Servicio Electoral (SERVEL)* , [en línea]

- <https://www.servel.cl/resultado-eleccion-presidencial-parlamentaria-y-de-cores-2013/> [10 de marzo 2017]
- <https://www.servel.cl/elecciones-presidenciales-analisis-de-resultados/> [10 de marzo 2017]

[26] *El Mercurio Online (EMOL)*, [en línea]

<http://www.emol.com/especiales/2013/actualidad/nacional/carrera-presidencial/> [10 de marzo 2017]

[27] Análisis de elecciones, [en línea]

<http://www.americaeconomia.com/analisis-opinion/elecciones-en-chile-el-triunfo-de-michelle-bachelet-y-la-segunda-vuelta> [10 de marzo 2017]

[28] *Documentación de uso de API de Twitter*, [en línea]

<https://dev.twitter.com/overview/api> [17 de Marzo 2017]

[29] *Lenguaje de programación Python*, [en línea]

<https://www.python.org/> [31 de Marzo 2017]

[30] *Circos, herramienta para gráficos*, [en línea]

<http://circos.ca//> [05 de Abril 2017]

[31] *Base de Datos MySql, herramienta informática,* [en línea]

- <https://www.oracle.com/lad/mysql/index.html> [05 de Abril 2017]
- <https://www.mysql.com/> [05 de Abril 2017]

[32] *Libre Office, herramienta de ofimatica* , [en línea]
<https://es.libreoffice.org/descubre/calc/> [05 de Abril 2017]

[33] *TRICEL, Tribunal Calificador de Elecciones*, [en línea] <http://www.tribunalcalificador.cl/resultados-electorales/> [05 de Abril 2017]

[34] *La Tercera, Matthei acusa a Parisi de tener deudas con trabajadores por cerca de $ 100 millones*, 21/10/2013, <http://www.latercera.com/noticia/matthei-acusa-a-parisi-de-tener-deudas-con-trabajadores-por-cerca-de-100-millones/>

[35] *24Horas.cl, Franco Parisi acusa de "matonaje político.ª Matthei*, 21 octubre 2013 <http://www.24horas.cl/politica/decisionfinal/franco-parisi-acusa-de-matonaje-politico-a-matthei–898213>

[36] *Imagen de la estructura del Tweet definida por Rafik Krikorian en anexo.* <https://www.linkedin.com/in/rkrikorian/>

[37] *Estadísticas de Twitter 2015* <http://comunidad.iebschool.com/iebs/redes-sociales/estadisticas-usuarios-twitter-como-son/>

[38] *Estadísticas de votantes 2016 del SERVEL* <https://www.servel.cl/wp-content/uploads/2017/03/Votantes_Edad_Sexo_Comuna_Municipales_2016.pdf>

[39] <http://www.emol.com/noticias/magazine/2009/08/13/371448/marco-enriquez-ominami-tambien-suma-votos-como-mr-twitter.html>

[40] <http://webcache.googleusercontent.com/search?q=cache:sEUFTRIlikoJ:www.24horas.cl/politica/decisionfinal/matthei-los-politiqueros-tienen-a-la-gente-harta-953062+&cd=1&hl=es&ct=clnk&gl=cl>

[41] <http://www.twitter.com/nombre_de_usuario>

[42] <https://www.papelenblanco.com/metacritica/la-ley-de-zipf-la-frecuencia-con-la-que-una-palabra-aparece-en-un-texto>

[43] <http://bibliotecadigital.ilce.edu.mx/sites/ciencia/volumen3/ciencia3/150/htm/sec_23.htm>

[44] <http://www.abc.es/ciencia/20131213/abci-misteriosa-predice-tamano-ciudades-201312131013.html>

[45] <https://ideas.repec.org/p/col/000102/003325.html>

[46] *Ricardo Baeza-Yates, Excavando la Web - artículo publicado en El Profesional de la Información*

<http://grix.tk/images/baeza2004.pdf>

[47] <https://us.pycon.org/2017/>

[48] <https://www.python.org/psf/grants/>

[49] <https://dev.twitter.com/docs/api/1.1/get/search/tweets>

[50] <https://api.twitter.com/1.1/search/tweets.json>

Apéndice A

Anexos de código

Investigación reproducible (se anexan códigos para explicaciones en las exposición de los experimentos) para garantizar la reproducción de estos resultados se anexan los siguientes códigos:

```
22  def eliminarAcentos(cadena):
23      d = {    '\xc1':'A',        '\xc0':'A',
24          '\xc9':'E',         '\xcd':'I',
25          '\xd3':'O',         '\xda':'U',
26          '\xdc':'U',         '\xd1':'N',
27          '\xc7':'C',         '\xed':'i',
28          '\xf3':'o',         '\xf1':'n',
29          '\xe7':'c',         '\xba':'',
30          '\xb0':'',          '\x3a':'',
31          '\xe0':'a',         '\xe1':'a',
32          '\xe2':'a',         '\xe3':'a',
33          '\xe4':'a',         '\xe5':'a',
34          '\xe8':'e',         '\xe9':'e',
35          '\xea':'e',         '\xeb':'e',
36          '\xec':'i',         '\xed':'i',
37          '\xee':'i',         '\xef':'i',
38          '\xf2':'o',         '\xf3':'o',
39          '\xf4':'o',         '\xf5':'o',
40          '\xf0':'o',         '\xf9':'u',
41          '\xfa':'u',         '\xfb':'u',
42          '\xfc':'u',         '\xe5':'a'
43          }
44
45      nueva_cadena = cadena
46      for c in d.keys():
47          nueva_cadena = nueva_cadena.replace(c,d[c])
48
49      return nueva_cadena
```

Figura A.1: Función de eliminación de tildes

```
18  nltk_stopwords_spanish = ['de', 'la', 'que', 'el', 'en', 'y', 'a', 'los', 'del', 'se', 'las',
        'por', 'un', 'para', 'con', 'no', 'una', 'su', 'al', 'lo', 'como', 'más', 'pero', 'sus', '
        le', 'ya', 'o', 'este', 'sí', 'porque', 'esta', 'entre', 'cuando', 'muy', 'sin', 'sobre',
        'también', 'me', 'hasta', 'hay', 'donde', 'quien', 'desde', 'todo', 'nos', 'durante', '
        todos', 'uno', 'les', 'ni', 'contra', 'otros', 'ese', 'eso', 'ante', 'ellos', 'e', 'esto',
         'mí', 'antes', 'algunos', 'qué', 'unos', 'yo', 'otro', 'otras', 'otra', 'él', 'tanto', '
        esa', 'estos', 'mucho', 'quienes', 'nada', 'muchos', 'cual', 'poco', 'ella', 'estar', '
        estas', 'algunas', 'algo', 'nosotros', 'mi', 'mis', 'tú', 'te', 'ti', 'tu', 'tus', 'ellas'
        , 'nosotras', 'vosotros', 'vosotras', 'os', 'mío', 'mía', 'míos', 'mías', 'tuyo', 'tuya'
        , 'tuyos', 'tuyas', 'suyo', 'suya', 'suyos', 'suyas', 'nuestro', 'nuestra', 'nuestros', '
        nuestras', 'vuestro', 'vuestra', 'vuestros', 'vuestras', 'esos', 'esas', 'estoy', 'estás',
         'está', 'estamos', 'estáis', 'están', 'esté', 'estés', 'estemos', 'estéis', 'estén', '
        estaré', 'estarás', 'estará', 'estaremos', 'estaréis', 'estarán', 'estaría', 'estarías', '
        estaríamos', 'estaríais', 'estarían', 'estaba', 'estabas', 'estábamos', 'estabais', '
        estaban', 'estuve', 'estuviste', 'estuvo', 'estuvimos', 'estuvisteis', 'estuvieron', '
        estuviera', 'estuvieras', 'estuviéramos', 'estuvierais', 'estuvieran', 'estuviese', '
        estuvieses', 'estuviésemos', 'estuvieseis', 'estuviesen', 'estando', 'estado', 'estada', '
        estados', 'estadas', 'estad', 'he', 'has', 'ha', 'hemos', 'habéis', 'han', 'haya', 'hayas'
        , 'hayamos', 'hayáis', 'hayan', 'habré', 'habrás', 'habrá', 'habremos', 'habréis', 'habrán
        ', 'habría', 'habrías', 'habríamos', 'habríais', 'habrían', 'había', 'habías', 'habíamos',
         'habíais', 'habían', 'hube', 'hubiste', 'hubo', 'hubimos', 'hubisteis', 'hubieron', '
        hubiera', 'hubieras', 'hubiéramos', 'hubierais', 'hubieran', 'hubiese', 'hubieses', '
        hubiésemos', 'hubieseis', 'hubiesen', 'habiendo', 'habido', 'habida', 'habidos', 'habidas'
        , 'soy', 'eres', 'es', 'somos', 'sois', 'son', 'sea', 'seas', 'seamos', 'seáis', 'sean', '
        seré', 'serás', 'será', 'seremos', 'seréis', 'serán', 'sería', 'serías', 'seríamos', '
        seríais', 'serían', 'era', 'eras', 'éramos', 'erais', 'eran', 'fui', 'fuiste', 'fue', '
        fuimos', 'fuisteis', 'fueron', 'fuera', 'fueras', 'fuéramos', 'fuerais', 'fueran', 'fuese'
        , 'fueses', 'fuésemos', 'fueseis', 'fuesen', 'sintiendo', 'sentido', 'sentida', 'sentidos'
        , 'sentidas', 'siente', 'sentid', 'tengo', 'tienes', 'tiene', 'tenemos', 'tenéis', 'tienen
        ', 'tenga', 'tengas', 'tengamos', 'tengáis', 'tengan', 'tendré', 'tendrás', 'tendrá', '
        tendremos', 'tendréis', 'tendrán', 'tendría', 'tendrías', 'tendríamos', 'tendríais', '
        tendrían', 'tenía', 'tenías', 'teníamos', 'teníais', 'tenían', 'tuve', 'tuviste', 'tuvo',
        'tuvimos', 'tuvisteis', 'tuvieron', 'tuviera', 'tuvieras', 'tuviéramos', 'tuvierais', '
        tuvieran', 'tuviese', 'tuvieses', 'tuviésemos', 'tuvieseis', 'tuviesen', 'teniendo', '
        tenido', 'tenida', 'tenidos', 'tenidas', 'tened']
```

Figura A.2: Lista de Stopwords

Apéndice B

Anexos Imágenes

Imagen de la estructura de un Tweet provista por Raffi Krikorian

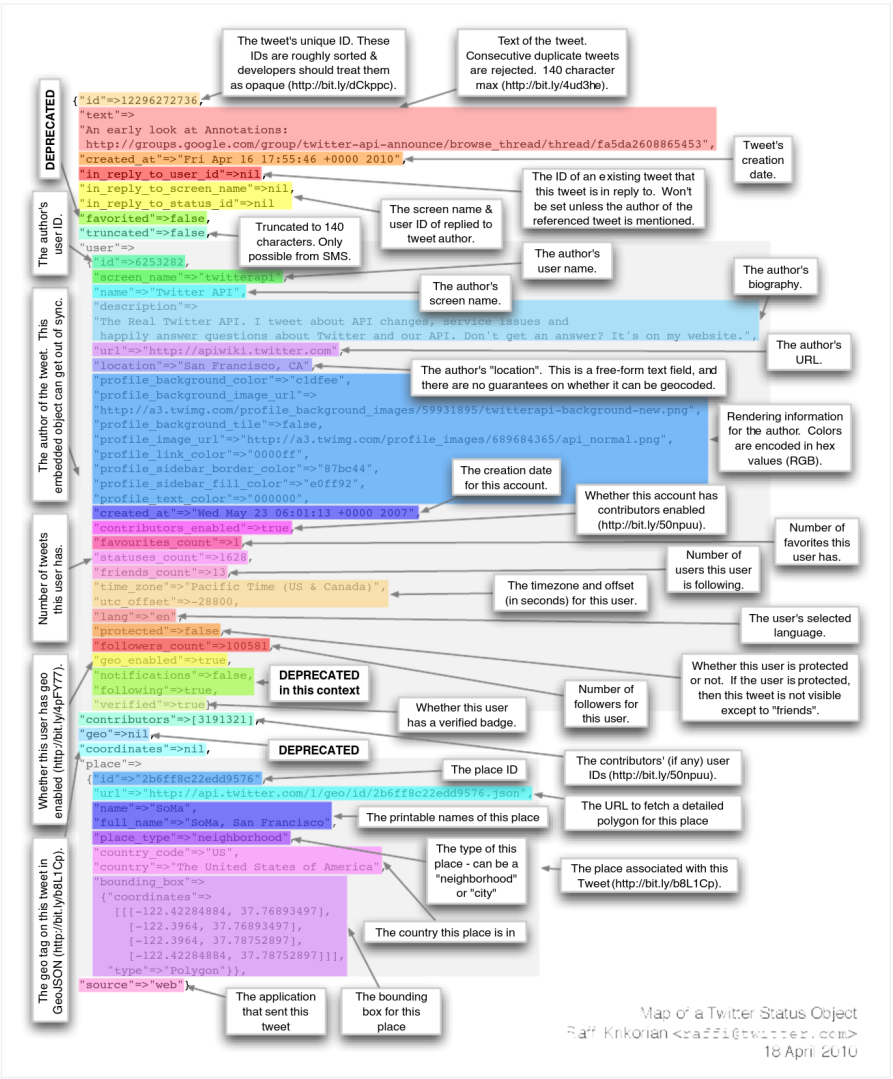

Figura B.1: Estructura de un Tweet

Apéndice C

Anexos de Gráficos

C.1 Co-menciones

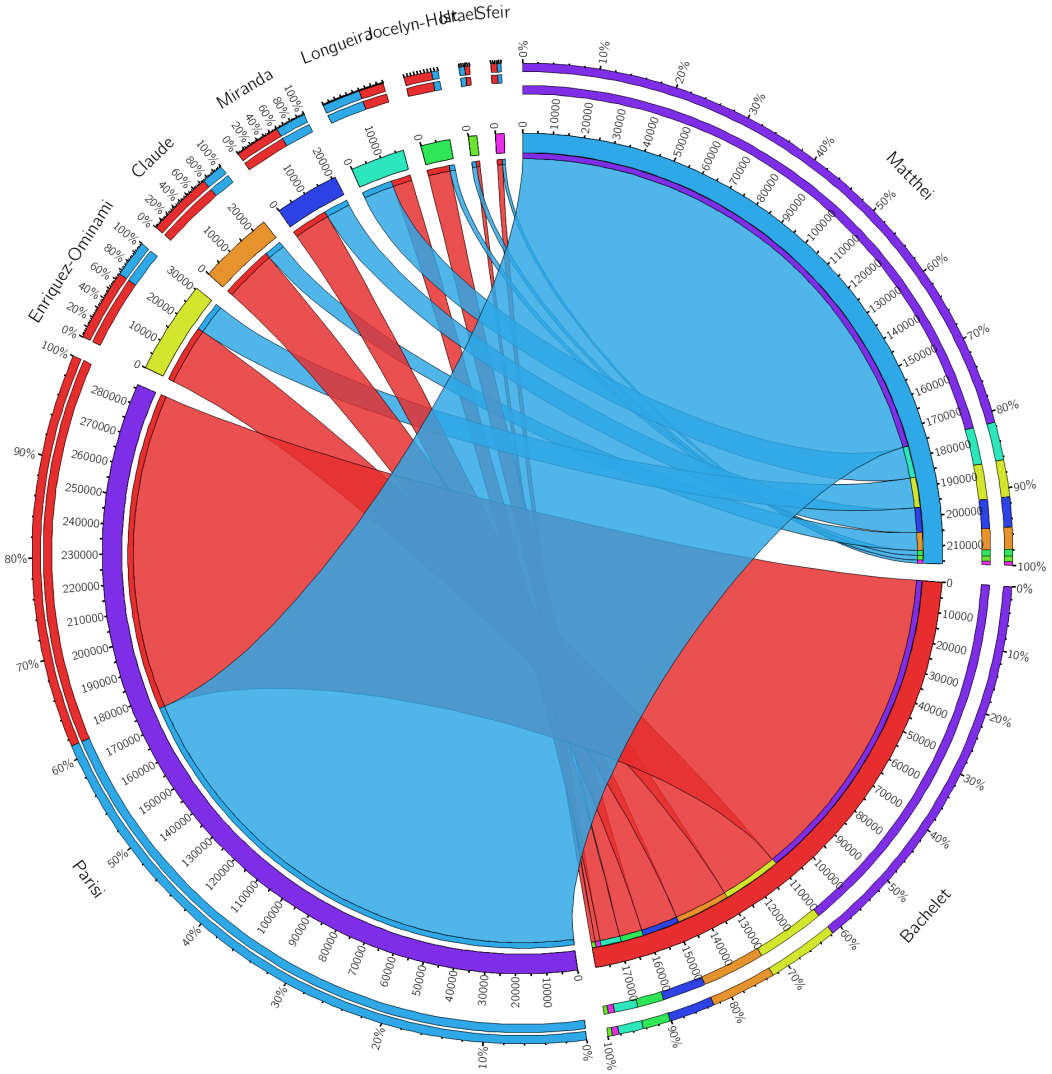

Figura C.1: Co-menciones Bachelet Matthei

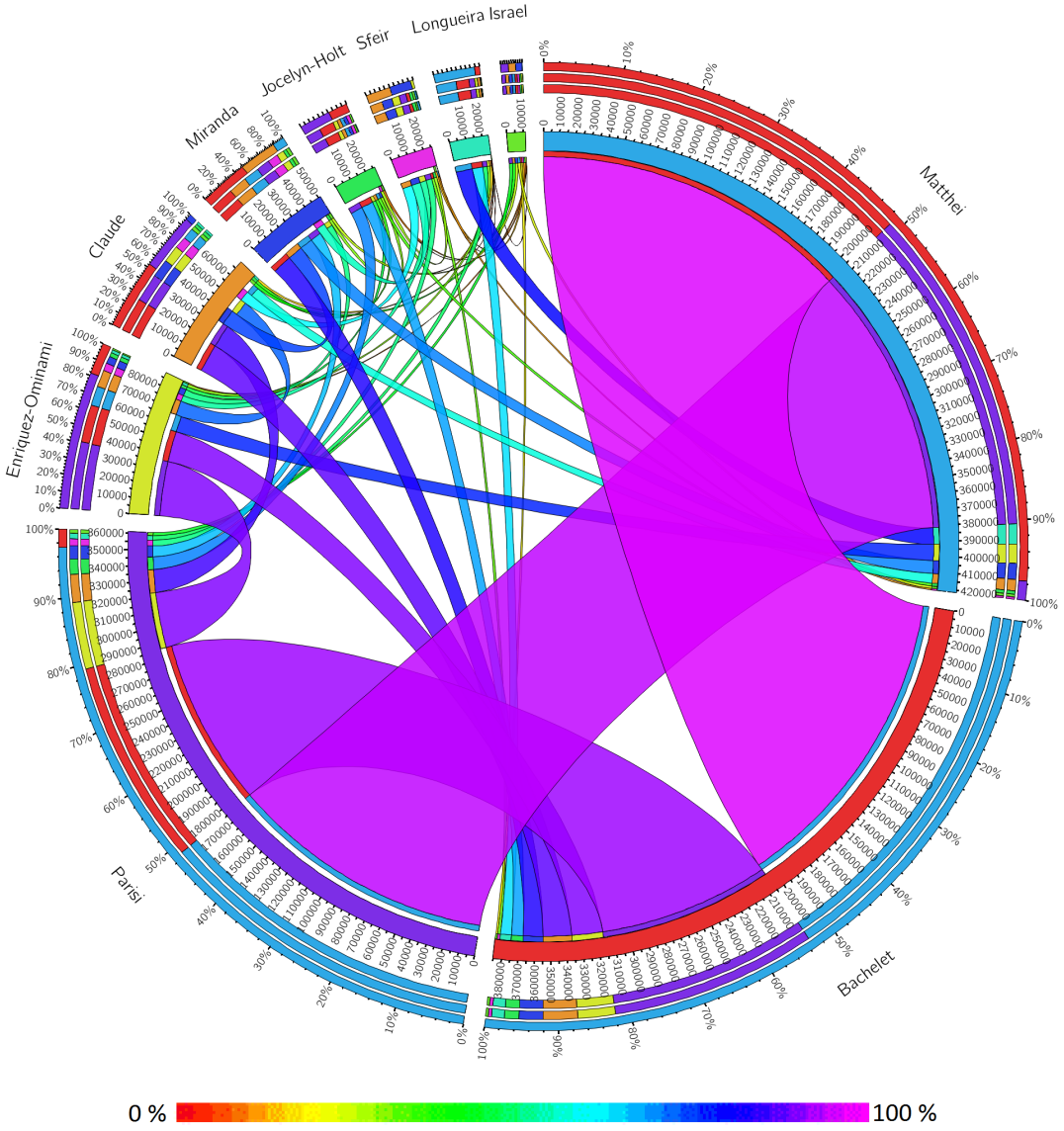

Figura C.2: Co-menciones todos los candidatos

C.2 Co-seguimiento

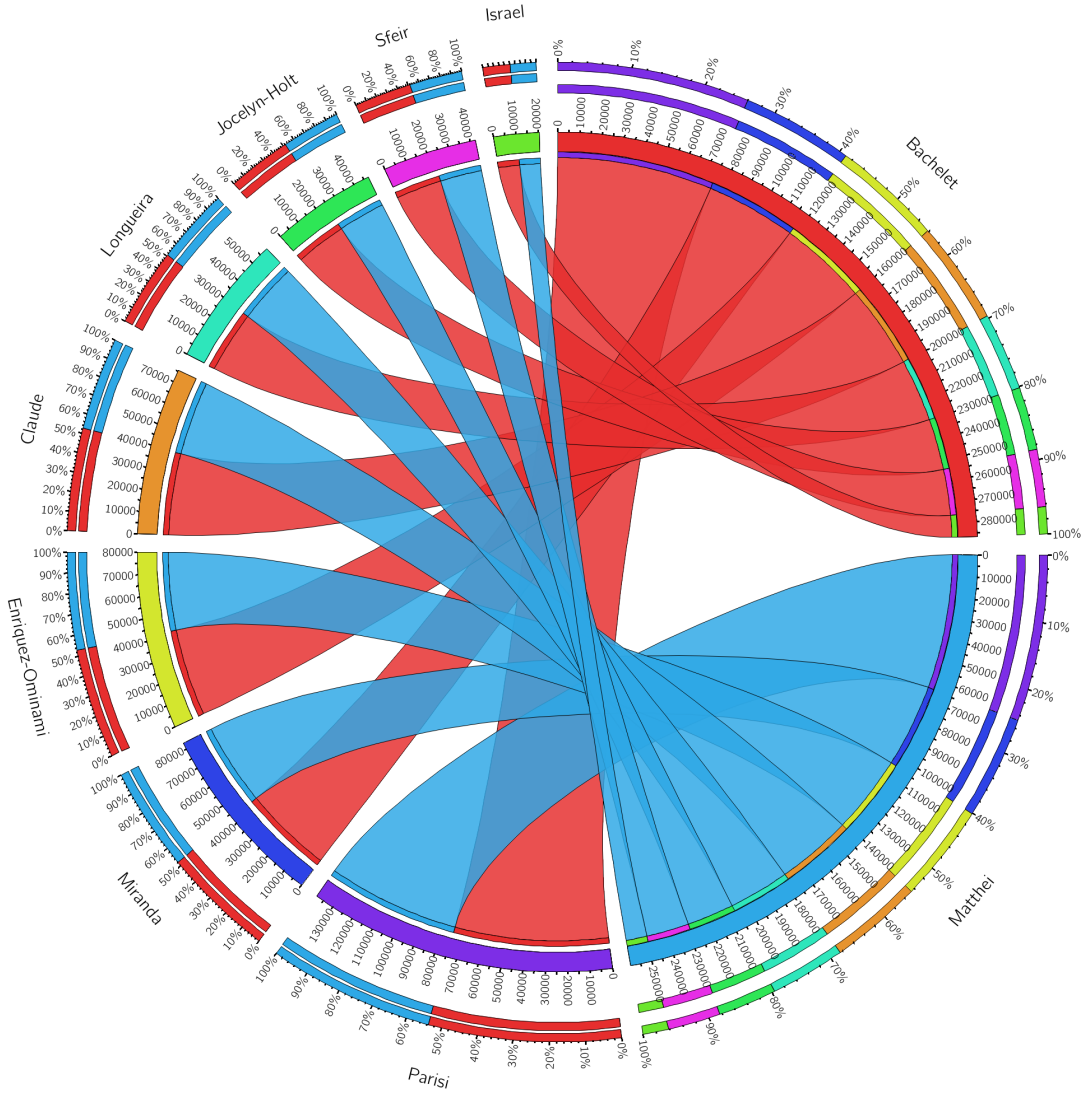

Figura C.3: Co-seguimiento Bachelet Matthei

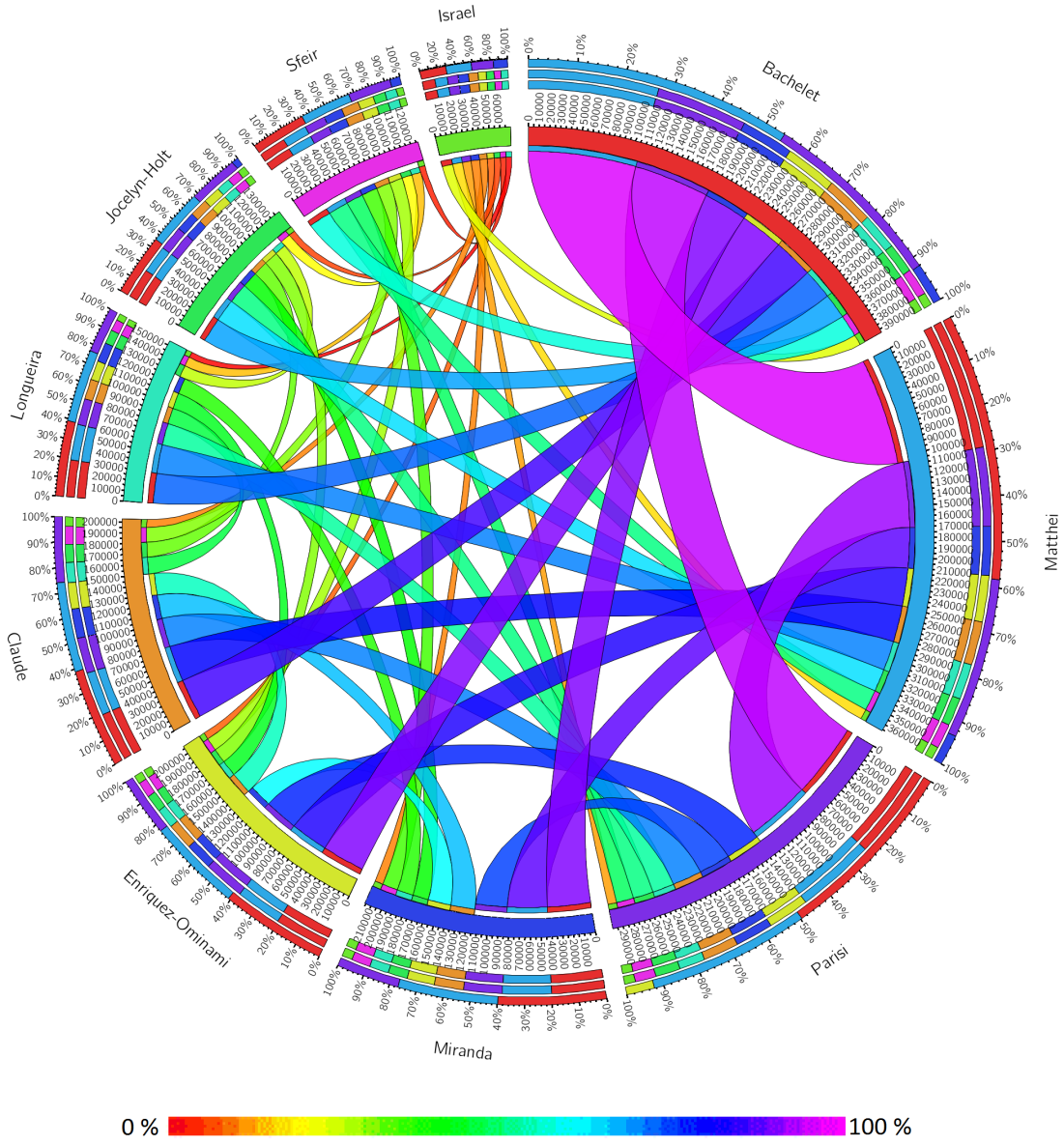

Figura C.4: Co-seguimiento todos los candidatos

C.3 Línea de Tiempo

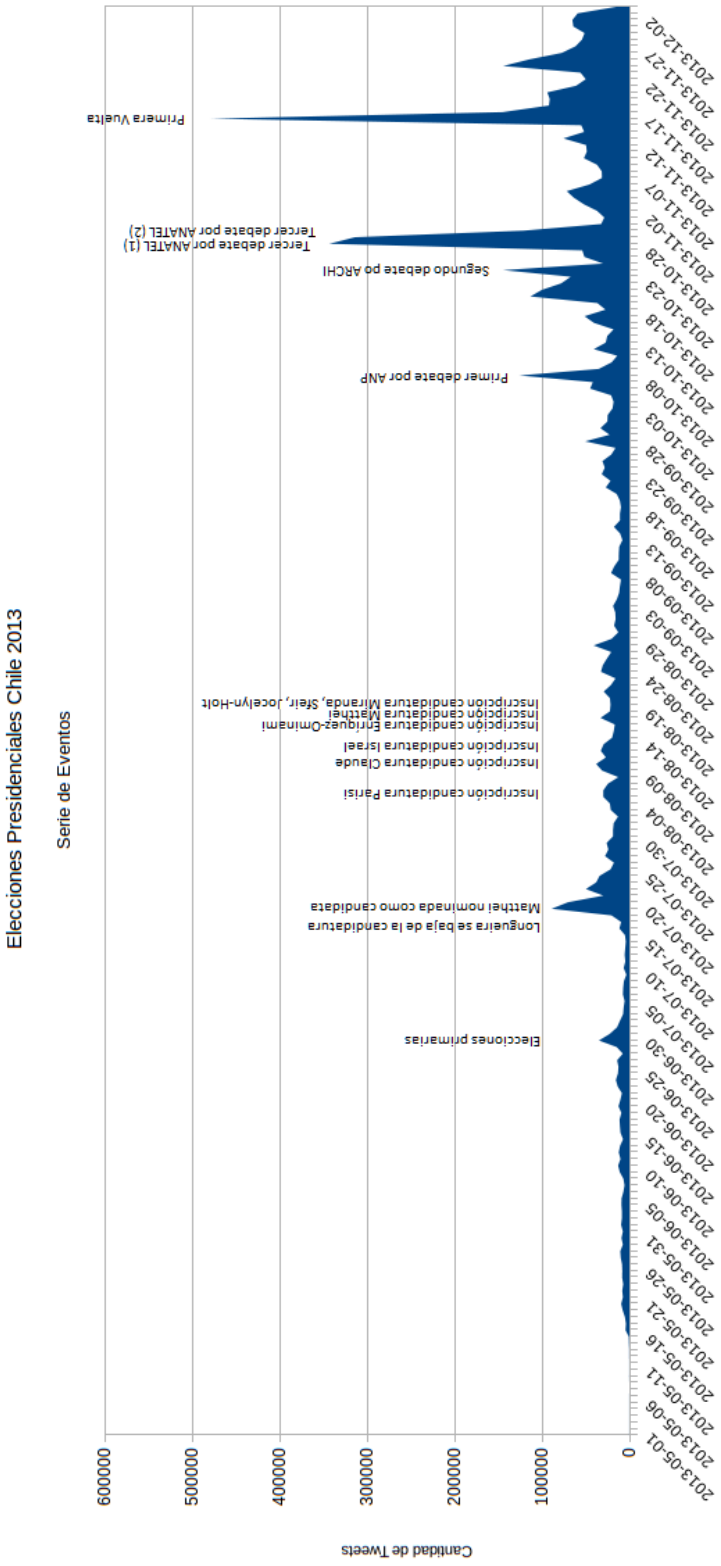

Figura C.5: Línea de Tiempo

C.4 Primer Periodo

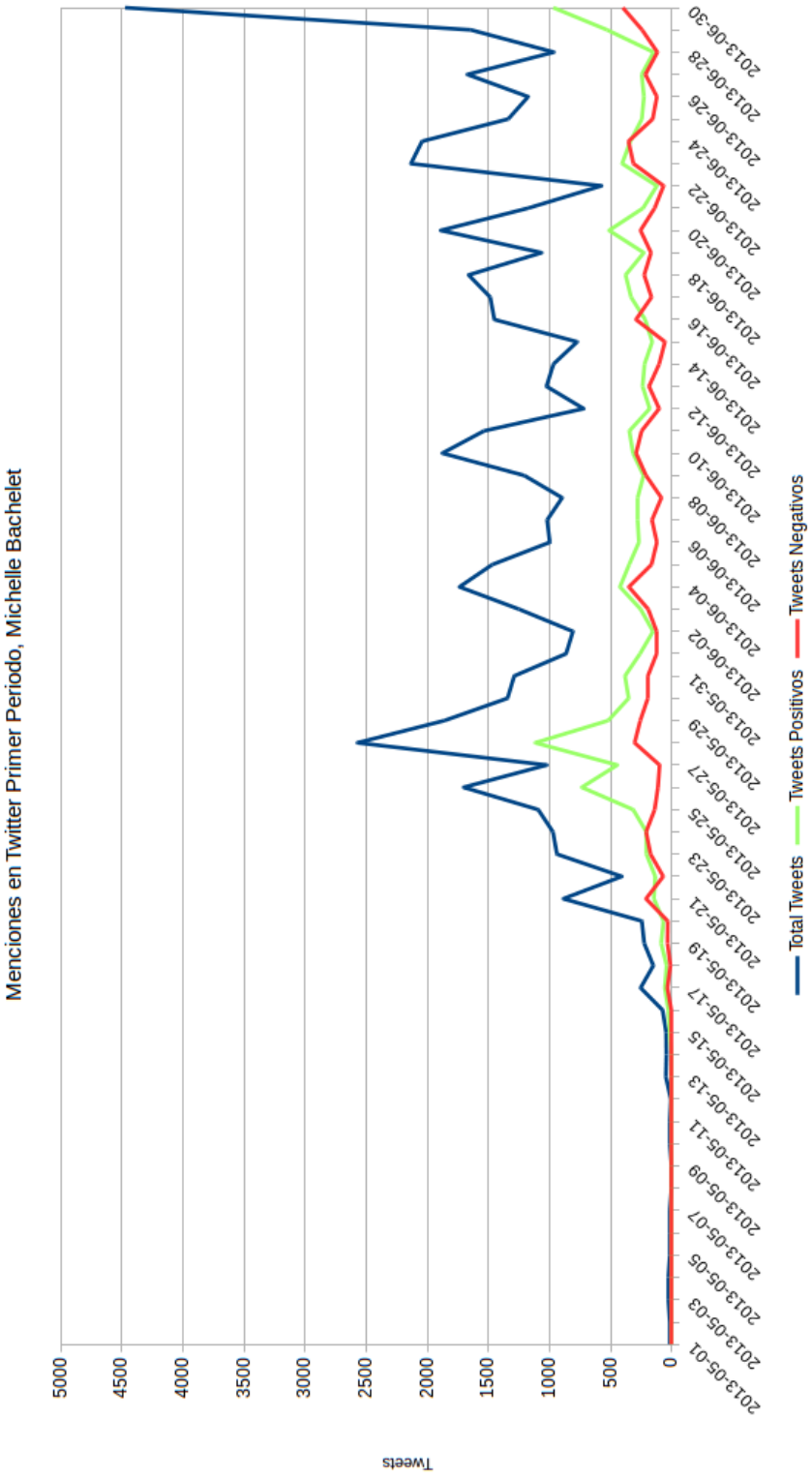

Figura C.6: Menciones en Twitter Primer Periodo, Michelle Bachelet

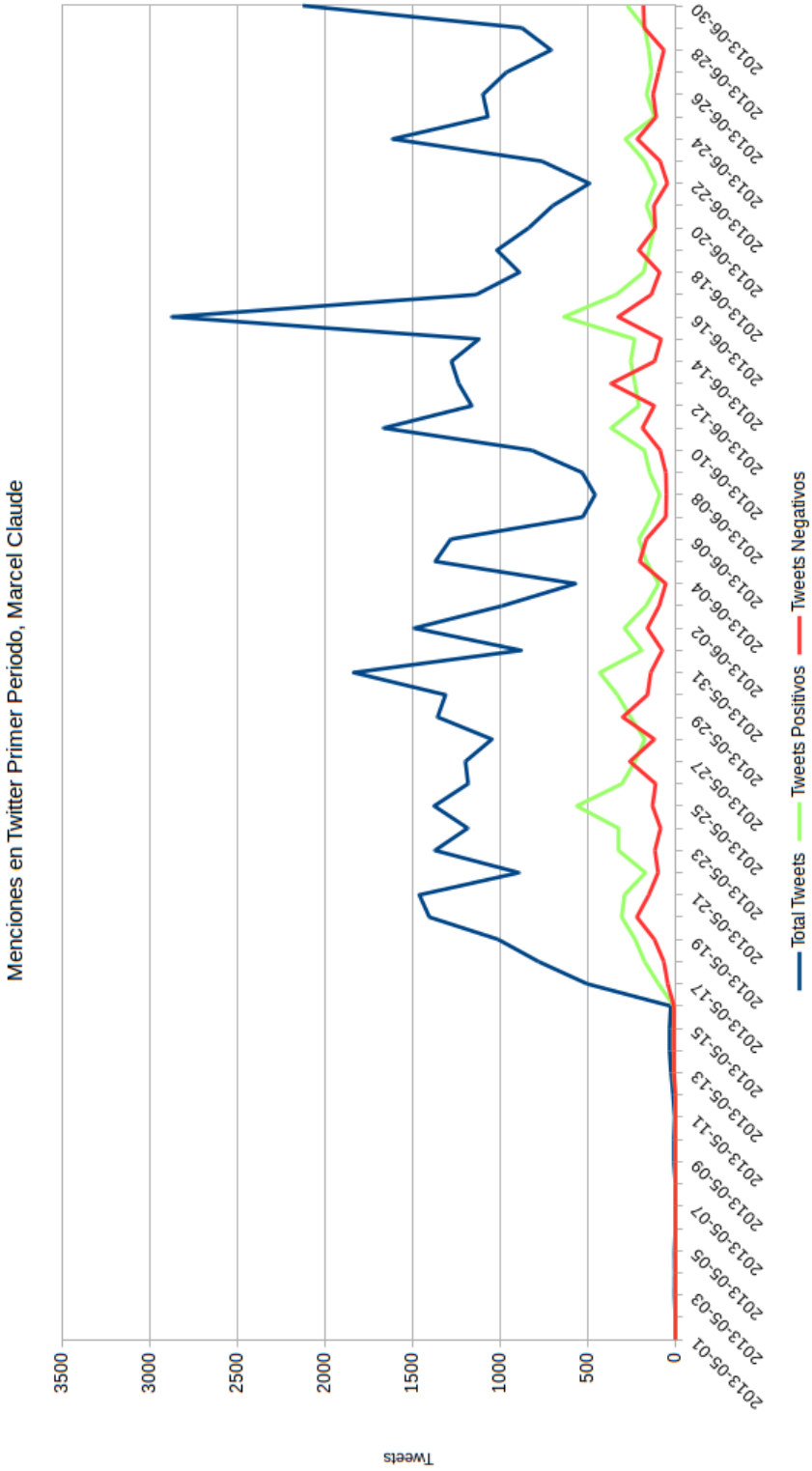

Figura C.7: Menciones en Twitter Primer Periodo, Marcel Claude

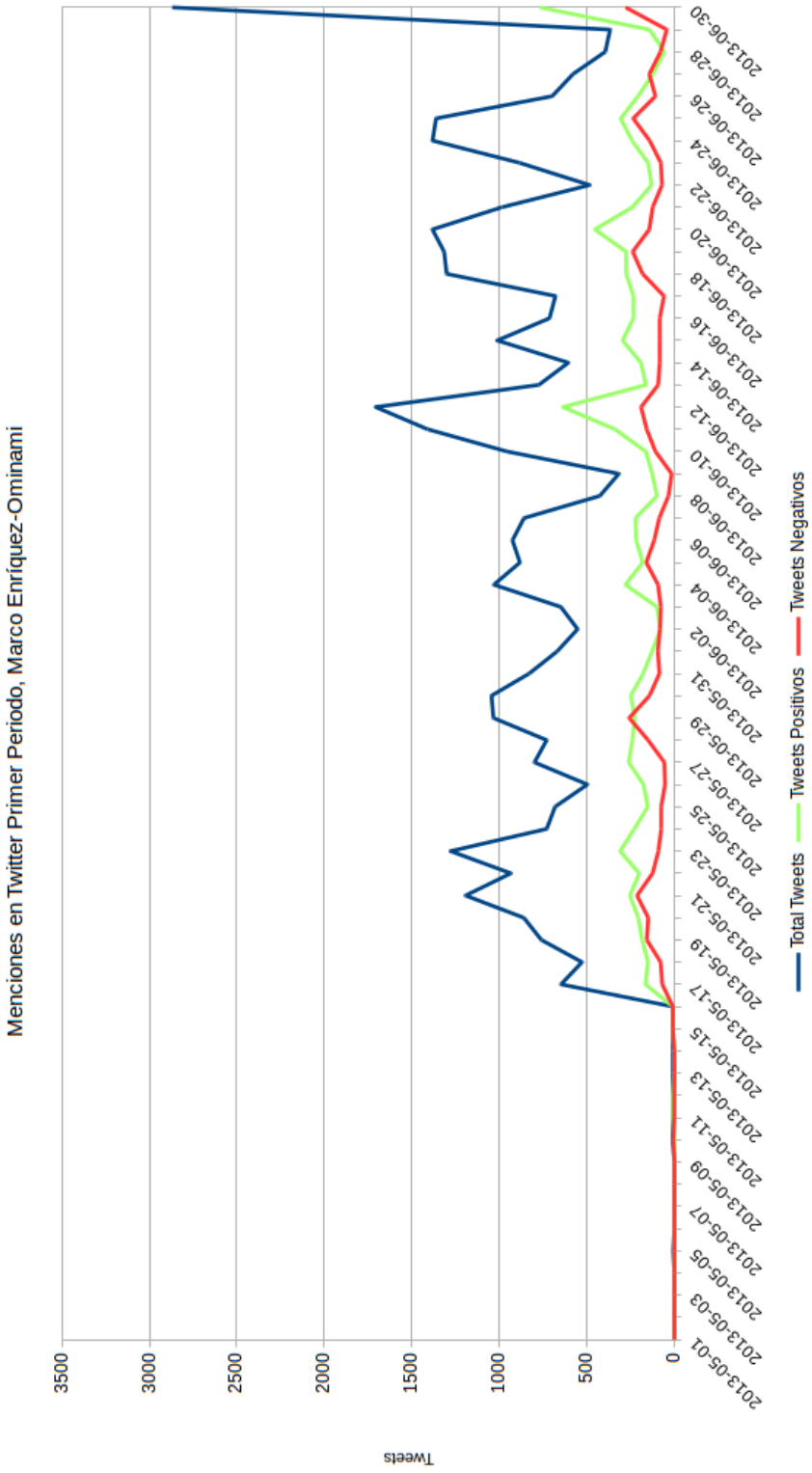

Figura C.8: Menciones en Twitter Primer Periodo, Marco Enriquez-Ominami

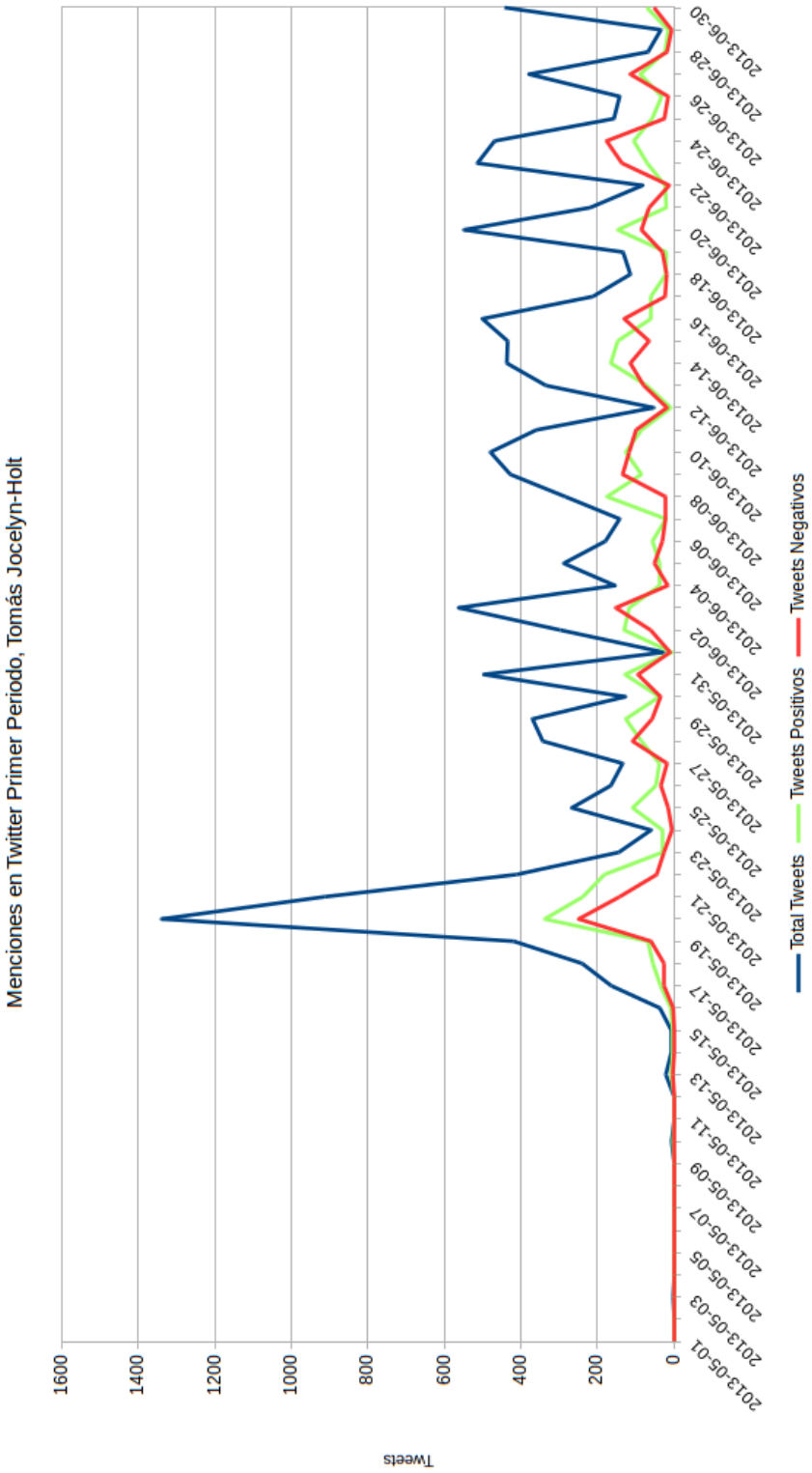

Figura C.9: Menciones en Twitter Primer Periodo, Tomás Jocelyn-Holt

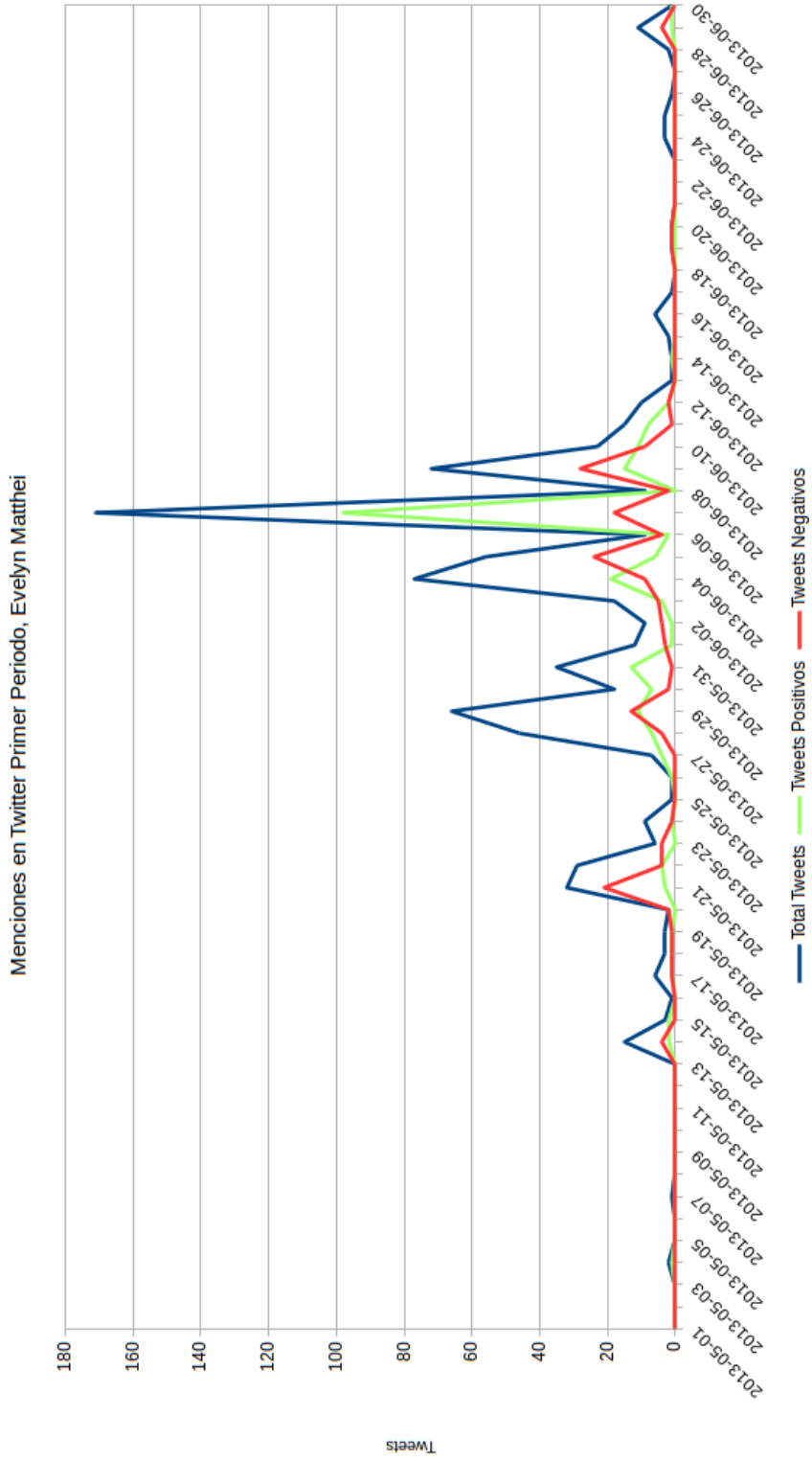

Figura C.10: Menciones en Twitter Primer Periodo, Evelyn Matthei

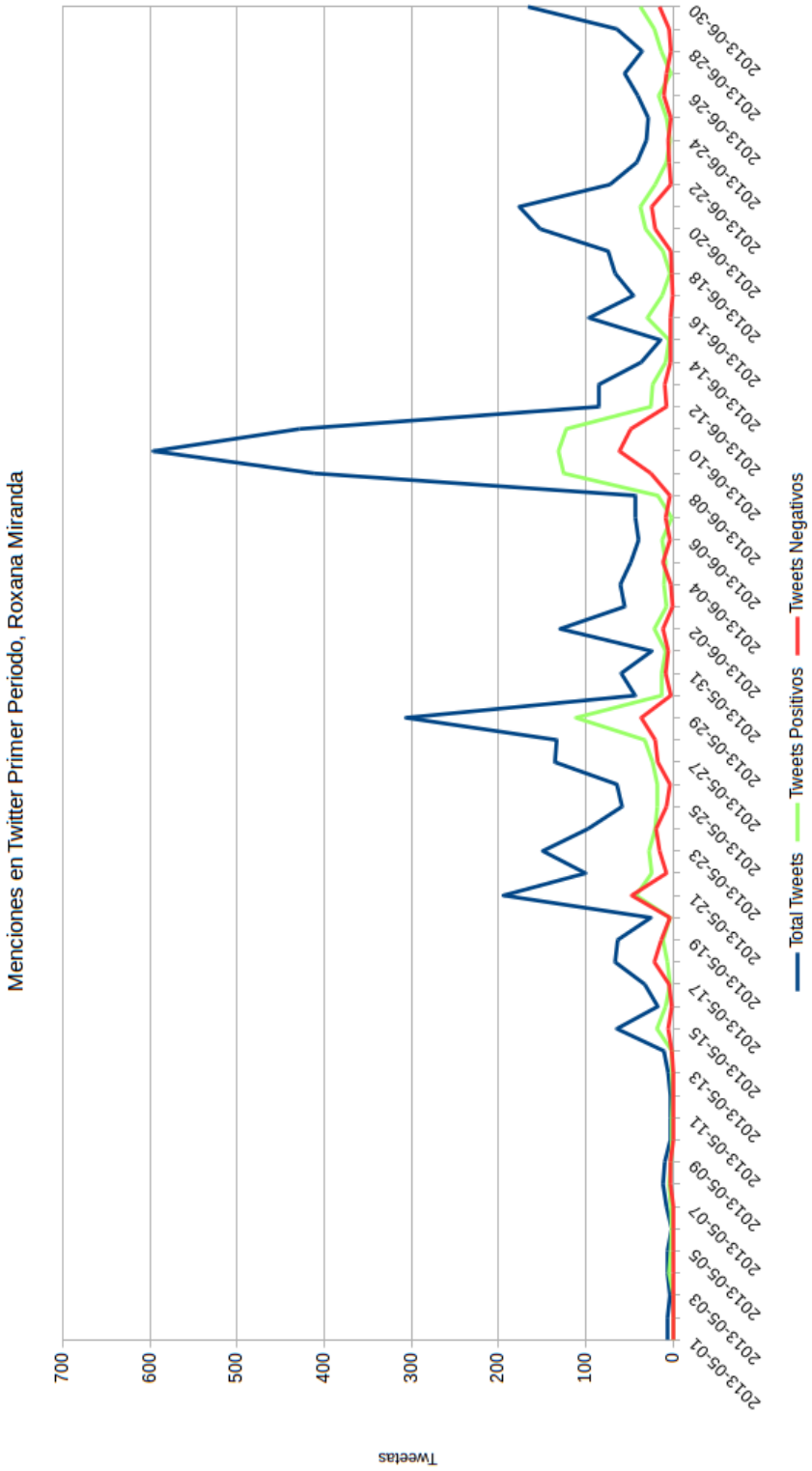

Figura C.11: Menciones en Twitter Primer Periodo, Roxana Miranda

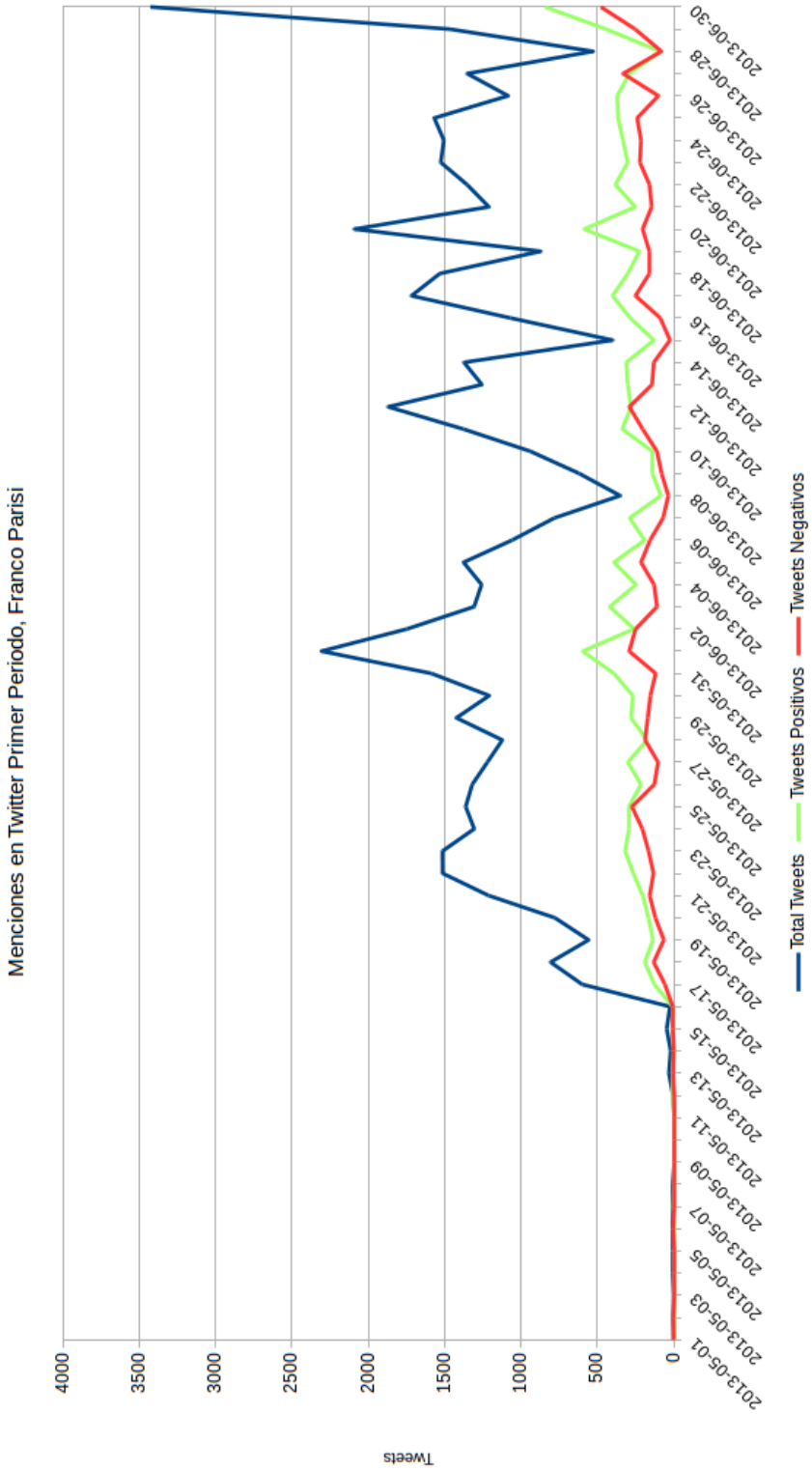

Figura C.12: Menciones en Twitter Primer Periodo, Franco Parisi

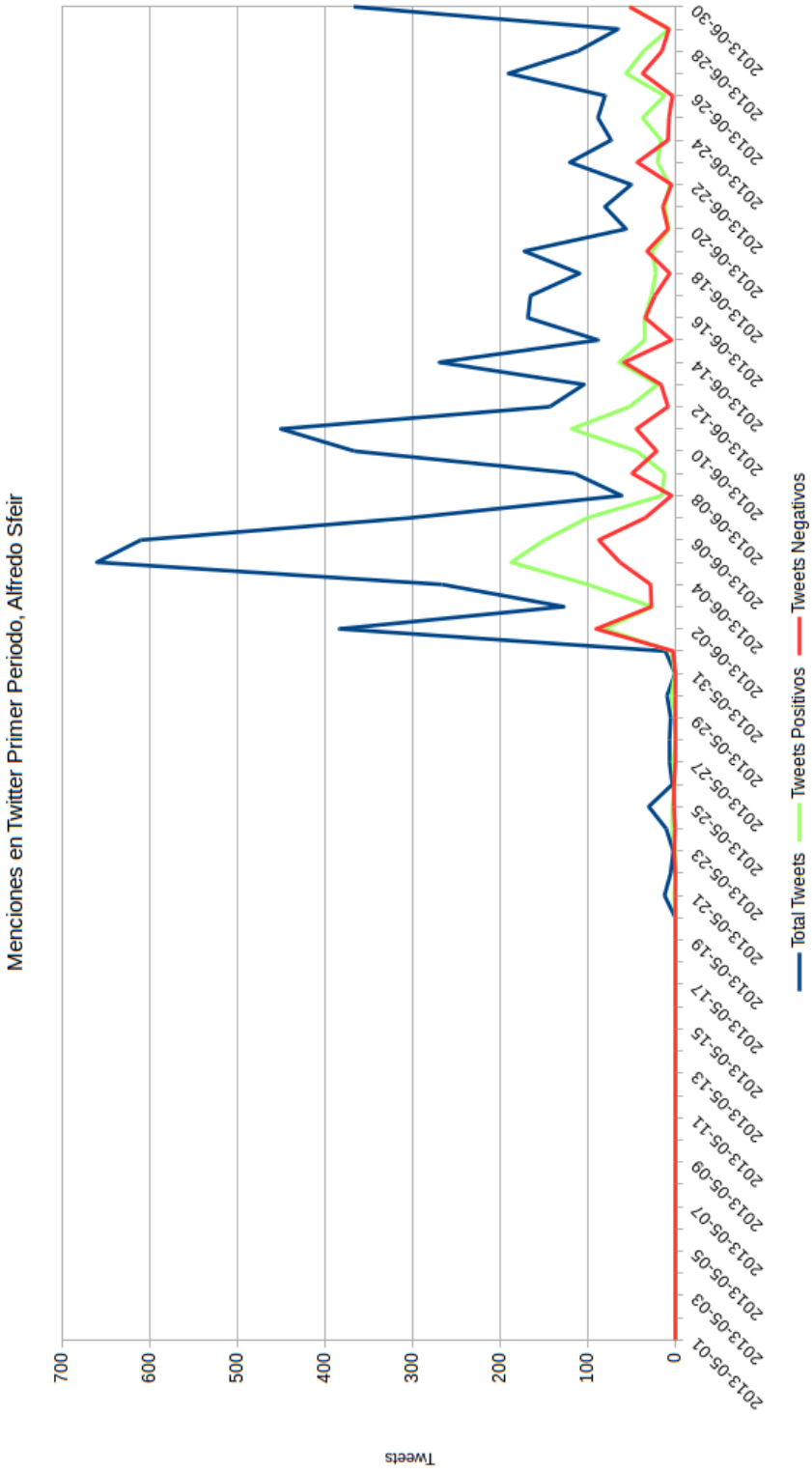

Figura C.13: Menciones en Twitter Primer Periodo, Alfredo Sfeir

C.5 Segundo Periodo

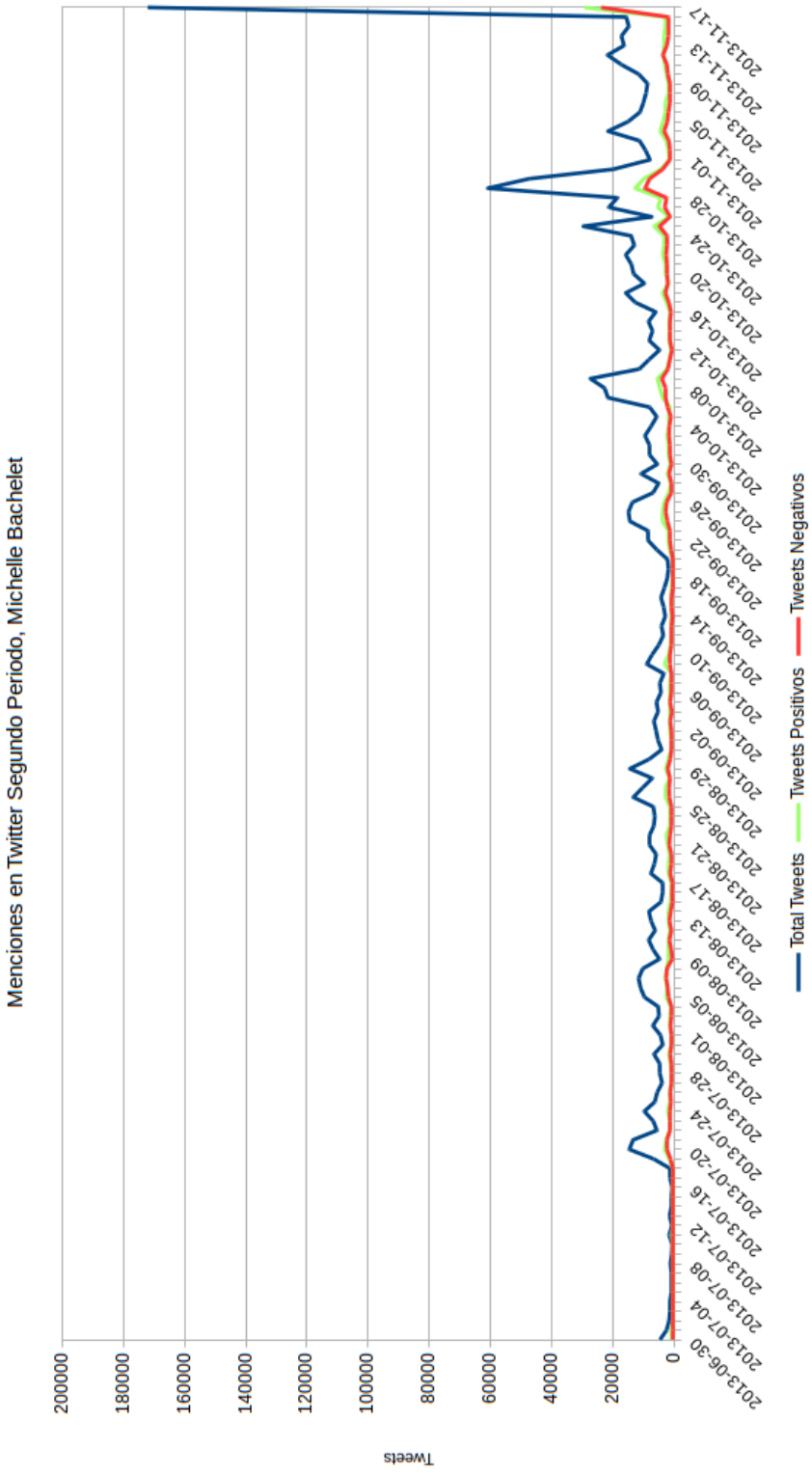

Figura C.14: Menciones en Twitter Segundo Periodo, Michelle Bachelet

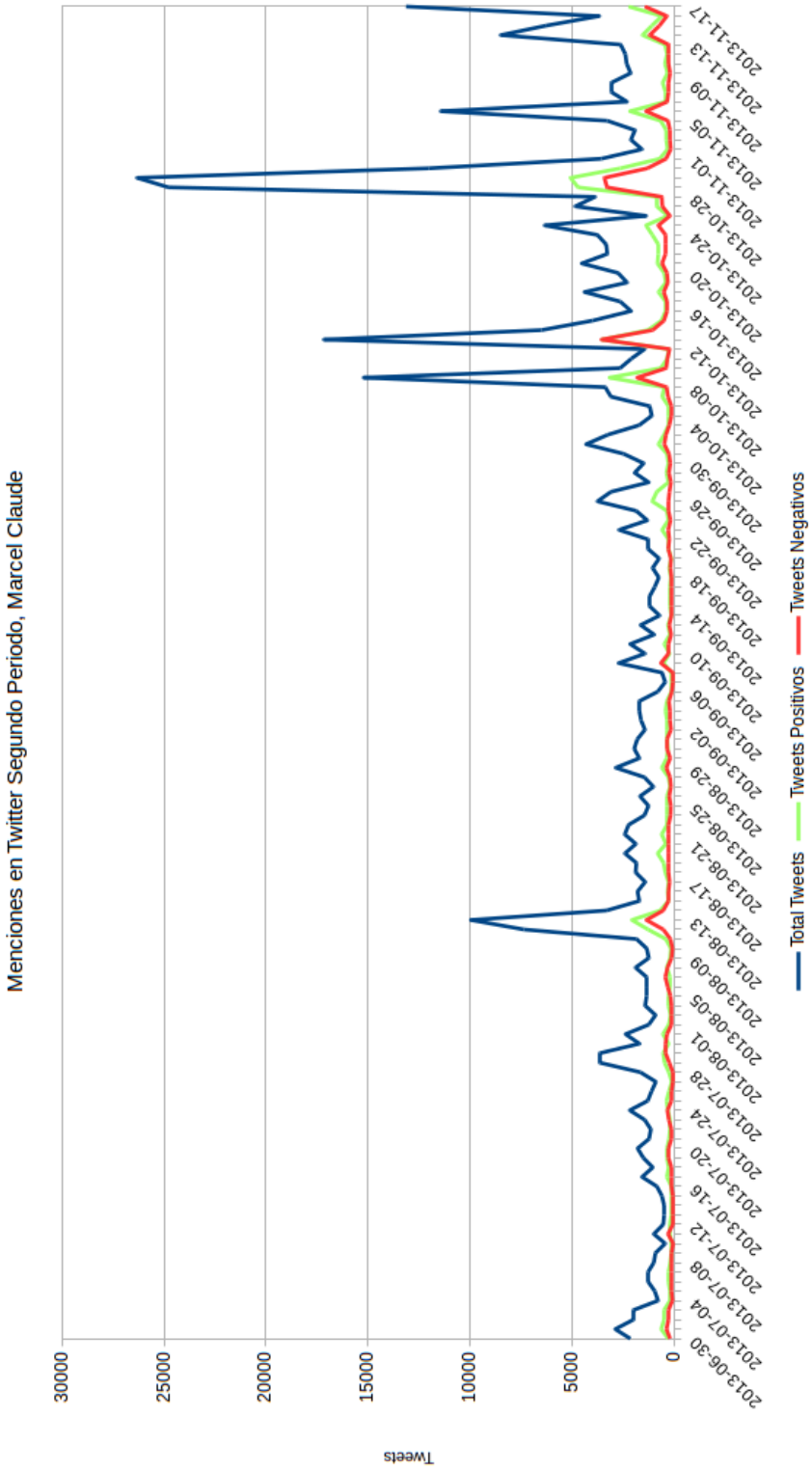

Figura C.15: Menciones en Twitter Segundo Periodo, Marcel Claude

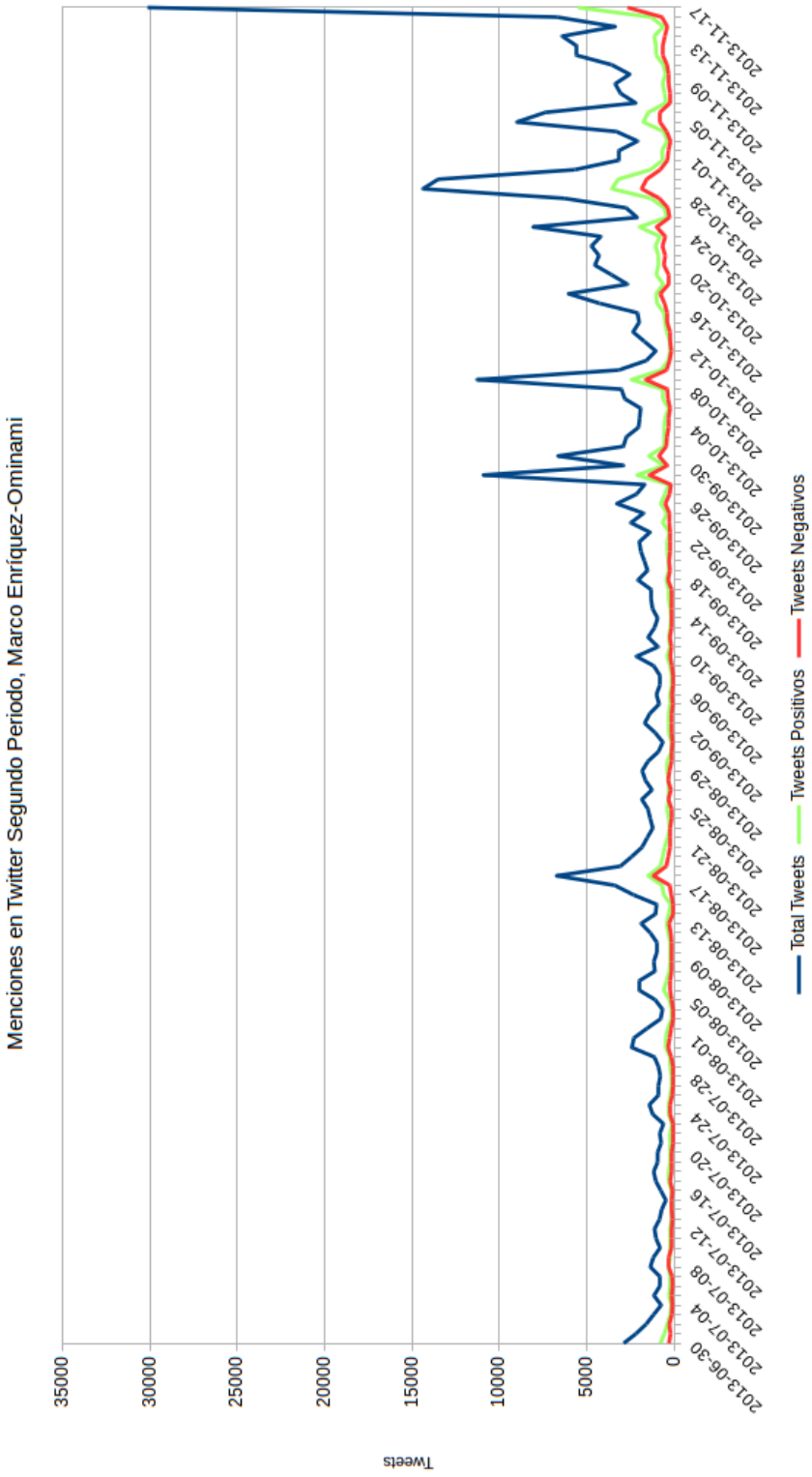

Figura C.16: Menciones en Twitter Segundo Periodo, Marco Enriquez-Ominami

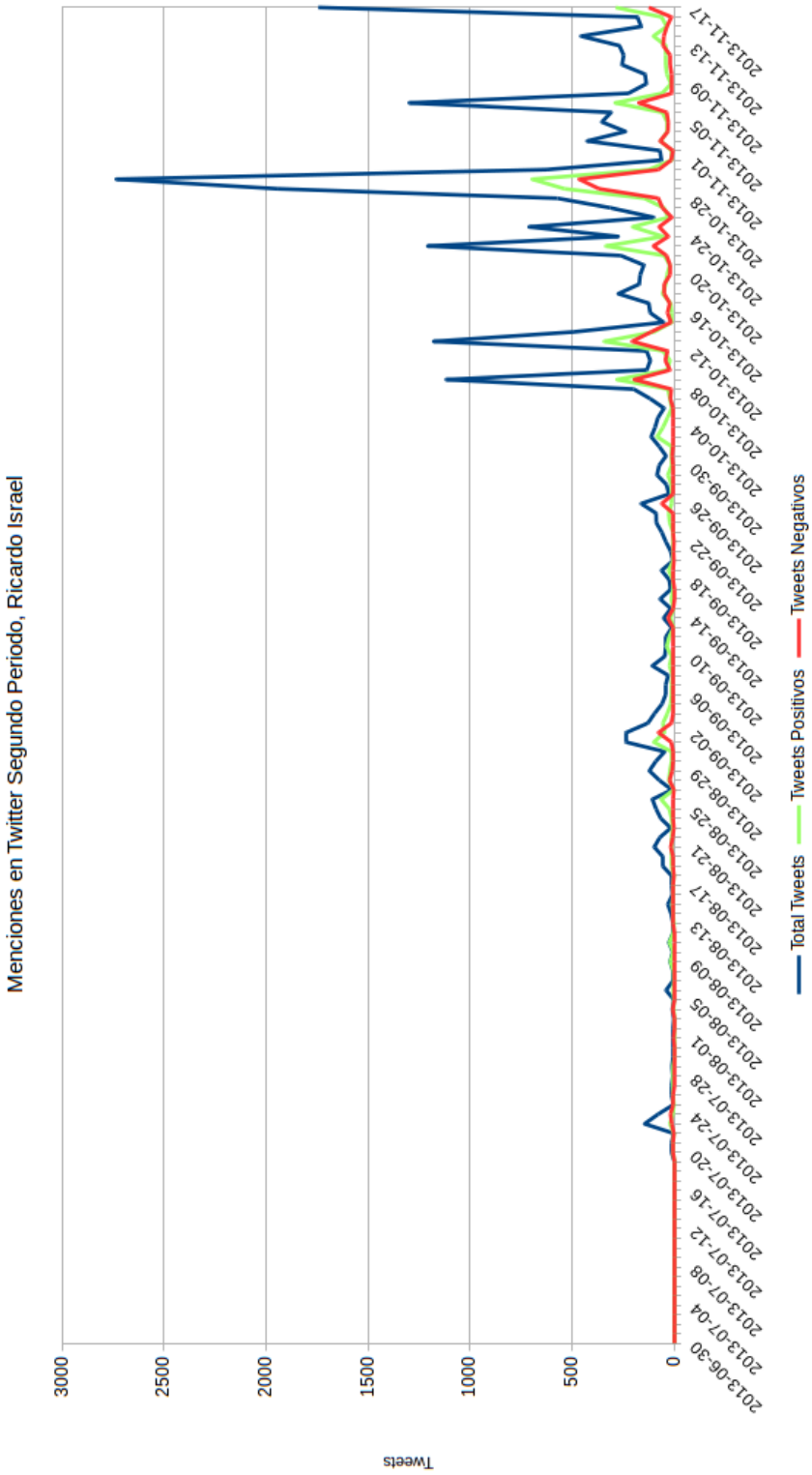

Figura C.17: Menciones en Twitter Segundo Periodo, Marco Ricardo Israel

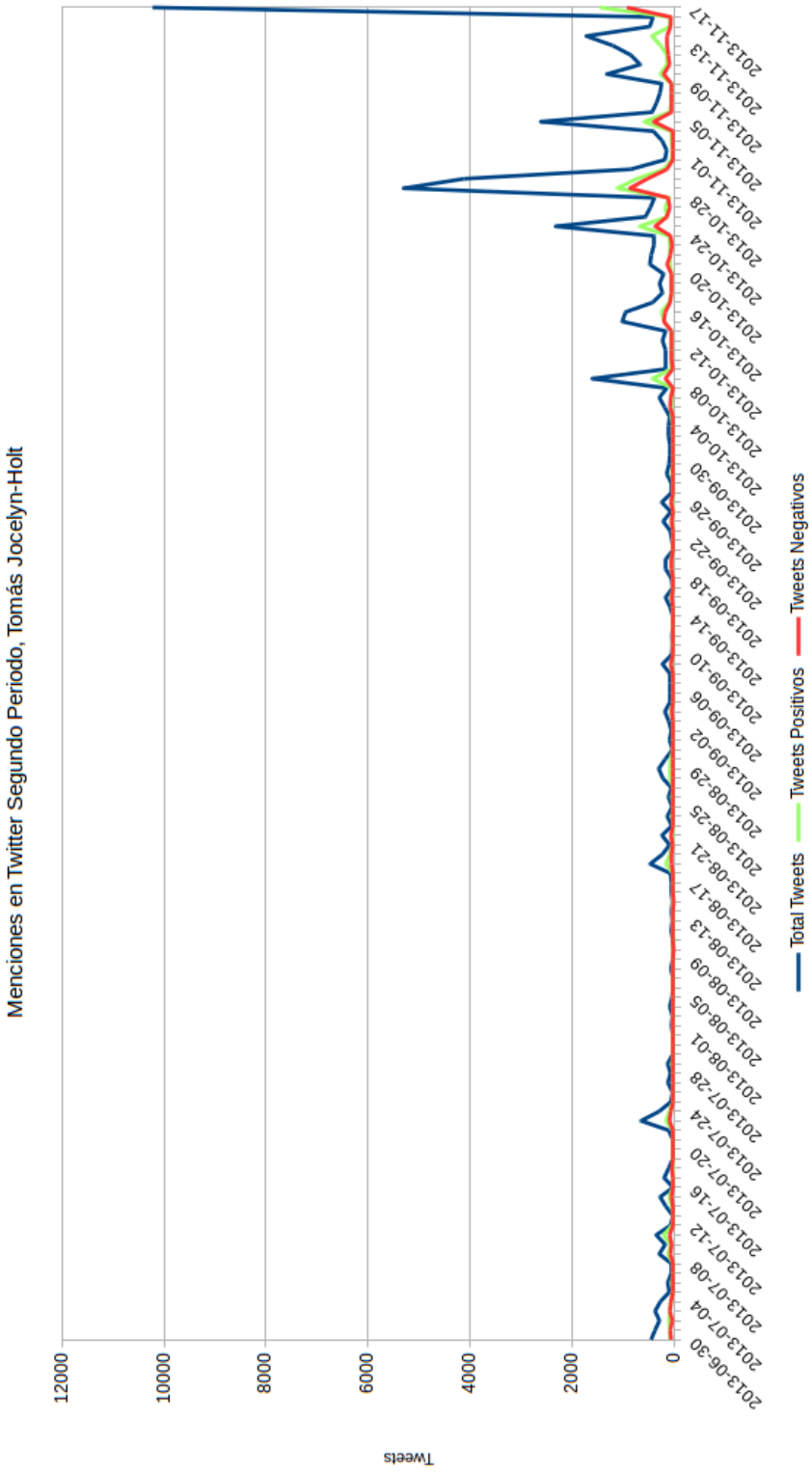

Figura C.18: Menciones en Twitter Segundo Periodo, Tomás Jocelyn-Holt

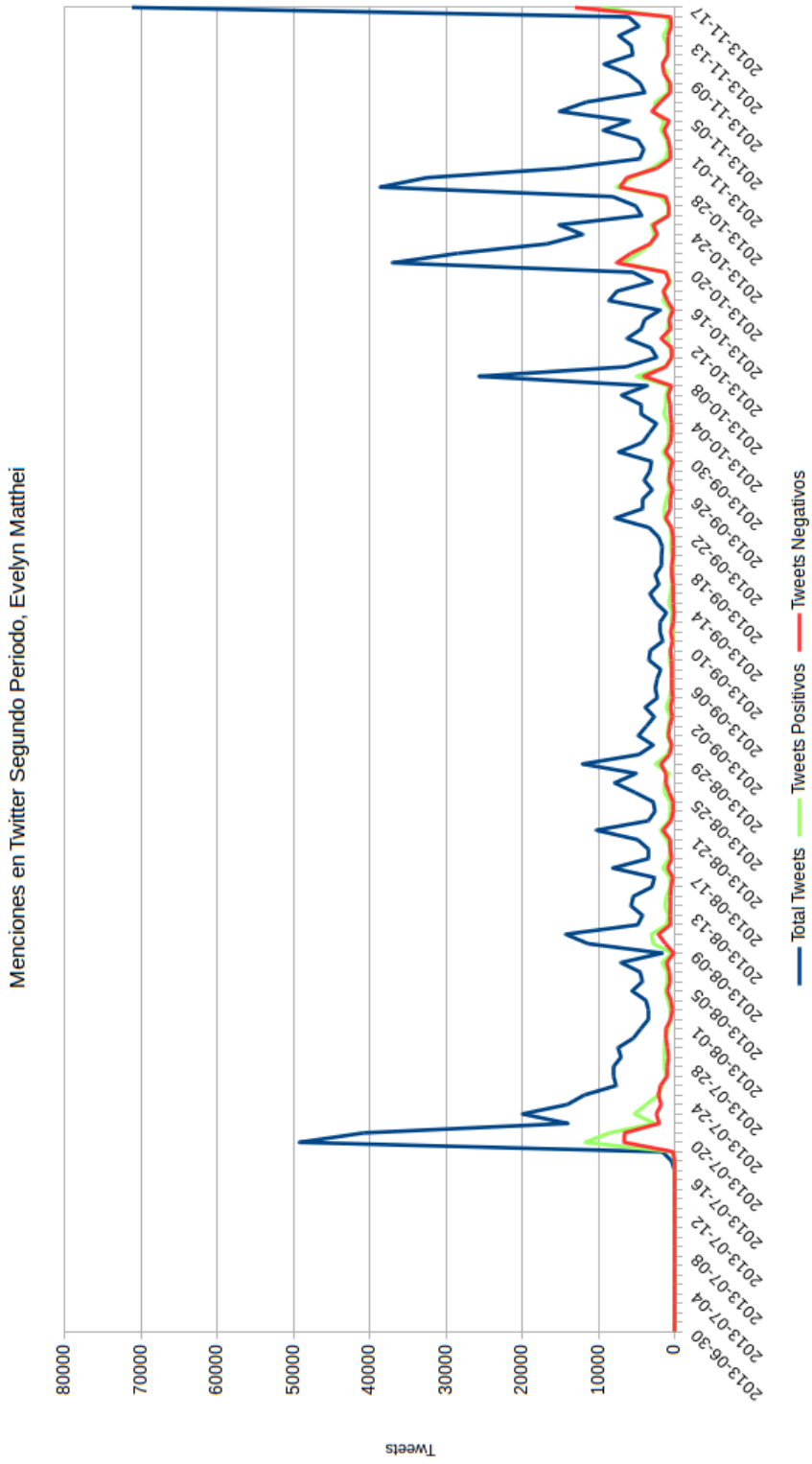

Figura C.19: Menciones en Twitter Segundo Período, Evelyn Matthei

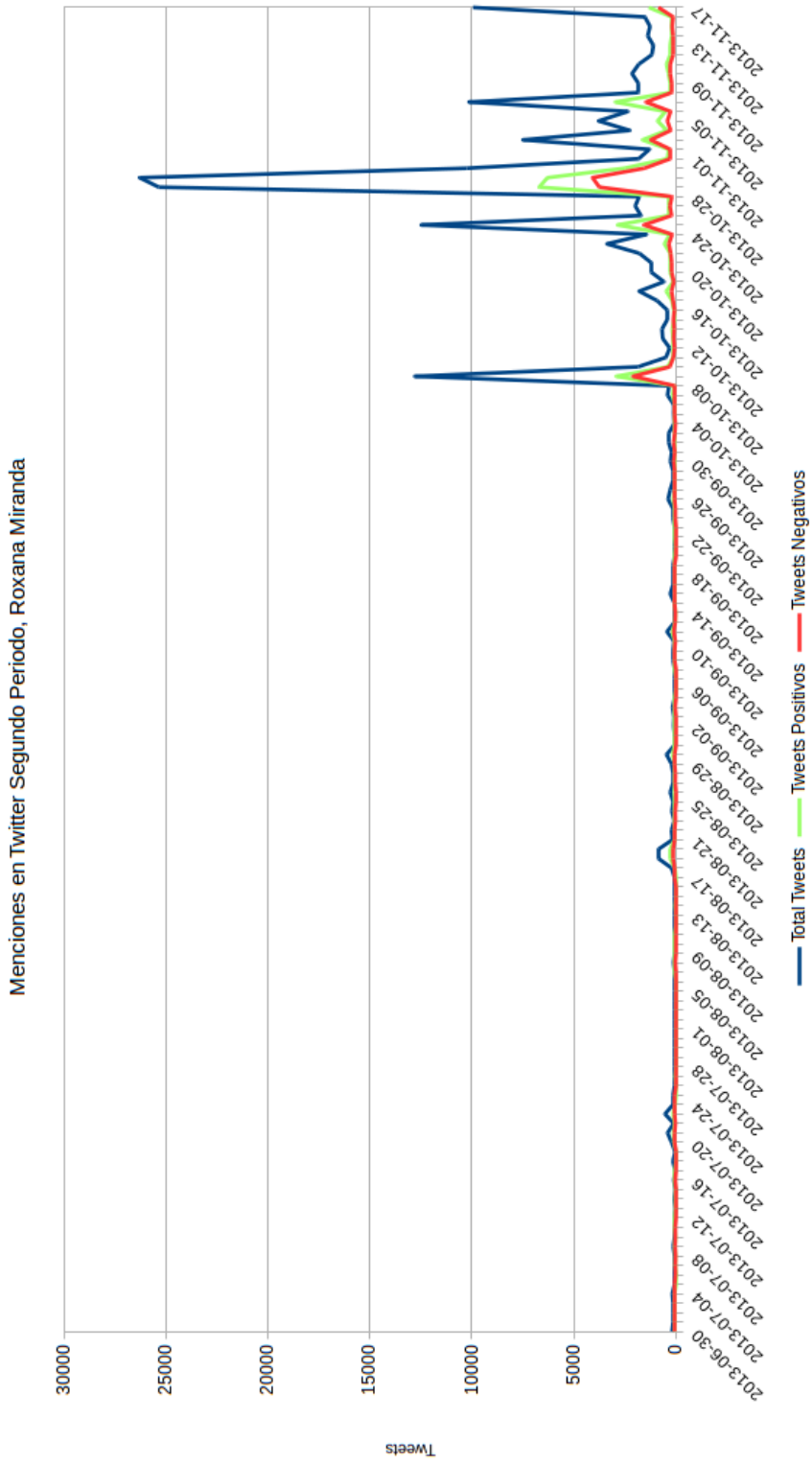

Figura C.20: Menciones en Twitter Segundo Período, Roxana Miranda

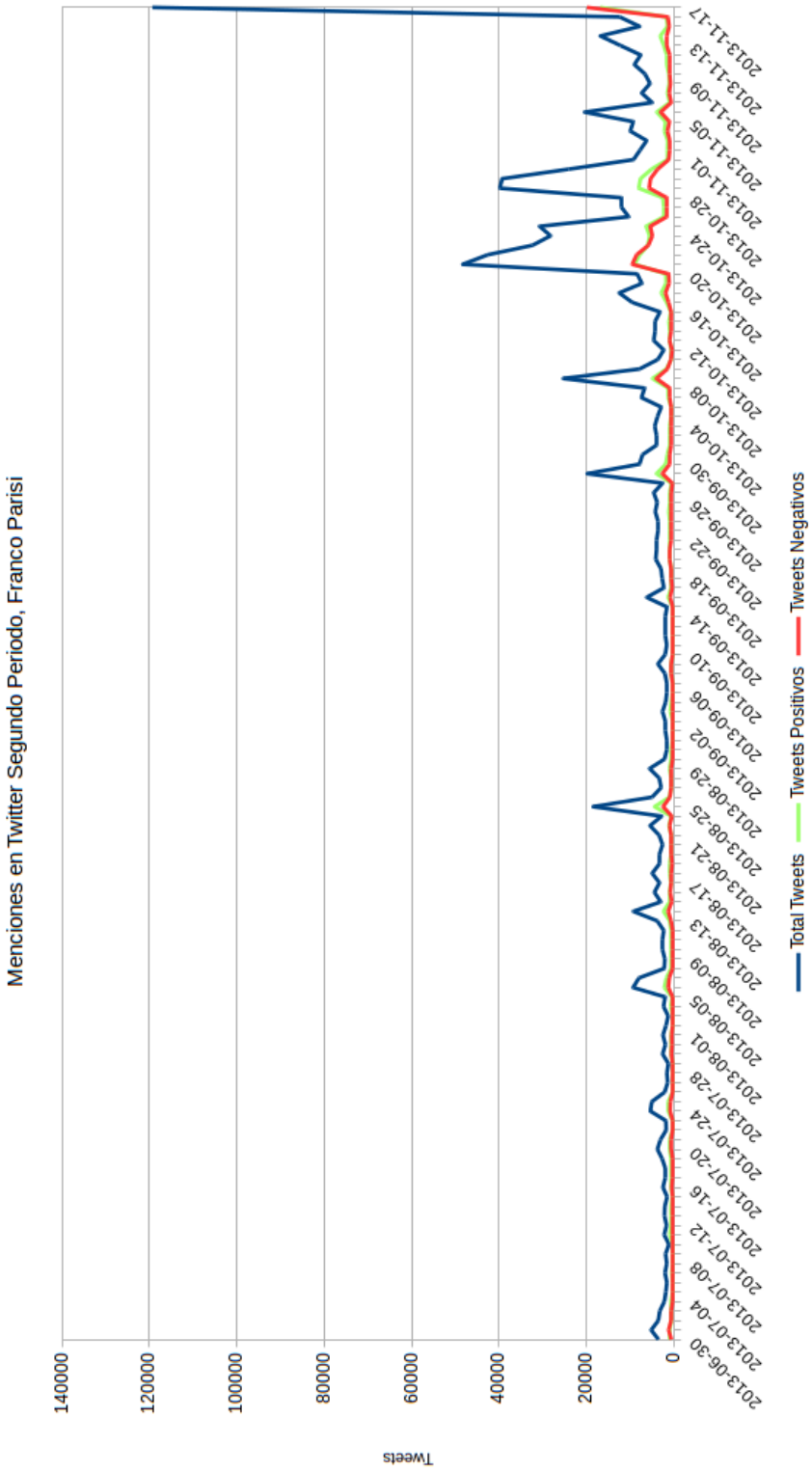

Figura C.21: Menciones en Twitter Segundo Período, Franco Parisi

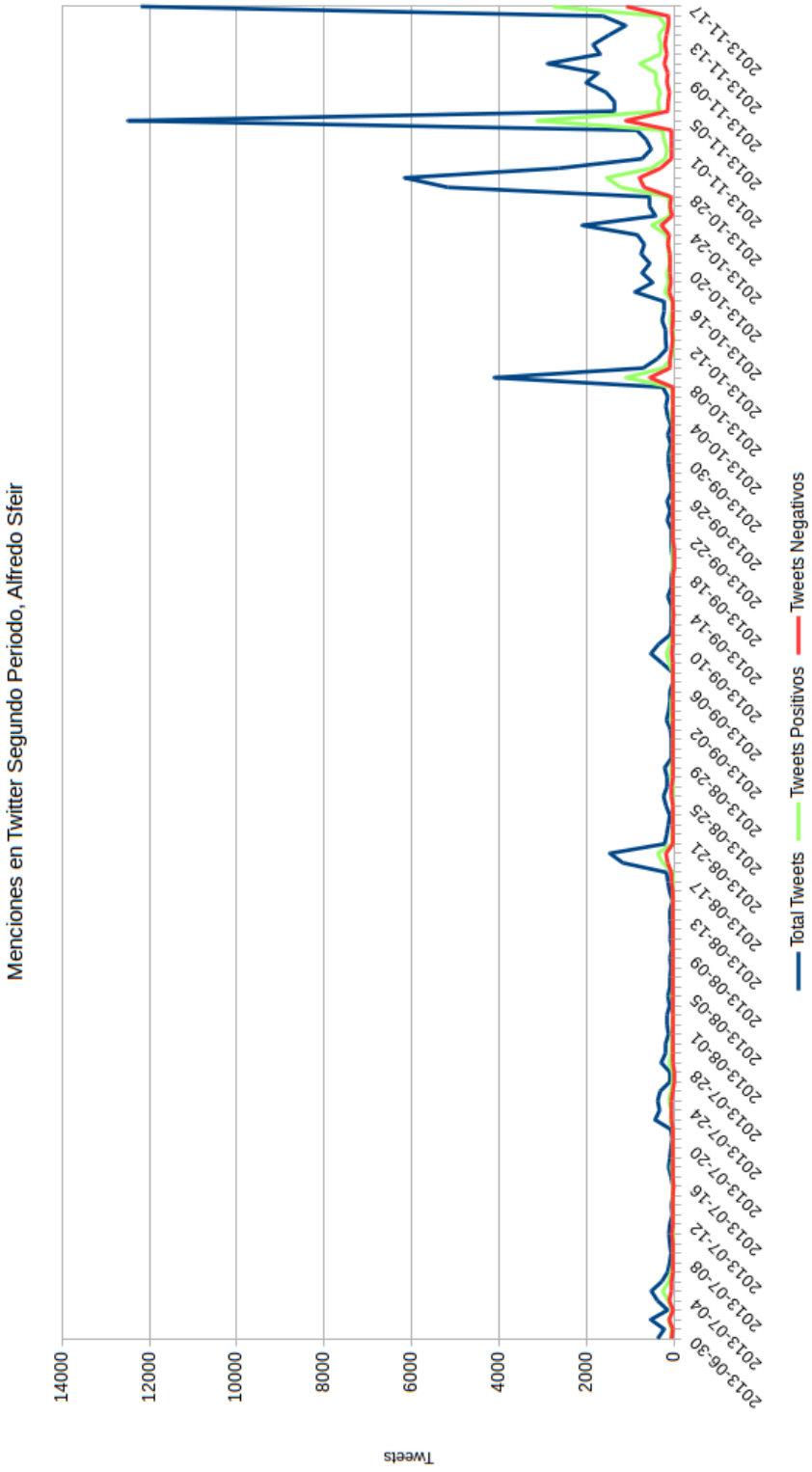

Figura C.22: Menciones en Twitter Segundo Período, Alfredo Sfeir

C.6 Tercer Periodo

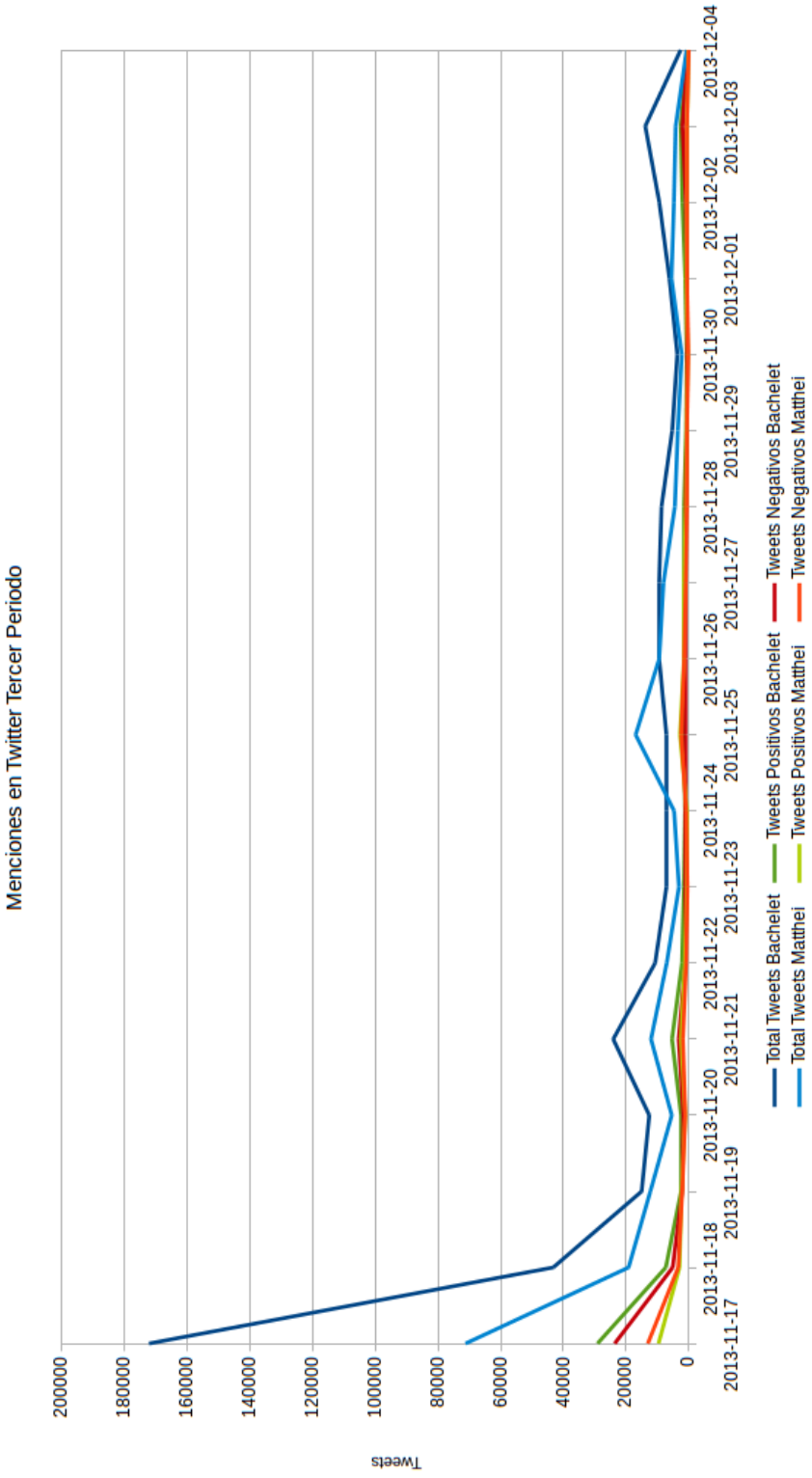

Figura C.23: Menciones en Twitter Tercer Período

C.7 Zipf

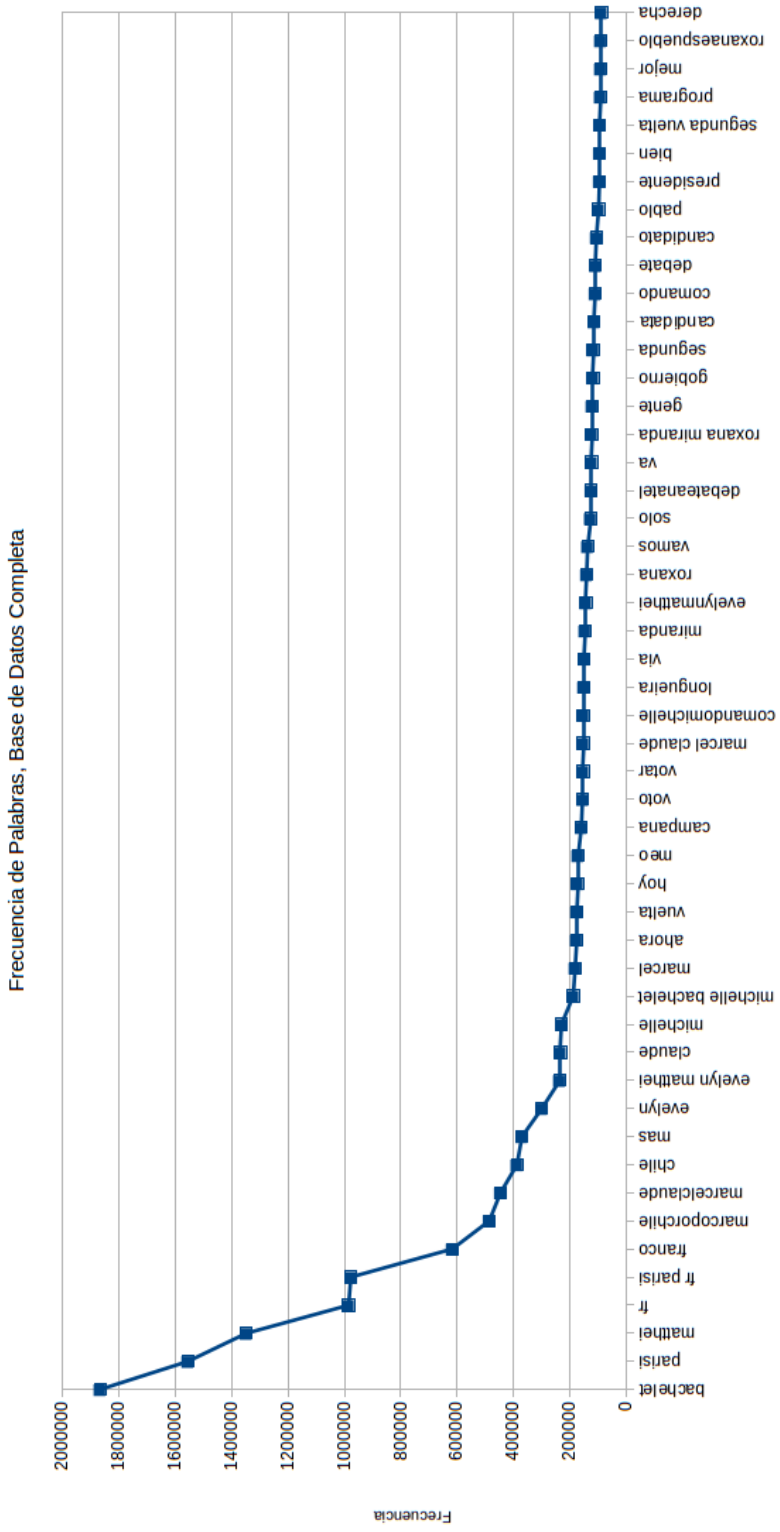

Figura C.24: Frecuencias de Palabras,Bases de Datos Completa

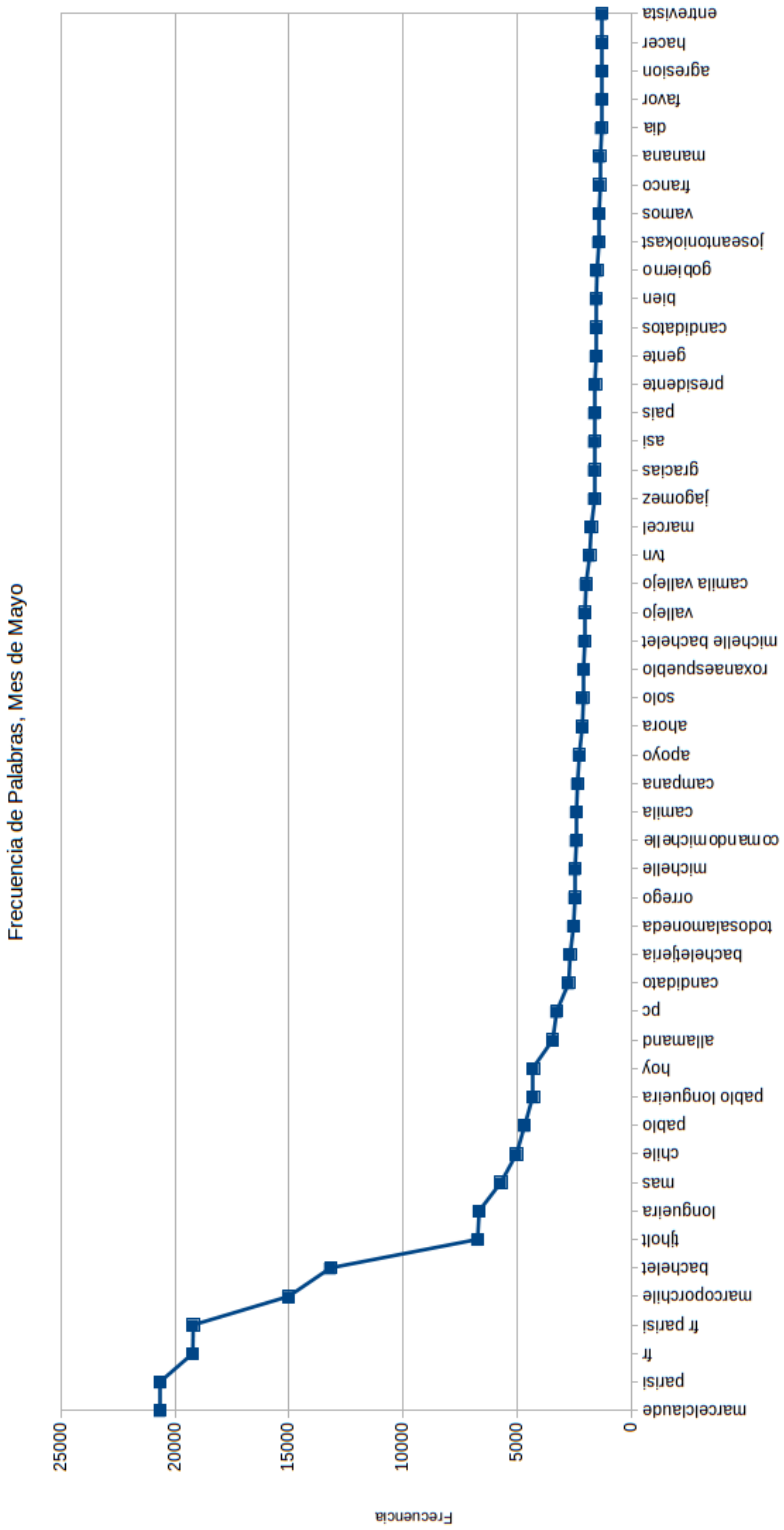

Figura C.25: Frecuencias de Palabras, Mes de Mayo

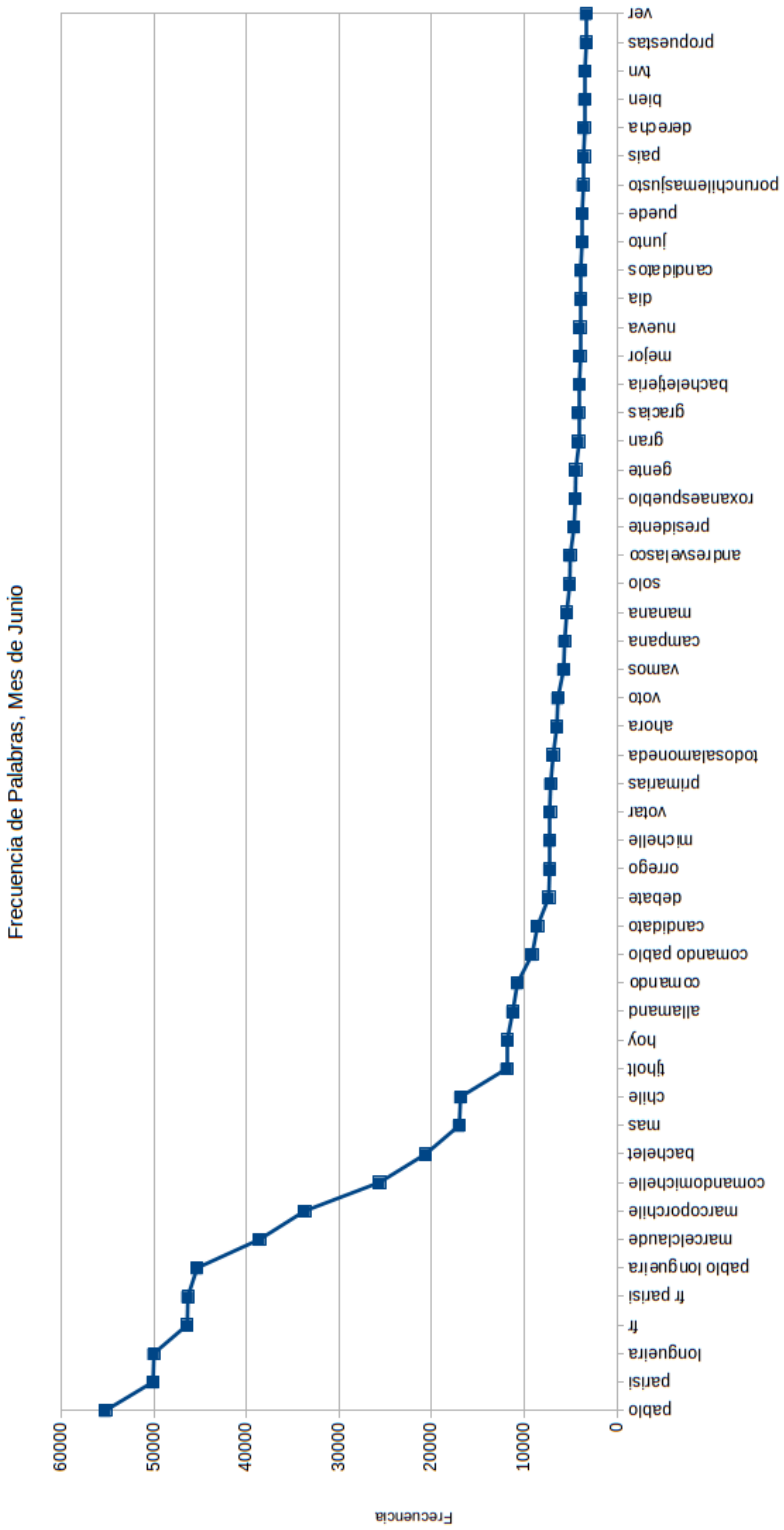

Figura C.26: Frecuencias de Palabras, Mes de Junio

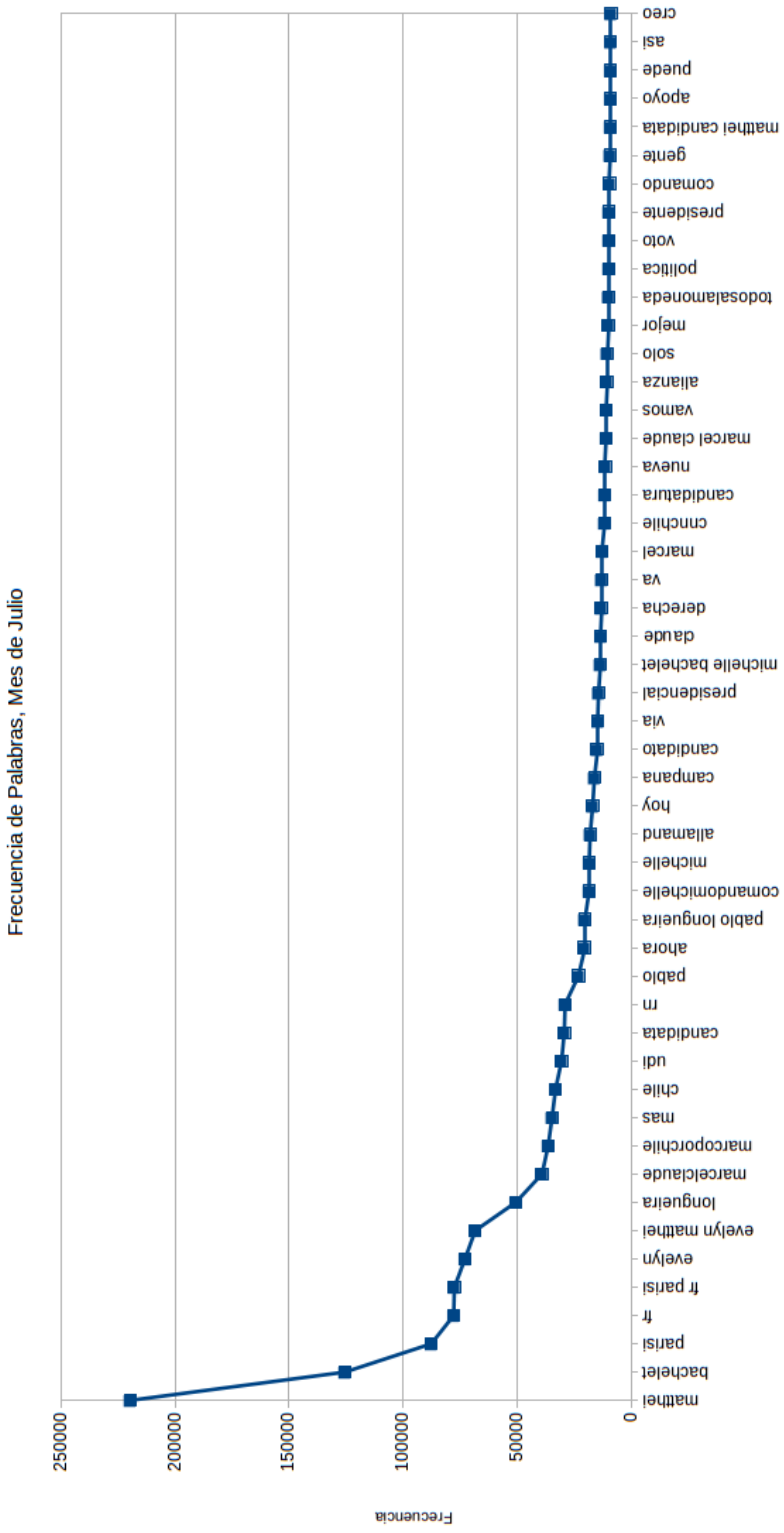

Figura C.27: Frecuencias de Palabras, Mes de Julio

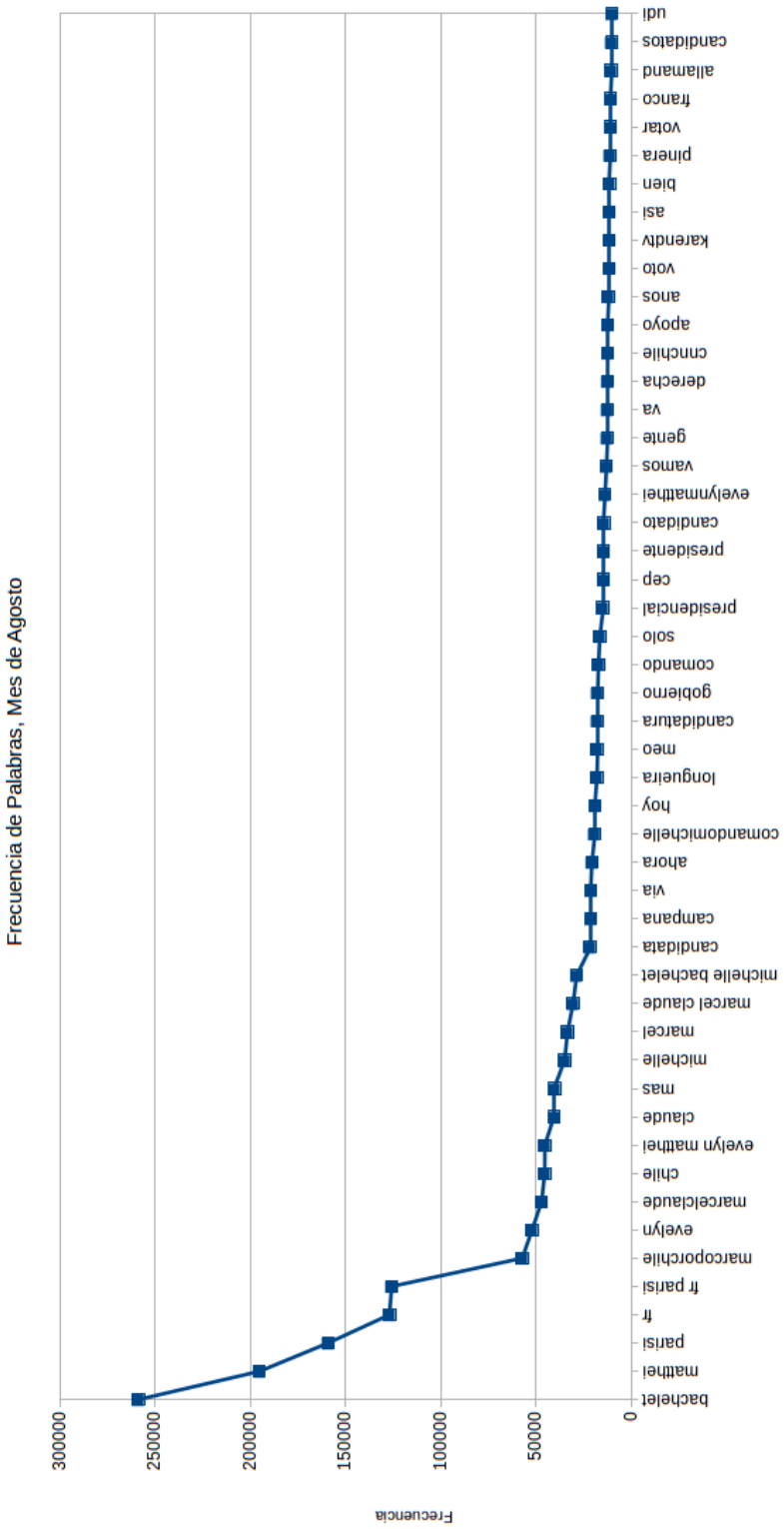

Figura C.28: Frecuencias de Palabras, Mes de Agosto

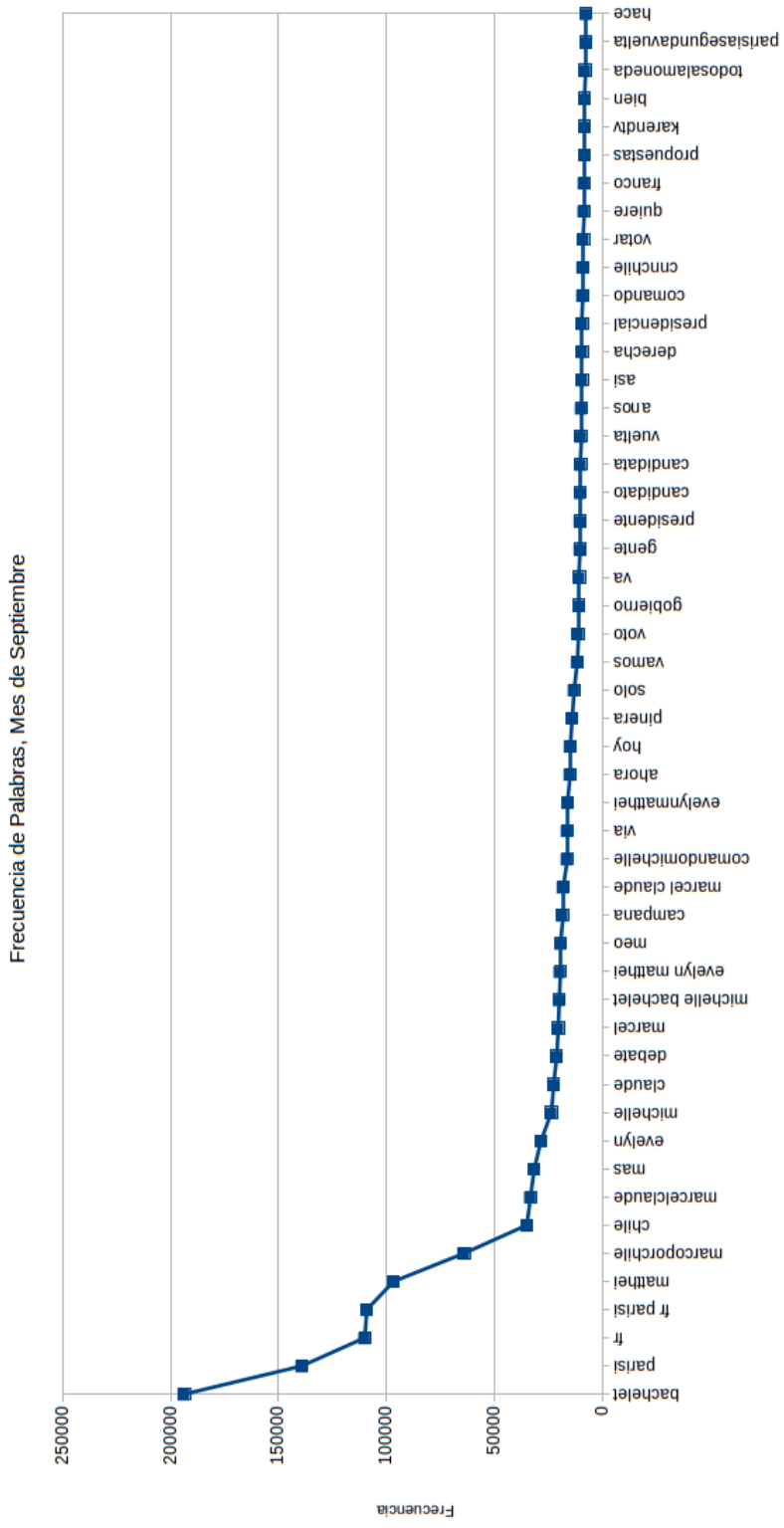

Figura C.29: Frecuencias de Palabras, Mes de Septiembre

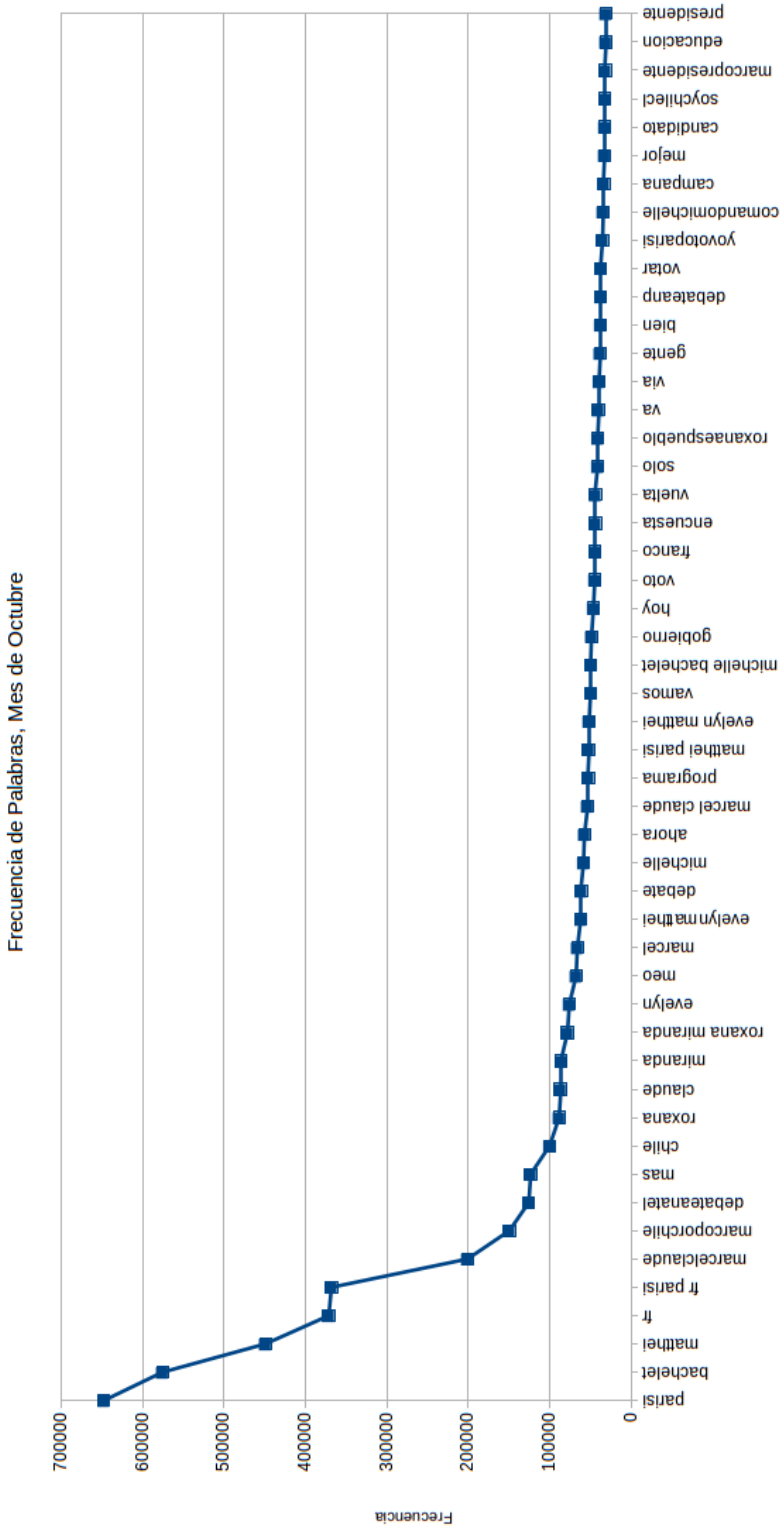

Figura C.30: Frecuencias de Palabras, Mes de Octubre

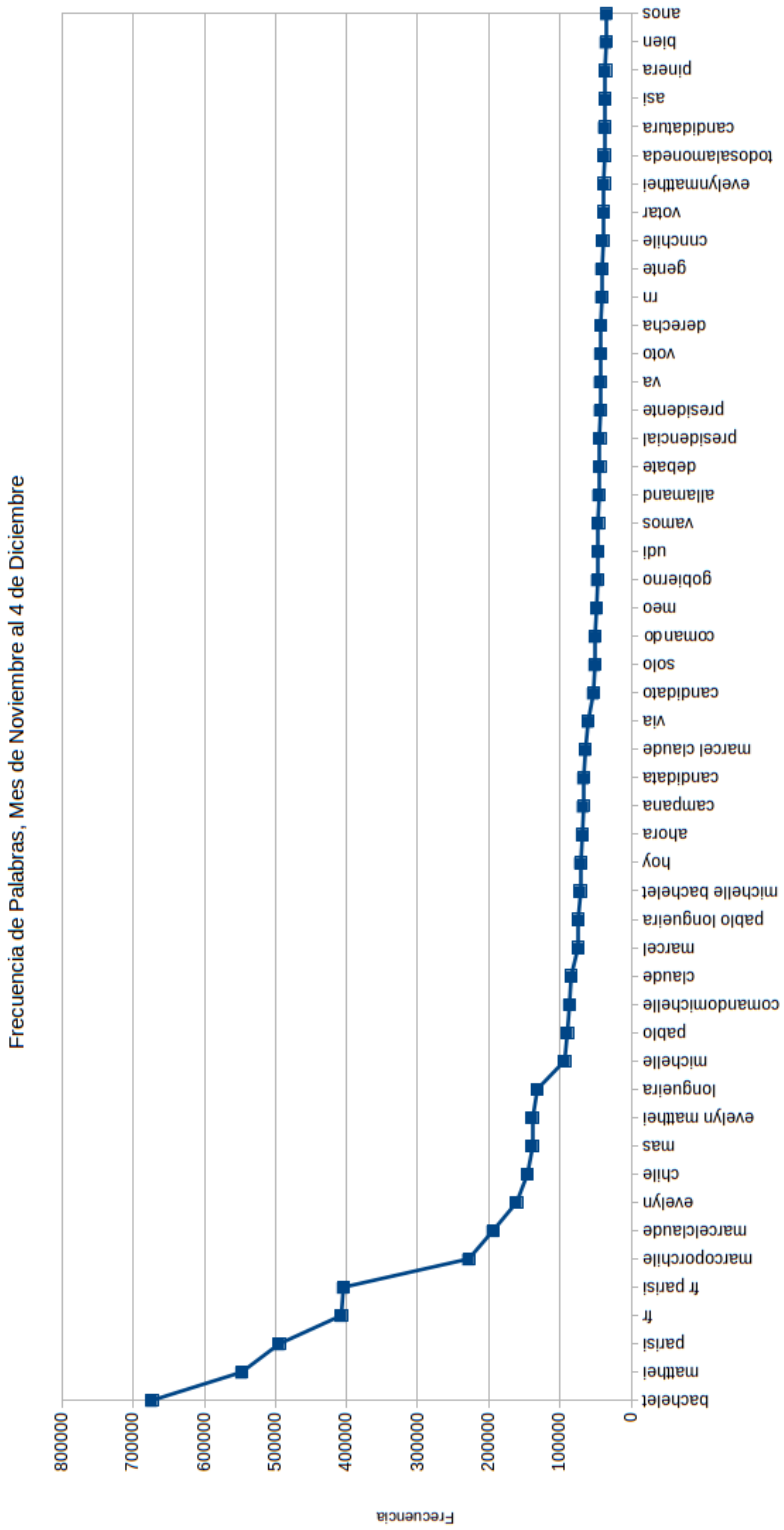

Figura C.31: Frecuencias de Palabras, Mes de Noviembre al 4 de Diciembre