UNIVERSIDAD DEL BÍO-BÍO

Facultad de Ciencias
Departamento de Estadística
Escuela Ingeniería Estadística

PROYECTO DE TÍTULO II, TITULADO:

"LEGITIMACIÓN Y DESLEGITIMACIÓN DEL CONFLICTO MAPUCHE DENTRO DE LA RED SOCIAL TWITTER"

MEMORIA PARA OPTAR AL TÍTULO DE:
INGENIERO ESTADÍSTICO

AUTOR: AVILEZ BOZO, JOSÉ MIGUEL

Profesor Guía: Faouzi Nadim. Tarik

Profesora CO-Guía: Díaz Costa. Elisabet

CONCEPCIÓN, Marzo de 2020

Indice

1. Introducción **5**

2. Objetivos del estudio **6**
2.1. Objetivo general 6
2.2. Objetivos especifivo 6

3. Marco Teórico **7**
3.1. Técnicas Exploratoria 7
3.1.1. Frecuencia de palabras 7
3.1.2. n-gramas 7
3.1.3. Análisis de frecuencia de palabras y documentos 8
3.1.4. Comparación del uso de palabras 8
3.1.5. Análisis de correlación (coeficiente phi) 9
3.2. Análisis de Sentimiento 9
3.3. Teoría de Grafos 9
3.3.1. Medidas de centralidad 10
3.4. Detección de comunidad 10
3.4.1. Algoritmo de Girvan y Newman 11
3.5. Análisis del Sentimiento en Twitter 12

4. Metodología **12**

5. Resultados **13**
5.1. Selección 13
5.2. Exploración 13
5.3. Limpieza 14
5.4. Transformación 15
5.5. Análisis exploratorio 15
5.6. Análisis de correlación (coeficiente de phi) 16
5.7. Relaciones entre palabras 17
5.8. Detección de grupos de palabras 21
5.9. Detección de comunidades 22
5.9.1. Detección de comunidades basada en el intervalo de borde (Newman-Girvan) 22
5.10. Diferencia entre grupos mediante correlación y log of odds ratio de las frecuencias 35
5.11. Análisis de sentimiento 51
5.12. Clasificación mediante variables relacionada con los usuarios 52

6. Conclusión **53**

7. Bibliografía **53**

Lista de tablas

Lista de gráfico

1. Estructura de un tweet. . . . 13
2. Tuits sin limpiar. . . . 15
3. Tuits aplicando la función de limpieza. . . . 15
4. Nube de Palabras de los términos con mayor frecuencia utilizado por los usuarios. . . . 16
5. Correlación entre algunos términos más frecuente. . . . 17
6. Recuento de los principales bigramas con una frecuencia mínima de 200. . . . 18
7. Grafo de no dirigido entre los términos con mayor frecuencia. . . . 19
8. Subgrafo de la componente con mayor grado. . . . 20
9. Subgrafo con grupos de palabras. . . . 21
10. Grafo de retweets 22
11. Sub Grafo de Retweets. . . . 23
12. Términos frecuentes por grupos 24
13. TF-IDF por grupos. . . . 25
14. tf-idf de bigramas por grupos 26
15. Frecuencia de bigramas por grupos 27
16. Distribución de polaridad grupo 1. . . . 29
17. Palabras positivas y negativas más comunes del grupo 1. . . . 29
18. Distribución de polaridad grupo 2. . . . 30
19. Palabras positivas y negativas más comunes del grupo 2. . . . 30
20. Distribución de polaridad grupo 3. . . . 31
21. Palabras positivas y negativas más comunes del grupo 3. . . . 31
22. Distribución de polaridad grupo 4. . . . 32
23. Palabras positivas y negativas más comunes del grupo 4. . . . 32
24. Distribución de polaridad grupo 5. . . . 33
25. Palabras positivas y negativas más comunes del grupo 5. . . . 33
26. Distribución de polaridad grupo 6. . . . 34
27. Palabras positivas y negativas más comunes del grupo 6. . . . 34
28. Comparación de las frecuencias de palabras entre grupo 1 y 2 mediante correlación. . . . 35
29. Comparación de las frecuencias de palabras entre grupo 1 y 2 mediante log of odds ratio de las frecuencias 36
30. Comparación de las frecuencias de palabras entre grupo 1 y 3 mediante correlación. . . . 37
31. Comparación de las frecuencias de palabras entre grupo 1 y 3 mediante log of odds ratio de las frecuencias 37
32. Comparación de las frecuencias de palabras entre grupo 1 y 4 mediante correlación. . . . 38
33. Comparación de las frecuencias de palabras entre grupo 1 y 4 mediante log of odds ratio de las frecuencias 38
34. Comparación de las frecuencias de palabras entre grupo 1 y 5 mediante correlación. . . . 39
35. Comparación de las frecuencias de palabras entre grupo 1 y 5 mediante log of odds ratio de las frecuencias 39

36. Comparación de las frecuencias de palabras entre grupo 1 y 6 mediante correlación. 40
37. Comparación de las frecuencias de palabras entre grupo 1 y 6 mediante log of odds ratio de las frecuencias 40
38. Comparación de las frecuencias de palabras entre grupo 2 y 3 mediante correlación. 41
39. Comparación de las frecuencias de palabras entre grupo 2 y 3 mediante log of odds ratio de las frecuencias 41
40. Comparación de las frecuencias de palabras entre grupo 2 y 4 mediante correlación. 42
41. Comparación de las frecuencias de palabras entre grupo 2 y 4 mediante log of odds ratio de las frecuencias 42
42. Comparación de las frecuencias de palabras entre grupo 2 y 5 mediante correlación. 43
43. Comparación de las frecuencias de palabras entre grupo 2 y 5 mediante log of odds ratio de las frecuencias 43
44. Comparación de las frecuencias de palabras entre grupo 2 y 6 mediante correlación. 44
45. Comparación de las frecuencias de palabras entre grupo 2 y 6 mediante log of odds ratio de las frecuencias 44
46. Comparación de las frecuencias de palabras entre grupo 3 y 4 mediante correlación. 45
47. Comparación de las frecuencias de palabras entre grupo 3 y 4 mediante log of odds ratio de las frecuencias 45
48. Comparación de las frecuencias de palabras entre grupo 3 y 5 mediante correlación. 46
49. Comparación de las frecuencias de palabras entre grupo 3 y 5 mediante log of odds ratio de las frecuencias 46
50. Comparación de las frecuencias de palabras entre grupo 3 y 6 mediante correlación. 47
51. Comparación de las frecuencias de palabras entre grupo 3 y 6 mediante log of odds ratio de las frecuencias 47
52. Comparación de las frecuencias de palabras entre grupo 4 y 5 mediante correlación. 48
53. Comparación de las frecuencias de palabras entre grupo 4 y 5 mediante log of odds ratio de las frecuencias 48
54. Comparación de las frecuencias de palabras entre grupo 4 y 6 mediante correlación. 49
55. Comparación de las frecuencias de palabras entre grupo 4 y 6 mediante log of odds ratio de las frecuencias 49
56. Comparación de las frecuencias de palabras entre grupo 5 y 6 mediante correlación. 50
57. Comparación de las frecuencias de palabras entre grupo 5 y 6 mediante log of odds ratio de las frecuencias 50
58. mediante el paquete syzhet. 51
59. Árbol de decisión 52

1. Introducción

Chile es un país sudamericano multicultural, en él conviven la nación chilena junto con diferentes naciones, correspondientes a distintos pueblos originarios. Dentro de estas naciones el pueblo mapuche ha tenido históricamente una mayor visibilización. Esto se ha debido a la confrontación de sus derechos indígenas, actualmente ante el Estado chileno, y previamente, ante sus colonizadores españoles (Bengoa, 2002). Dentro de este contexto, se ubica en su historia reciente el caso del comunero mapuche Camilo Catrillanca, quien fue muerto de un disparo en su cabeza por integrantes de un grupo militar de la policía chilena . Este caso fue muy bullado en el país y alcanzó incluso repercusión internacional, producto de que la policía realizó un montaje con las evidencias sobre su muerte, con el fin de desligar sus responsabilidades ante el asesinato, situación fuera de todo protocolo policial y uso racional de la fuerza (El Mostrador, 2019). Dada la importancia de este caso en relación al conflicto entre el Estado chileno y el pueblo mapuche, se evidencia en la literatura que las investigaciones abordan esta temática, pero escasamente se han analizado las opiniones de las redes sociales como Twitter sobre este tópico, e incluso no existen investigaciones que utilicen el Análisis de Sentimientos en esta red social y que se asocien a este tipo de problemática, sin embargo existen otras investigaciones sobre la misma problemática que recogen análisis de videos en Youtube (Maldonado, 2011). De este modo, surge el objetivo de la investigación que busca caracterizar los grupos de influencia en esta red y sus posibles interrelaciones, por medio del análisis de sentimientos. El estudio de Análisis de Sentimientos, enmarcado en el procesamiento del Lenguaje Natural es entendido como: 'Range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications' (Liddy, 2001, p.1). Así, el objetivo del tratamiento computacional de opiniones, sentimientos y la subjetividad textual (Wiebe, 1994), considera para nuestro caso el análisis de las opiniones referidas a la causa mapuche. Situación que se releva dado que permite entender las demandas del ejercicio del poder, mediante la legitimación y deslegitimación de determinadas posiciones, puesto que dichas posturas influyen en la construcción de opiniones que luego se reproducen, en la misma red social Twiiter, o bien dichas opiniones funcionan como caja de resonancia en otras redes sociales, tales como Facebook, Instagram o Whatsapp, entre otras. Así,el activismo cibernético se caracteriza por un proceso de comunicación horizontal, sin liderazgo e incontrolable (Tascón y Quintana,2012). Igualmente, las producciones de Twitter se reproducen posteriormente en canales de televisión de señal abierta, construyendo la realidad por medio de la alineación o desalineación con dichas posiciones (Martin & White, 2005). La aplicabilidad de los resultados del estudio refieren a la apertura de un nuevo campo poco abordado en nuestro medio, creando distintas posibilidades de construcción de corpus lingüísticos sobre la materia, favoreciendo el análisis de las audiencias para entender las dinámicas de opinión frente a la construcción y/o modificación de leyes sobre la relación entre el Estado y las naciones originarias. Igualmente, analizando las redes de relaciones entre los grupos de influencia, al analizar sus posibles argumentos y la red de argumentos expuestos y considerando su dialogicidad, como también al orientar acciones desde el reconocimiento del Estado hacia las naciones originarias y de qué modo pudieran articularse sus relaciones con este y otros pueblos originarios; o bien posibilitando,

la creación de mejoras en las políticas públicas que dialoguen con la realidad social del pueblo mapuche y de otros pueblos. Sobre investigaciones que utilizan el Análisis de Sentimientos puede observarse en la literatura el estudio de Sidorov, Galicia y Camacho (2016) que analiza un corpus emocional en español basado en tweets. Dicha investigación plantea que no ha surgido el método más apropiado para clasificar tweets en español en el ámbito de Análisis de Sentimientos y que faltarían aún más estudios de este tipo para solucionar el problema. No obstante, se recomienda abordar la metodología Knowledge Discovery in Databases (KDD) o minería de datos (MD). La que se utiliza para la extracción de información en gran cantidad de datos para su conocimiento y comprensión (Rodríguez & García, 2016). Su procedimiento utiliza distintas etapas: selección, exploración, limpieza, transformación, técnicas estadísticas, evaluaciones e interpretación de resultados (Landa, 2016). Este tipo de metodología ha favorecido el análisis de la gran cantidad de datos disponibles en las redes sociales, permitiendo revelar estructuras interconectadas, estableciendo relaciones de términos o frases claves, para descubrir temas, subtemas y entidades semánticas relevantes (Kuz & amp, Falco & amp; Giandini, 2016). Antecedentes de investigaciones sobre la aplicación de este tipo de metodología se observa en redes sociales como Facebook, Youtube y Twitter. En el ámbito de la temática de la salud a través de twitter, se observa lo planteado por Islam (2019) quien descubrió la relación entre yoga y veganismo mediante el modelado de temas. En adelante abordaremos el Marco Teórico inscrito en las ciencias de la computación, ligado a la Minería de Datos y el Análisis de Sentimientos. Luego, presentaremos la Metodología KDD, la que permite seguir el análisis de datos correspondiente. Finalmente, mostraremos los resultados y conclusiones desarrolladas en este trabajo.

2. Objetivos del estudio

2.1. Objetivo general

Visualizar las opiniones de los usuarios de Twitter hacia la comunidad mapuche.

2.2. Objetivos especifivo

- Identificar la connotación que tienen los tweets con respecto a la comunidad mapuche.
- Analizar los tipos de problemas que enfreta la comunidad mapuche.
- caracterizar el rol de los tweets en la divulgación de la causa mapuche.
- Usar técnicas de Text Mining para tratar el problema.

3. Marco Teórico

Dentro de la estadística y la ciencia de la computación existe un campo llamado minería de datos que permite trabajar con grandes volúmenes de conjuntos de datos (información), buscando patrones y tendencias. En el último tiempo, se ha explorado una nueva técnica ligada al análisis de textos denominada minería de textos. Ted Kwartler (2017) lo define como proceso de extracción de información relevante a partir de textos. En este tipo de trabajo, surge el Análisis de Sentimientos (AS) que busca examinar las opiniones manifestadas en textos mediante el análisis computacional (Dubiau y Ale, 2013). Este tipo de análisis de texto busca facilitar la toma de decisiones, examinando opiniones en sus vertientes tanto positivas o negativas. También pretende estudiar el efecto de este tipo de opiniones sobre una determinada temática (Duran, 2016). Esta perspectiva de análisis en redes sociales examina elementos básicos de una red o grafos, que corresponden a una red de nodos o actores. Estos se corresponden en Twitter con los individuos o grupos de personas. Sus vínculos o bordes, son los lazos que existen entre dos o más nodos. Por su parte, los flujos, aportan las direcciones del vínculo, la que puede ser bidireccional o unidireccional (Cordón, 2007). Este procedimiento metodológico puede utilizarse en redes sociales, puesto que estas se manifiestan como una estructura social integrada por un conjunto de usuarios (personas, organizaciones, etc.) los que están relacionados de acuerdo a algún criterio (amistad, parentesco, relación profesional, creencias, etc.). La metodología en cuestión, permite también modelar relaciones o interacciones entre cualquier clase de individuos, tales como personas, grupos u organizaciones. Caracterizando esta estructura en una red en términos de nodos o vértices (actores individuales, personas o cosas dentro de la red) y los bordes o enlaces (relaciones o interacciones) que los conectan. Esta red a menudo se visualiza a través de grafos o matrices debido a la naturaleza cualitativa de estas interacciones. Según Aguirre (2011) los objetivos principales del análisis de redes sociales son: Identificar los actores más influyentes y descubrir los grupos de actores cohesionados, aplicando técnicas de detección de comunidades. Las técnicas más utilizadas en esta metodología gira en torno al análisis de sentimiento.

3.1. Técnicas Exploratoria

3.1.1. Frecuencia de palabras

A la hora de entender que caracteriza un texto o un cojunto de texto, es interesante estudiar qué palabras emplea y con qué frecuencia. Esta es una tarea común en la minería de textos es mirar las frecuencias de palabras.

3.1.2. n-gramas

Sabemos que en el lenguaje se crea por combinaciones de palabras, es decir, determinadas palabras tienden a seguir a otras inmediatamente, o que tiende a coexistir dentro de los mismos textos.(Silge & Robinson,2017)

Las frecuencias de palabras son una herramienta simple para mirar palabras individuales,en lugar de dividir un texto en palabras individuales, podemos dividirlas en grupos,por ejemplo, dos palabras, tres palabras o más. De esta forma, podemos capturar cierta información que no se visualiza con palabras individuales. Este

método de dividir las palabras en grupos se conoce como análisis 'n-gram'. Con dos palabras, se conoce como bigrama, con tres palabra es un trigrama.

3.1.3. Análisis de frecuencia de palabras y documentos

Una pregunta central en la minería de texto y el procesamiento del lenguaje natural es cómo cuantificar los términos claves de un documento. Una medida de cuán importante puede ser una palabra es su frecuencia de término (tf). Sin embargo, hay palabras en un documento que ocurren muchas veces pero que pueden no ser importantes denominados ?palabras vacías ? como ?el', 'es', 'de', etc.

En la fórmula:

$$tf(t,d) = \frac{n_{ij}}{\sum_1 n_{ij}} = \frac{n_{ij}}{|d_i|}$$

donde n_{ij} es el número de veces que aparece el término t_j en el documento d_i.

Otro enfoque es observar la frecuencia de documentos inversa (idf) de un término, que disminuye el peso de las palabras de uso común y aumenta el peso de las palabras que no se usan mucho en una colección de documentos, Se define:

$$idf(t,D) = log(\frac{D}{n_j})$$

donde D es el número total de documentos y n_j es el número de documentos que contienen el término t_j.

Esto se puede combinar con la frecuencia de término para calcular qué tan importante un término en un documento (tf-idf). El índice tf-idf está designado a medir la importancia de una palabra para un documento en una colección (o corpus) de documentos. Por ejemplo, se puede identificar características propias de cada novela en una colección de novelas (Silge and Robinson, 2017). La fórmula matemática para esta medida es:

$$tfidf(t,d,D) = tf(t,d)idf(t,D)$$

Donde t es el término, d es cada documento, D el total de documento y tf-idf es el peso asignado a ese término t en el documento correspondiente.

3.1.4. Comparación del uso de palabras

Para diferenciar el uso de palabras que utiliza cada usuario o grupo, es decir, palabras que utiliza mucho un autor y que no utiliza el otro. Una manera de hacer este análisis es mediante el log of odds ratio de las frecuencias.
Se define :

$$\text{log of odds ratio} = log\left(\frac{(\frac{n_k+1}{N+1})_{Grup,1}}{(\frac{n_k+1}{N+1})_{Grup,2}}\right) \quad (1)$$

Donde n_k el número de veces que aparece el término k en los textos de cada autor o grupos y N el número total de términos de cada autor.

3.1.5. Análisis de correlación (coeficiente phi)

El coeficiente phi corresponde a una medida común para la correlación binaria. El foco del coeficiente phi es cuánto más probable es que aparezcan tanto la palabra X como la Y, o que ninguna aparezca, que una aparece sin la otra (Silge and Robinson, 2017). Consideremos la siguiente tabla:

	Tiene la palabra y	**Sin palabra y**	
Tiene la palabra y	n_{11}	n_{10}	$n_{1.}$
Sin palabara y	n_{01}	n_{00}	$n_{0.}$
	$n_{,1}$	$n_{,0}$	n

El coeficiente phi es definido como:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{,0}n_{,1}}}$$

El coeficiente phi es equivalente a la correlación de Pearson, pero con datos binarios.

3.2. Análisis de Sentimiento

El análisis de sentimiento conocido también como minería de opinión corresponde a una de las funciones de la minería de datos (Kwartler, 2017) y consiste en identificar y extraer el sentimiento de un documento (actitud del autor). Este análisis clasifica la polaridad de un texto, es decir si la opinión expresada es positiva, negativa o neutra. También puede clasificar estados emocionales, como enfado, miedo, sorpresa, etc. El análisis de sentimiento está basado en lexicones que corresponde a diccionario donde las palabras están definidas como positivas o negativas. Se realiza la clasificación mediante el léxico "NRC", que fue creado por el Consejo Nacional de Investigación de Canadá (Martínez, 2017). Se utiliza en una multitud de contextos como el análisis de sentimiento, comportamiento de los consumidores, marketing de productos y campañas políticas. En el análisis de prensa, realizado eminentemente a través de análisis de contenido, incluye frecuencia de palabras, correlaciones y el análisis de sentimientos (Gil, 2010).

3.3. Teoría de Grafos

Un grafo es un conjunto de nodos y un conjunto de líneas entre pares de nodos, esto permite representar adecuadamente la estructura de una red, donde estas líneas pueden ser dirigidas, en el sentido que la conexión es importante, denominándose arcos, o bien líneas no dirigidas, la conexión indica un sentido bidireccional que se llaman aristas. En la literatura un grafo se define como un conjunto de nodos y un conjunto de líneas (aristas) que conectan los nodos. Esto a veces se escribe matemáticamente como una pareja de conjuntos (V, E) donde:

- V es un conjunto distinto de vació.

- E es un conjunto de pares de elementos de V.

- Se denota G (V, E)

3.3.1. Medidas de centralidad

Las medidas de centralidad permiten estudiar qué nodos son los más centrales, los nodos más importantes son los que poseen un mayor poder o bien los más prestigiosos. Las medidas de centralidad son una buena aproximación al análisis de los grafos, y permiten evaluar las dimensiones reales del prestigio y del poder (Velázquez y Aguilar, 2005). Entre las medidas más usadas en minería de datos encontramos:

- **Degree o grados**

  La medida más simple de centralidad se basa en la noción de que un nodo que tiene más vínculos es más prominente que los nodos con pocos o ningún vínculo.

- **Cercanía (closeness)**

  Son nodos que, a pesar de tener pocas conexiones, sus arcos permiten llegar a todos los puntos de la red más rápidamente que desde cualquier otro punto. Representan una excelente posición para monitorear el flujo de información de toda la red. Es decir, establece la distancia media de un nodo con el resto de los nodos de la red. Lo que se expresa en las ecuaciones:

$$c(x) = \frac{1}{\sum_y d(x,y)}$$

- **Intermediación (betweenness)**

  La intermediación es una medida que cuantifica la frecuencia o el número de veces que un nodo actúa como un puente a lo largo del camino más corto entre otros dos nodos. La centralidad intermedia de un nodo viene dada por la expresión:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

  Donde σ_{st} es el número total de caminos más cortos al nodo s al nodo t y $\sigma_{st}(v)$ es el número de esos caminos que pasan v.

3.4. Detección de comunidad

La detección de comunidades tiene como objetivo identificar, utilizando la topología de un grafo, grupos de nodos densamente conectados entre sí y que comparten características comunes o tengan un rol similar dentro de nodos de un grafo. A pesar de que su objetivo parece intuitivo, la detección de comunidades posee un grave

problema: no existe una definición de comunidad universales aceptada, más allá de la noción de que debe haber más aristas entre los vértices de una comunidad que con los vértices de otras comunidades (Fortunate, 2010). Existe una cantidad de algoritmos que identifican comunidades, debido a la inexistencia de una definición consistente de comunidad. En la literatura existen criterios comunes y sus respectivos algoritmos.

3.4.1. Algoritmo de Girvan y Newman

El algoritmo de Girvan y Newman es uno de los métodos más conocidos para detectar comunidades, dado que permite considerar redes dirigidas y ponderadas. Girvan y Newman consideran la idea de asumir que las aristas son un valor alto en betweenness son enlaces entre grupos y posibles enlaces entre comunidades, mientras que las aristas de menor valor son aristas que conectan miembros dentro de un clúster.(Luke,2015) El algoritmo consiste:

- Calcular la intermediación para todas las aristas del grafo.
- Elimina la arista que obtuvo el mayor valor de intermediación (por el que pasan un mayor número de caminos más cortos).
- Re-calcula la medida de intermediación para el resto de las aristas del grafo.
- repetir desde el segundo paso, hasta que no quede ninguna arista.

Modularidad.

La función de calidad más popular se denomina Modularidad y fue propuesta por Newman y Girvan en 2004. es una función que mide la calidad de una partición concreta de una red en comunidades:
Se define como la diferencia entre el número de enlaces existentes en los grupos y el número de enlaces esperado en una red aleatoria equivalente.La siguiente fórmula es la utilizada para calcular la Modularity (Q) de una partición determinada:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(C_i C_J)$$

En donde:

- A_{ij} : Es la matriz de adyacencia del grafo
- k_i : Es el grado del nodo i
- k_j : Es el grado delnodo j
- m : Es número de aristas o enlaces
- $\frac{k_i k_j}{2m}$: Es el número esperado de aristas entre dos nodos i y j
- La función δ vale 1 si los nodos i y j están en la misma comunidad ($C_i = C_j$) y cero en otro caso.

Q ∈ [-1,1]. Cuanto mayor es su valor, mejor es la partición, es decir, las comunidades encontradas están densamente conectadas internamente (hay más enlaces de los que cabría esperar aleatoriamente) y dispersamente conectadas entre sí En una red aleatoria, Q=0.
En la práctica, una modularidad de 0.3 es un buen valor.
Se usa tanto para comparar la calidad de distintas particiones como diseñar métodos de descubrimiento de comunidades que traten de maximizar su valor,

3.5. Análisis del Sentimiento en Twitter

En la última década, el análisis de sentimiento (SA, sentiment analysis), ha despertado un creciente interés. Resulta un gran reto para las tecnologías del lenguaje, ya que obtener buenos resultados es mucho, más difícil de lo que muchos creen. La tarea de clasificar automáticamente en un sistema de sentimientos positivos o negativos, opinión o subjetividad (Pang & lee,2008) un texto escrito en un lenguaje natural es muy compleja por la dificultad de los expertos para acordar asignar a un texto un determinado sentimiento. Además, la interpretación personal de un individuo o conjunto personas se ve afectada por diversos factores, culturales, geográficos y sus experiencias personales. Otros desafíos que plantea la clasificación de textos en sentimientos se relaciona con las dificultades de redacción de los autores de los mensajes y el tamaño del texto, sobre todo para redes sociales como Facebook o Twitter. Los mensajes publicados en Twitter constituyen un material de gran interés para detectar tendencias de opinión entre los usuarios. El hecho de que se hagan públicas opiniones, ideas y debates propicia una cierta asimilación a una conversación informal. En el contexto de la comunicación política, el análisis de contenido y los estudios cuantitativos de los mensajes de Twitter permiten identificar patrones de comportamiento entre los usuarios y puntos de inflexión en las corrientes de opinión (Jungherr, 2015).

4. Metodología

La metodología empleada corresponde a la llamada metodología KDD, esta metodología creada para organizar los diversos componentes y repertorios empleados en el análisis de Lenguaje Natural, particularmente en lo referido a la selección de datos, la exploración de estos, la limpieza y luego la transformación de los mismos, es la que facilita llegar a la aplicación de la técnica de minería de datos, la que a su vez permite utilizar tanto técnicas predictivas, como técnicas descriptivas, para pasar a la evaluación e interpretación de resultados, y finalizando con la etapa de difusión y uso de modelos (Landa , 2016). Antes de pasar a revisar las etapas del método KDD vinculadas a esta investigación, presentaremos algunas características del género de red social en cuestión.

Estructura de Tweet

Twitter es una red social de microblogging en la que los usuarios expresan sus opiniones en texto de máximo 280 caracteres denominados twits (tweets en inglés). Un tuit está normalmente conformado por 3 partes:

- Texto del usuario: Contenido real del texto, este refleja la opinión del usuario.

Figura 1: Estructura de un tweet.

- Referencias: Se cita en el texto a uno más usuarios registrados en Twitter, por ejemplo @Daniel será una referencia al usuario con el nombre "Daniel".
- Etiquetas(hashtags): Son etiquetas que clasifican el mensaje dentro de un grupo de tendencia, eso facilita que la publicación sea vista por el público correcto ya sea pueden filtrar las publicaciones en base a estas etiquetas. Pasaremos ahora a revisar las etapas del método KDD y las iremos aplicando inmediatamente para establecer los respectivos resultados de esta investigación.

5. Resultados

5.1. Selección

La selección de la información se realizó por medio de la aplicación ?Application Programming Interfaces? (API). Con dicha aplicación fue posible recopilar e integrar las fuentes de datos existentes, además de identificar y seleccionar las variables relevantes en los datos, para finalmente aplicar las técnicas de muestreo adecuadas. La aplicación utilizada permitió desarrollar análisis estadísticos de tipo descriptivos y correlaciones de datos, asociados a las opiniones sobre la causa mapuche. Asimismo, esta aplicación permite capturar y luego disponer de la información de los usuarios de Twitter y sus respectivas opiniones. La búsqueda de los Twitt de usuarios en la red Twitter se realizó teniendo como población objetivo a quienes hubiesen empleado el término Mapuche en lengua española. Los tweets seleccionados corresponden a mensajes producidos entre marzo y agosto del año 2019, respectivamente. La población objetivo está referida a los usuarios de twitter, que en sus tweets tiene relacionado el término de búsqueda 'mapuche' y en lenguaje español.

5.2. Exploración

La etapa de exploración de la información se basa en la utilización de técnicas de análisis exploratorio de datos. Para nuestro caso se utilizó la librería rtweet del

Software estadístico R, la cual se conecta con la API de Twitter. De igual manera, es relevante señalar que para obtener la base de datos es necesario registrarse como desarrollador en el sitio web: https://dev.twitter.com/apps y crear un Twitter App, forma que nos permitirá interactuar con el sistema operativo o con otro programa. Luego, el programa aporta una serie de claves y tokens de identificación para realizar la comunicación con la respectiva librería de rtweet, con el fin de desarrollar los primero análisis de correlaciones existentes en la información seleccionada. Esta función Entrega los tweets que coinciden con una consulta de búsqueda proporcionada por el usuario. Solo devuelve los datos de 6-9 días y un máximo de 18.000 tweets. No se consideran los retweets . Los tweets se han considerado desde 03 de marzo del 2019 hasta 12 agosto del año 2019. También se consideró otro parámetro adicional ?include_rts=TRUE ?de la función ?searh-tweets?, este nos entrega los retweets, se han considerado desde 03 de marzo del 2019 hasta 12 agosto del año 2019. En minería de texto el proceso de limpieza consiste en eliminar del texto todo aquello que no aporta información al análisis, por lo que se eliminan patrones no informativos, signos de puntuaciones, espacios innecesarios nombre de usuario o direcciones web. También se eliminan aquellas palabras que no tienen significado propio: artículos, preposiciones, conjunciones, deformaciones del lenguaje.(Kwartler,2017)

5.3. Limpieza

Esta etapa consiste en detectar y tratar la presencia de valores atípicos. Lo segundo, corresponde a imputar la presencia de información faltante o valores perdidos. Finalmente, si se requiere puede ser necesaria la limpieza para quitar datos erróneos e irrelevantes para los fines establecidos. En nuestro caso estamos interesado en eliminar diferente símbolos y caracteres no informativo, que no son relevantes para el análisis. A continuación, se presentan los patrones que queremos eliminar de nuestros tweets.

- Nombres de usuarios (@): En Twitter, cada usuario dispone de un alias precedido del símbolo @, por ejemplo @NombreUsuario. Cuando un usuario escribe un tweet, puede mencionar a otros usuarios con este alias.
- Links a páginas web(http/s): Es común que los usuarios hagan referencias enlaces de páginas web. Estos no aportan información al analizar frecuencias de los términos y el análisis de sentimiento.
- Hashtag (#): Son términos incluidos en los tweets con el objetivo de etiquetar sus mensajes. Al hacer click sobre el hashtag el usuario es redireccionado al conjunto de tweets que contienen la misma etiqueta.
- Números y otros: También se eliminan caracteres numéricos, caracteres de control y espacios innecesarios.
- Mayúscula: Para estandarizar todos los tweets se transforma todas las palabras a minúsculas.

Para limpiar los tweets, se crea una función en R. se presenta los resultados al aplicar la función:

```
 [1] "@soychilecl LAS TANQUETAS SON DE LOS MAPUCHE?."                                                    $
 [2] "Falleció Obispo Sergio Contreras, amigo del pueblo mapuche.\nDon Sergio abrió los recintos de$
 [3] "Precisiones:\nDebe decir: Devuelve una pequeña parte de las tierras que pertenecen a comunida$
 [4] "@jorgeramosnews Estimado Don Jorge, no se trata de criticar por criticar, pero... ¿no le pare$
 [5] "@grdisenos58 @MarthaCajigas @jorgeramosnews Y con el pueblo mapuche. Para la derecha internac$
 [6] "@crisgarciat Lamentable señora. Usted está cegada por su falsa campaña. Su Dios estaría muy d$
 [7] "La movilización mapuche no cesa aunque la policía no quiera retirarse https://t.co/5oSrEd5uqL$
 [8] "Comunidad mapuche manifiesta ocultamiento de información por muerte de lonko Juan de Dios Men$
 [9] "Veo x aqui q le encaman tant vergas al Mapuche o Loco A. Nahuelpan,par Dinenno un jugador fra$
[10] "@elkaiser63 Mapuche en la Bombonera es un tuitazo, pero Q te tiraron en el cuello es Q te tir$
[11] "@elkaiser63 A la larga la calidad línea por línea del equipo le gana a las individualidades. $
[12] "Muy triste x la muerte de Sergio Contreras. Una persona notable. C/lvocación social, mirada c$
[13] "@Fernandazucchel Lo tenemos claro, de que no les gusta que los llamen Nuestro Pueblo Mapuche,$
```

Figura 2: Tuits sin limpiar.

```
 [1] " las tanquetas son de los mapuche "                                                $
 [2] "falleció obispo sergio contreras amigo del pueblo mapuche don sergio abrió l$
 [3] "precisiones debe decir devuelve una pequeña parte de las tierras que pertene$
 [4] " estimado don jorge no se trata de criticar por criticar pero no le parece q$
 [5] " y con el pueblo mapuche para la derecha internacional violentar los derecho$
 [6] " lamentable señora usted está cegada por su falsa campaña su dios estaría mu$
 [7] "la movilización mapuche no cesa aunque la policía no quiera retirarse "     $
 [8] "comunidad mapuche manifiesta ocultamiento de información por muerte de lonko$
 [9] "veo x aqui q le encaman tant vergas al mapuche o loco a nahuelpan par dinenn$
[10] " mapuche en la bombonera es un tuitazo pero q te tiraron en el cuello es q t$
[11] " a la larga la calidad línea por línea del equipo le gana a las individualid$
[12] "muy triste x la muerte de sergio contreras una persona notable c social mira$
[13] " lo tenemos claro de que no les gusta que los llamen nuestro pueblo mapuche $
```

Figura 3: Tuits aplicando la función de limpieza.

5.4. Transformación

Esta etapa apunta a utilizar técnicas de reducción o aumento de la dimensión de los datos para favorecer su análisis. Después de limpiar los tweets para eliminar la variabilidad queremos determinar las palabras más utilizadas en los tweets y después obtener algunas relaciones entre los términos. Primero tenemos que separar el texto en pequeñas partes denominadas token, ya sea palabras individuales, palabras compuestas, hashtags, sentimiento, etc. Esto permite procesar cada elemento, para luego eliminar los stopwords y determinar la frecuencia de cada token. Para dividir el texto en tokens individuales y transformarlo en una estructura de datos ordenado. Para ello, utilizamos la unnest_tokens() función del paquete tidytext .

5.5. Análisis exploratorio

Después del limpiar los tuits para eliminar la variabilidad queremos determinar las palabras mas utilizadas en los tuits y después obtener algunas relaciones entres los términos .

Primero tenemos que separar el texto en pequeñas partes denominadas token, ya sea palabras individuales, palabras compuestas, hashtags, sentimiento, etc. Esto permite procesar cada elemento, para luego eliminar los stopwords y determinar la frecuencia de cada token.

Para dividir el texto en tokens individuales y transformarlo en una estructura de datos ordenado. Para ello, utilizamos la unnest_tokens() función del paquete tidytext .

Esta función tiene dos argumentos básicos , el primero el nombre de la variable de salida y el segundo , el texto original que queremos dejarlo en token(palabras individuales).

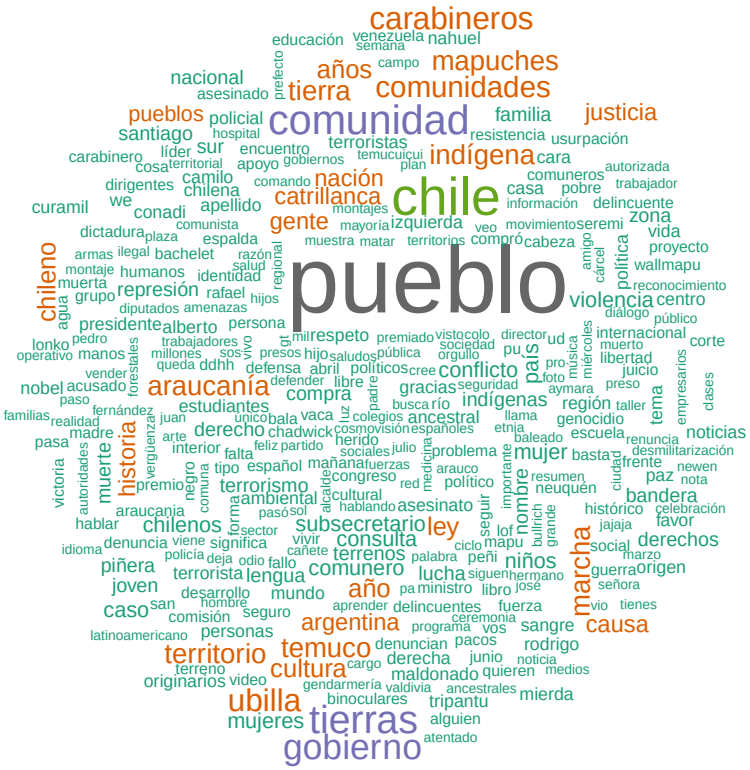


Figura 4: Nube de Palabras de los términos con mayor frecuencia utilizado por los usuarios.

La nube de palabras muestra de forma cualitativa los términos con mayor frecuencia usados en los tweets por los usuarios. La frecuencia del término está dada por el tamaño de la letra. Observamos que los términos pueblo,comunidad, historia tierra, gobierno , etc . Son lo que presentan una mayor frecuencia.

5.6. Análisis de correlación (coeficiente de phi)

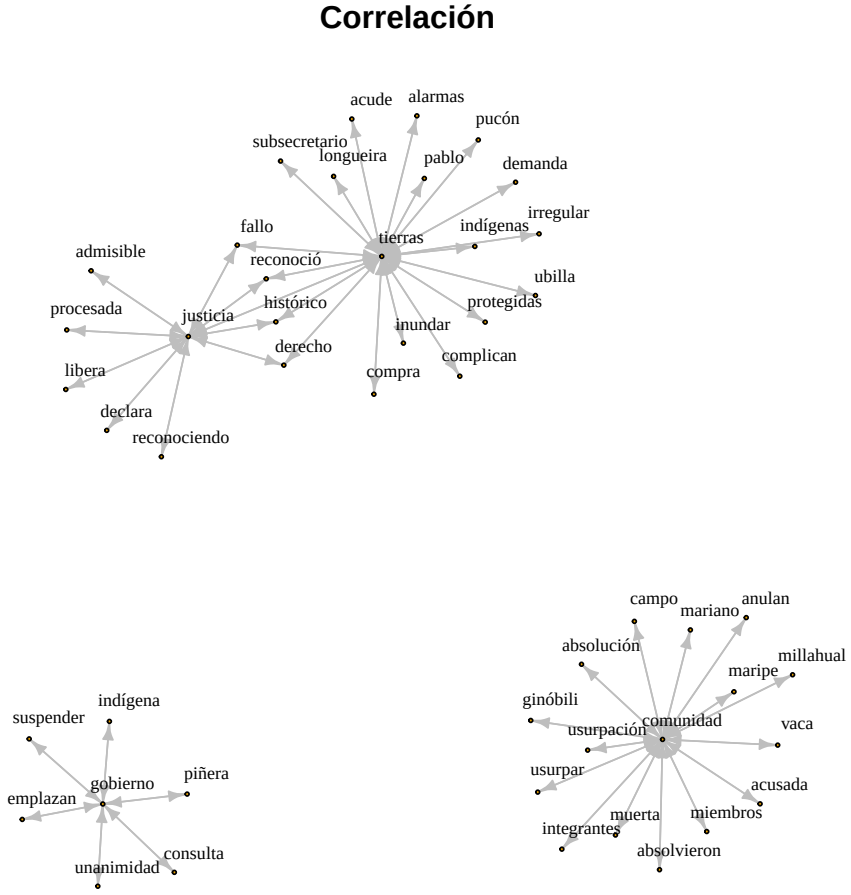


Figura 5: Correlación entre algunos términos más frecuente.

5.7. Relaciones entre palabras

Hasta el momento se han considerado a las palabras como unidades individuales, independiente. Sabemos que en el lenguaje se crea por combinaciones de palabras, es decir, determinadas palabras tienden a seguir a otras inmediatamente , o que tiende a coexistir dentro de los mismos textos.
La función 'unnest_tokens()' del paquete tidytext permite separar el texto por n-gramas, siendo cada n-grama una secuencia de n palabras consecutivas.

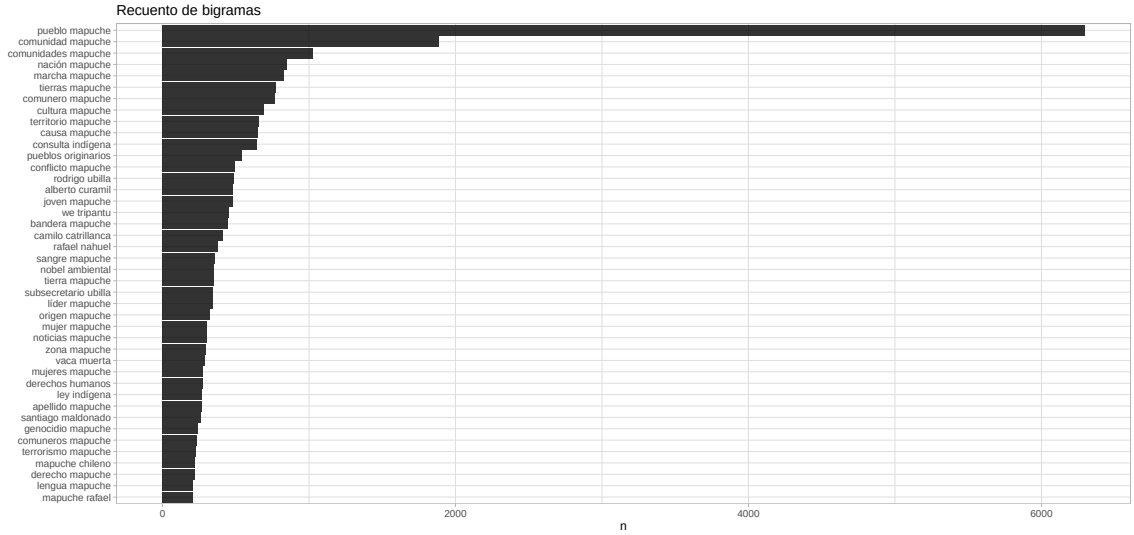
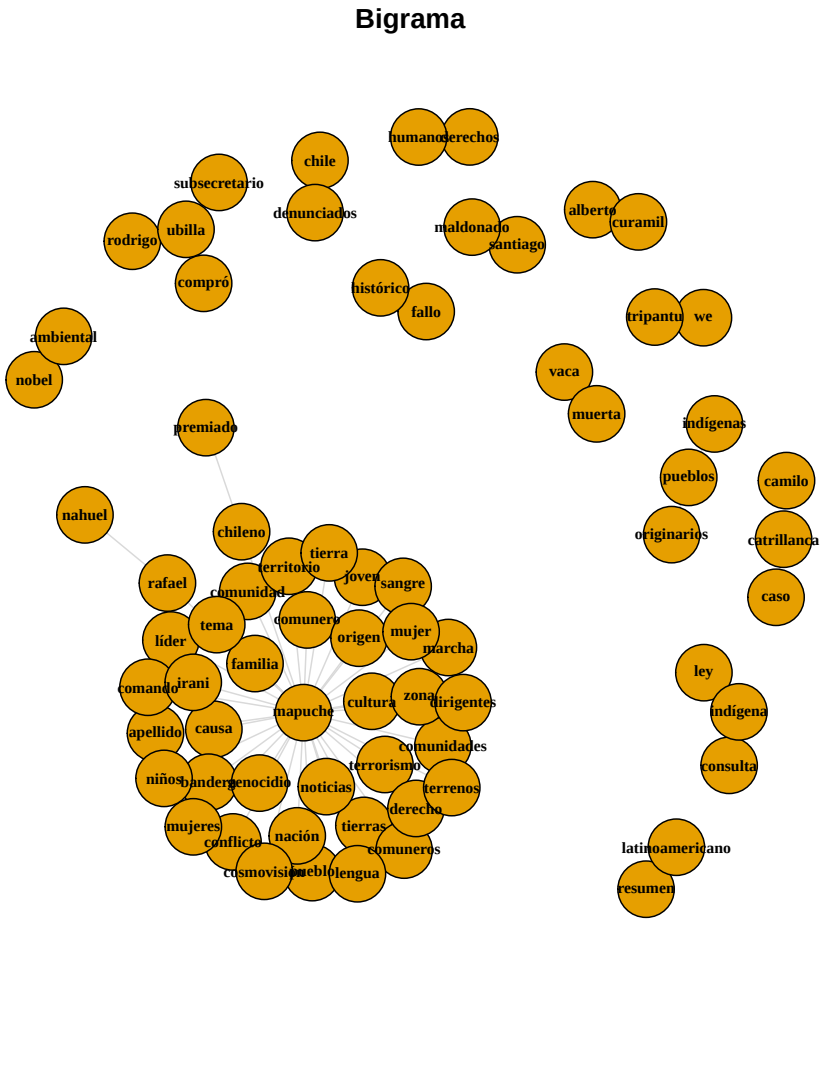


Figura 6: Recuento de los principales bigramas con una frecuencia mínima de 200.

Se puede apreciar en los primeros bigramas, sobre el pueblo o comunidades mapuche sobre conflicto territorial.



Recuento: 150

Figura 7: Grafo de no dirigido entre los términos con mayor frecuencia.

El gráfico muestra un grafo no dirigido , con una frecuencia mayor a 150. Podemos ver con más claridad algunas relaciones de palabras importantes. Por ejemplo, se encuentran nombres mapuches como el 'caso de Camilo Catrillanca' y Alberto Curamil que gano el premio ambiental Goldman , también se observa el nombre de 'Santiago Maldonado' que apoyó el reclamo de los pueblos originarios por sus tierras ancestrales.Por otro lado, esta el nombre del subsecretario Rodrigo Ubilla que compro tierras ancestrales en Temuco. Estos son acontecimiento recientes. También se puede observar acerca del pueblo mapuche que es un pueblo originario indígena.

A continuación se extrae el mayor componente Se extrae el mayor componente conectado al grafo para obtener una mejor visualización.

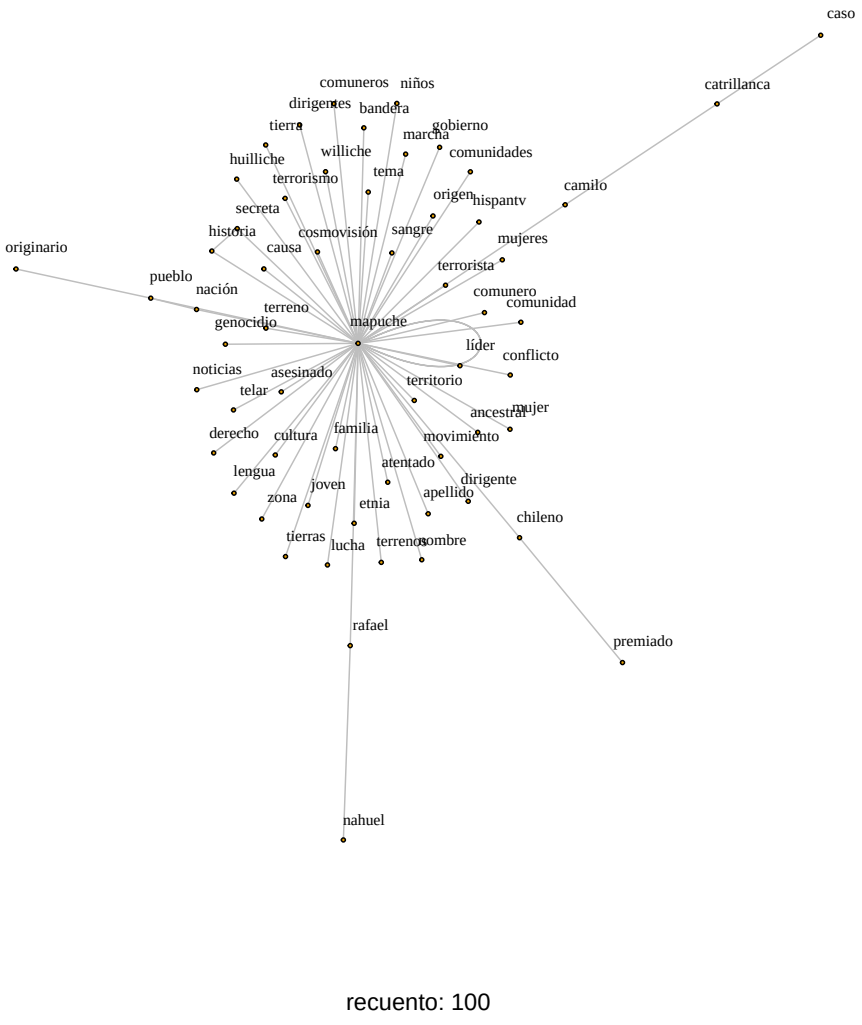


Figura 8: Subgrafo de la componente con mayor grado.
El subgrafo muestra la frecuencia mayor a 100 con relación al término de búsqueda 'mapuche'. Se observar la relaciones sobre el pueblo mapuche originario , sobre el conflicto sobre las tierras ancestrales , sobre causas históricas del territorio. Por otra parte, observan relaciones del mapuche Camilo Catrillanca , Alberto Curamil y Rafel Nahuel.

5.8. Detección de grupos de palabras

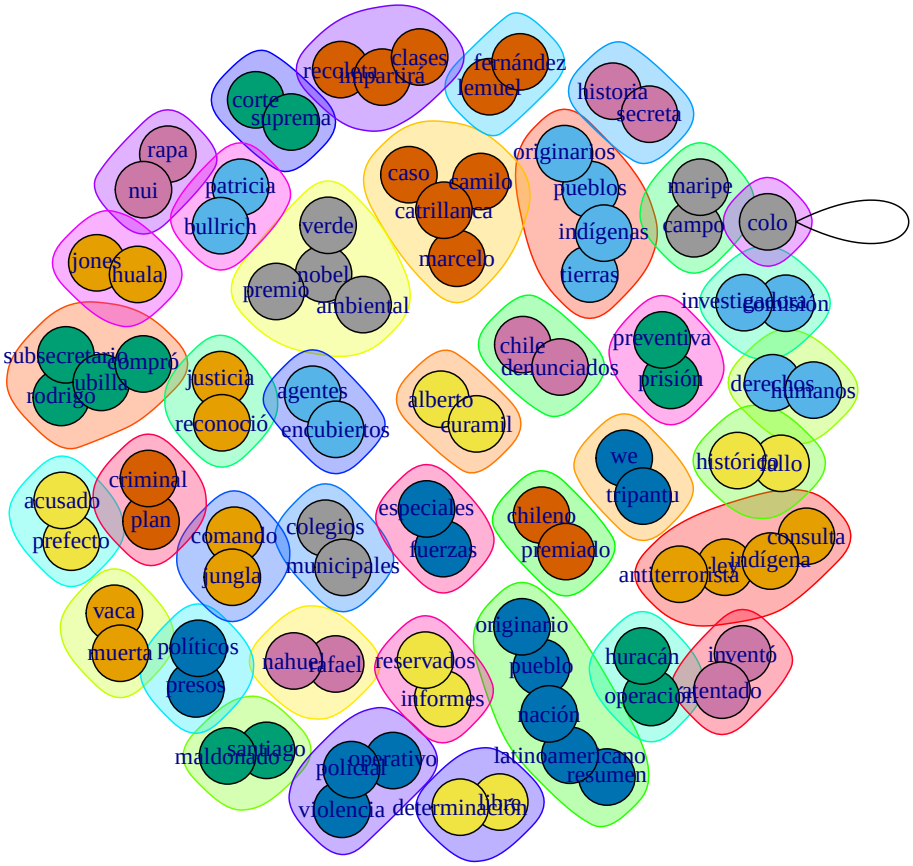


Figura 9: Subgrafo con grupos de palabras.
En el subgrafo se puede apreciar los grupos de palabras mas frecuentes.

5.9. Detección de comunidades

¿Tiene alguna estructura el grafo de RTs(retweets)? Al hacer un RT de un usuario en un contexto politico significa en cierta forma que estamos compartiendo la opinión de esa persona. Se espera que tengamos muchos RTs dentro de una comunidad con la misma opinión sobre la conversación y pocos hacia afuera.
En nuestro caso, queremos ver si existen grupos de usuarios que opinan a favor o en contra sobre el conflicto mapuche.

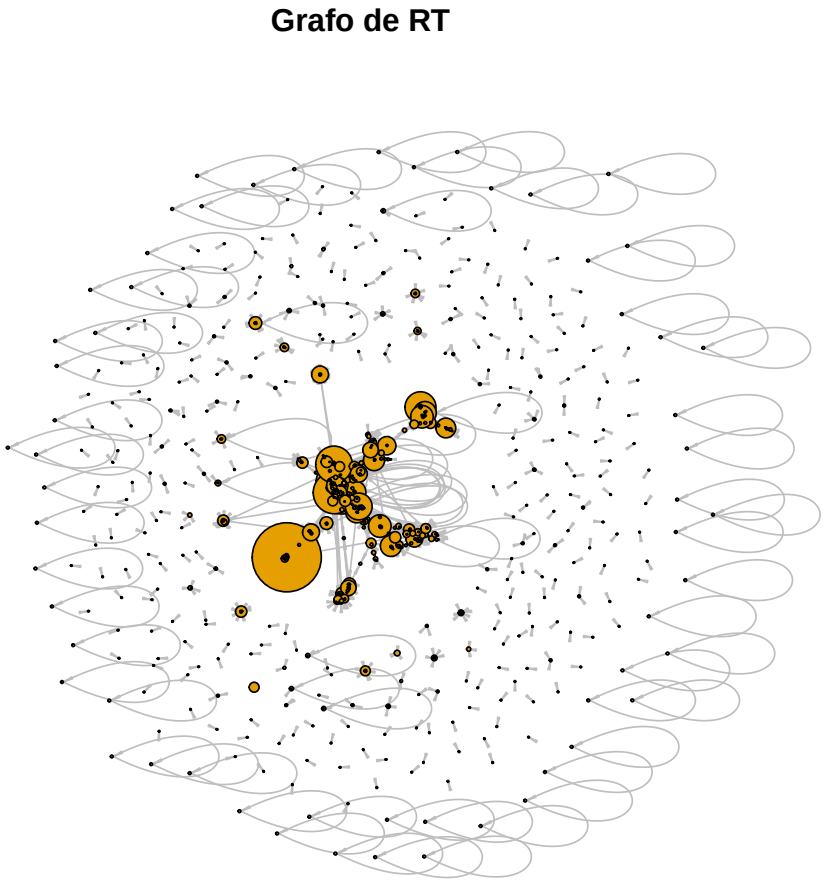


Figura 10: Grafo de retweets .
Grafo de retweets de la comunidades en estudio.

5.9.1. Detección de comunidades basada en el intervalo de borde (Newman-Girvan) .

Se aplica el algoritmo de Newman-Girvan implementado en el paquete igraph de r para determinar si existen grupos. Existen una medida de bondad de ajuste para ver si los grupos de nodos densamente conectados entre si, conocida como modularidad .

La modularidad es una estadística corregida por azar y se define como la fracción de vínculos que se encuentran dentro de los grupos dados menos la fracción esperada si los vínculos se distribuyen al azar. .(Luke,2015).
Se observa mediante el algoritmo existen 377 grupos con una modularidad de un 0.87, sabemos mayor es su valor, mejor es la partición, es decir, las comunidades encontradas están densamente conectadas internamente

A continuación, para Identificar la connotación que tienen los tweets con respecto a la comunidad mapuche y, por consiguiente, caracterizar. Se identifican aquellos grupos(comunidad) con más de 30 usuarios para poder representar gráficamente.

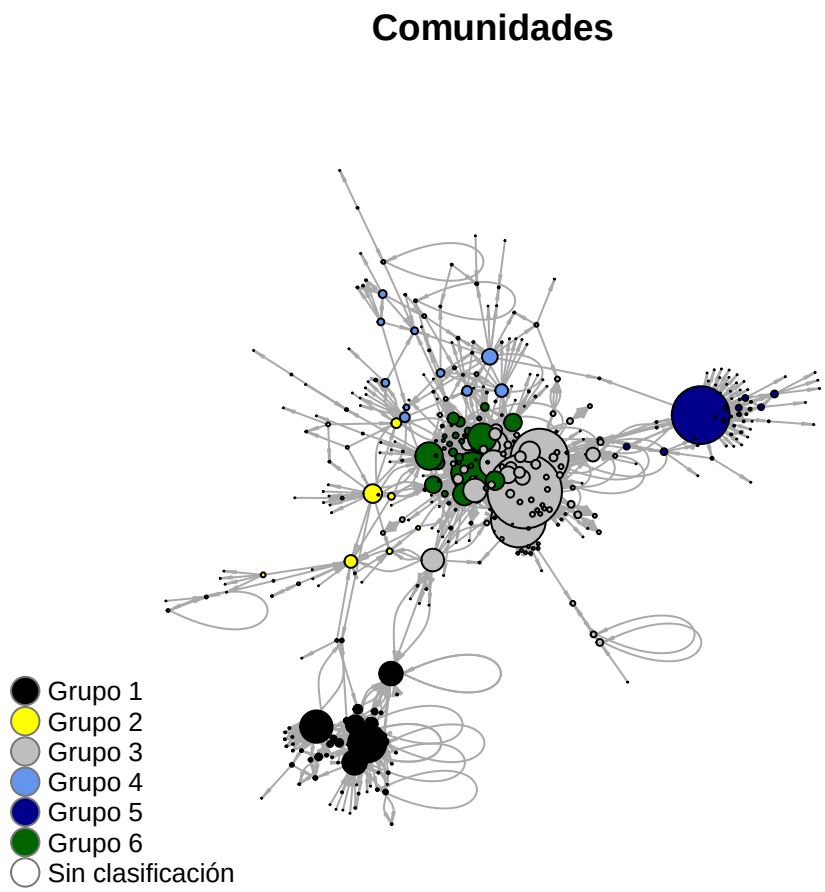


Figura 11: Sub Grafo de Retweets.
Muestra el sub grafo de retweets de la comunidades en estudio.

A continuación, se aplica un análisis exploratorio para obtener algunas ideas sobre la percepción de cada grupo sobre el conflicto mapuche.

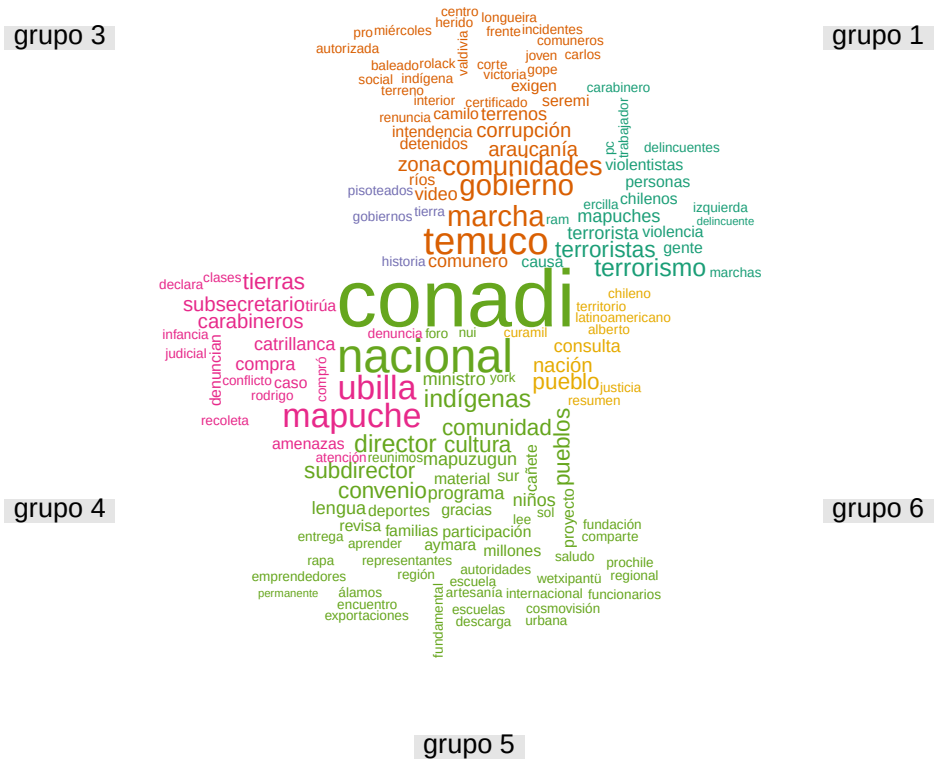


Figura 12: Términos frecuentes por grupos .

La nube de palabra muestra los términos con mayor frecuencia utilizada por los usuarios de cada grupo. La frecuencia del término esta dada por el tamaña de la letra.

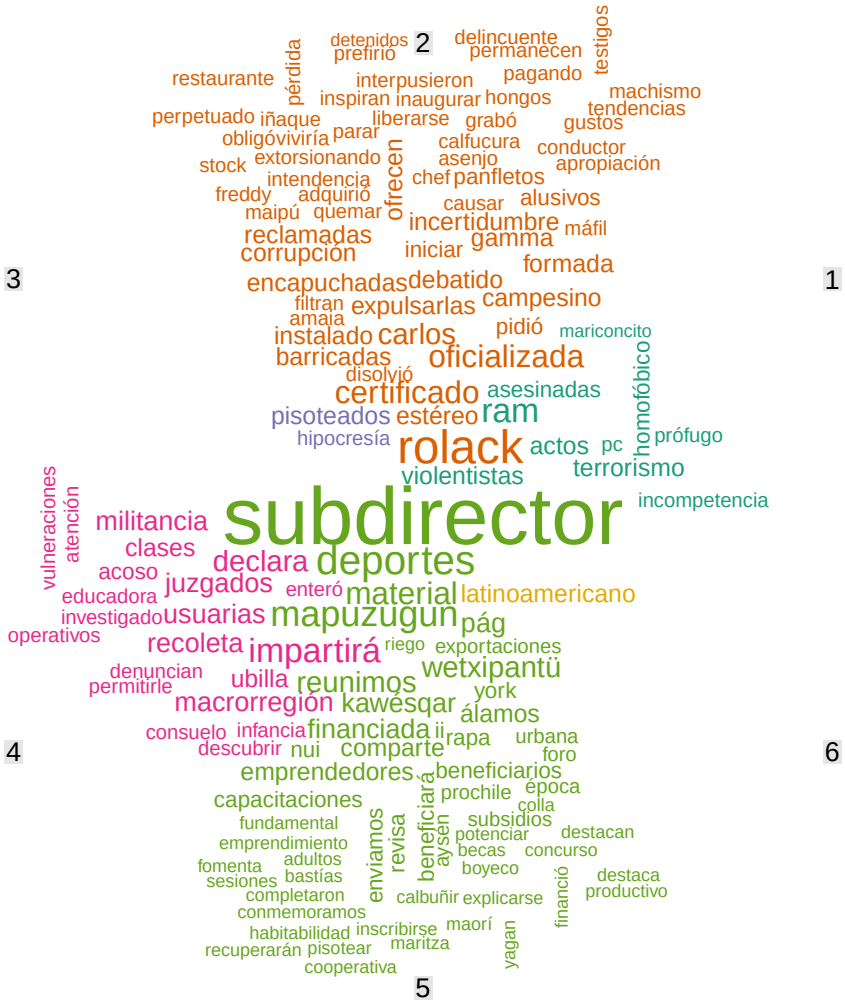


Figura 13: TF-IDF por grupos.
se identifica las palabras que son importantes para cada grupo mediante la medida tf-idf.

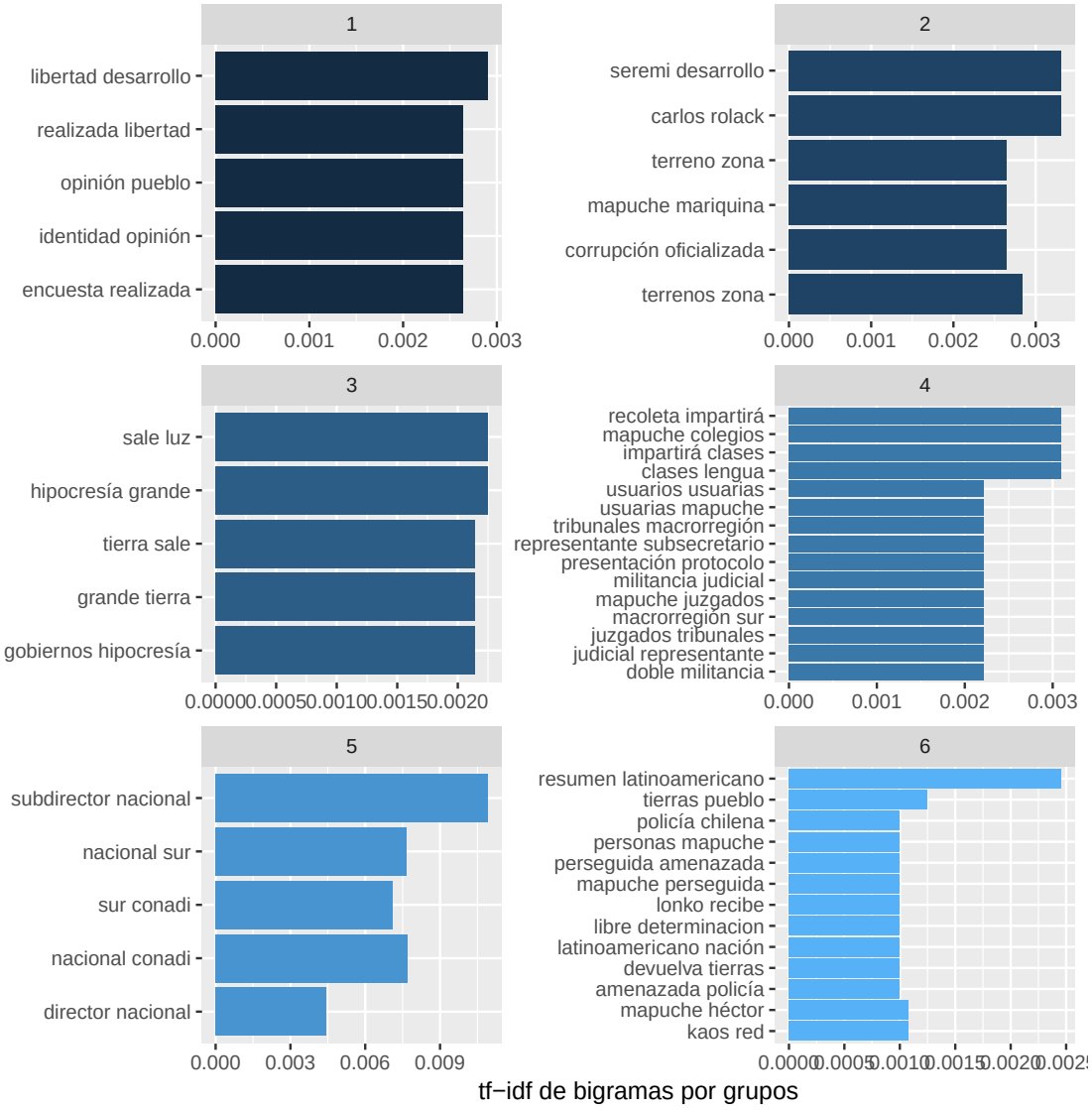


Figura 14: tf-idf de bigramas por grupos .
Muestra los bigramas que son más importantes por cada grupo.

Figura 15: Frecuencia de bigramas por grupos .

La nube de palabra muestra los bigramas con mayor frecuencia utilizada por los usuarios de cada grupo. La frecuencia del término esta dada por el tamaña de la letra.

Mediante los gráficos 12, 13, 14 y 15 se observa que el grupo uno contiene términos negativos. Por ejemplo, en el gráfico 12, se observan términos más frecuente como terrorismo, violencia ,delincuente,lo que alude que ese grupo no apoya el conflicto mapuche. El grupo 2 y 3 contiene términos más frecuente como tierra , historia , gobierno , lo que se puede intuir sobre una visión favorable y negativa sobre el gobierno. El cuarto grupo habla sobre temas actuales sobre el conflicto mapuche . El quito grupo habla sobre la CONADI(Corporación Nacional de Desarrollo Indígena) una corporación del gobierno y el sexto grupo contiene termino que son positivo acerca el conflicto mapuche y actuales. Para verificar lo anterior , se aplica análisis de sentimiento mediante el paquete syuzhet.

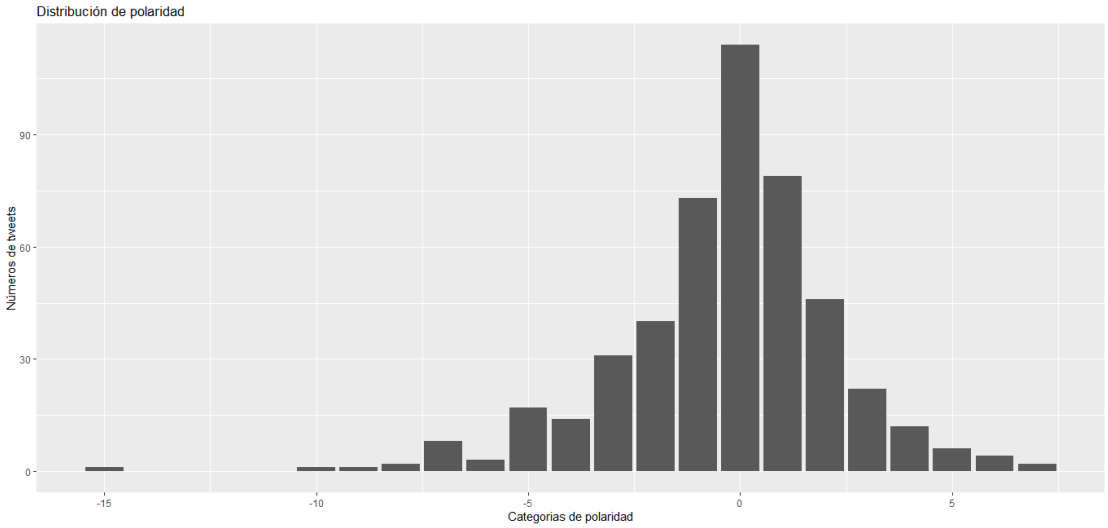


Figura 16: Distribución de polaridad grupo 1.

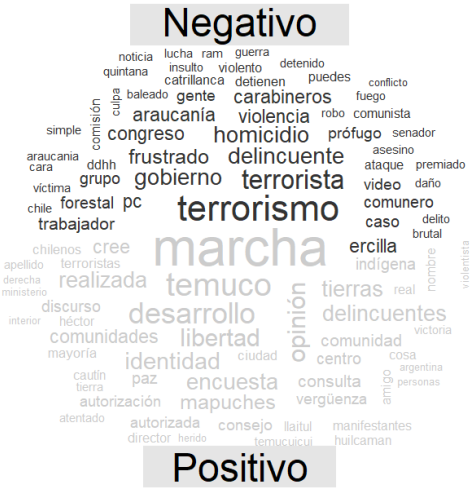


Figura 17: Palabras positivas y negativas más comunes del grupo 1.

El tamaño del texto de una palabra en la figura 17 es proporcional a su frecuencia dentro de su sentimiento. Podemos usar esta visualización para ver las palabras positivas y negativas más importantes.

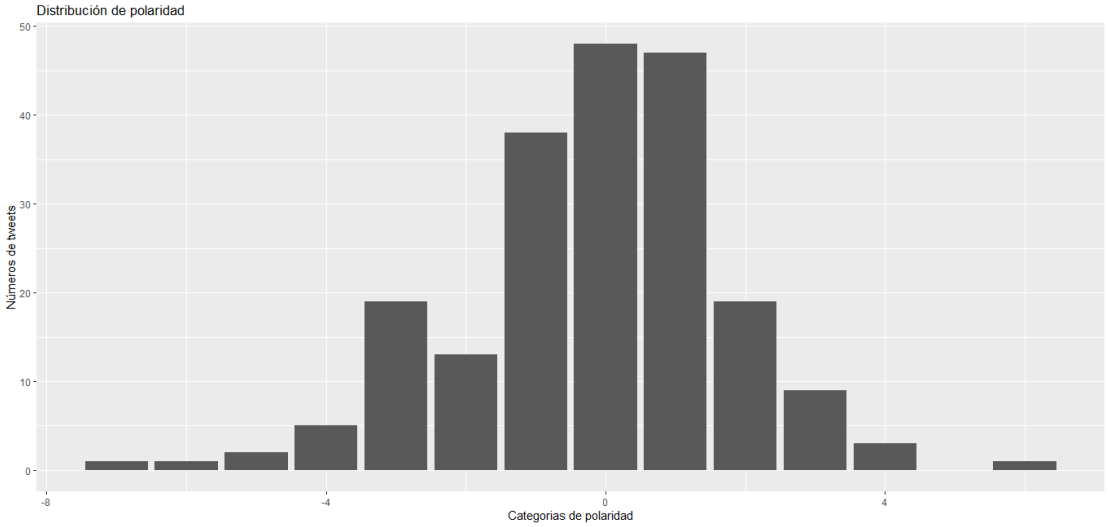


Figura 18: Distribución de polaridad grupo 2.

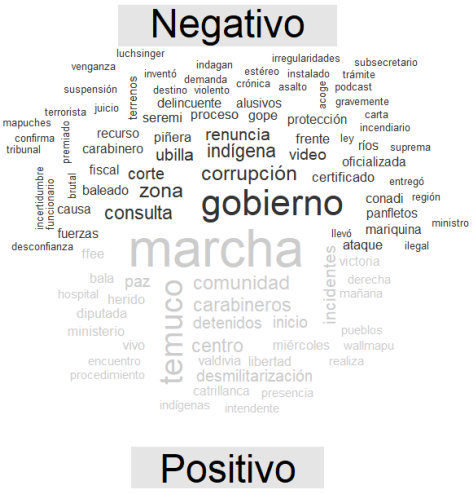


Figura 19: Palabras positivas y negativas más comunes del grupo 2.

El tamaño del texto de una palabra en la figura 19 es proporcional a su frecuencia dentro de su sentimiento. Podemos usar esta visualización para ver las palabras positivas y negativas más importantes.

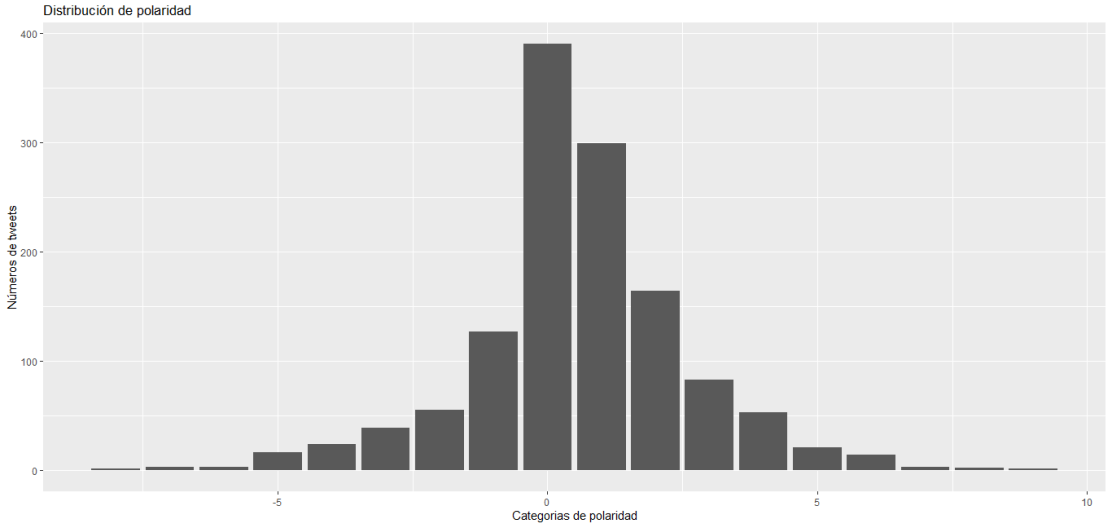


Figura 20: Distribución de polaridad grupo 3.

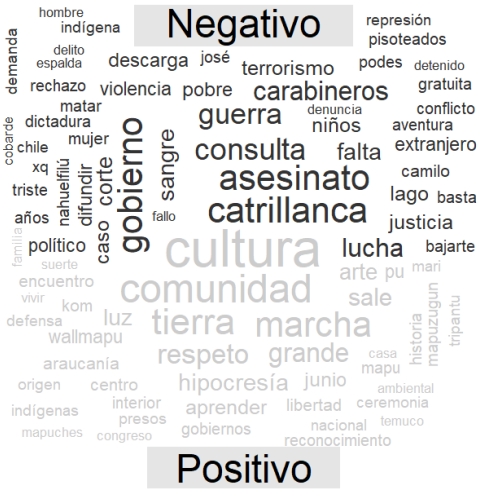


Figura 21: Palabras positivas y negativas más comunes del grupo 3.

El tamaño del texto de una palabra en la figura 21 es proporcional a su frecuencia dentro de su sentimiento. Podemos usar esta visualización para ver las palabras positivas y negativas más importantes.

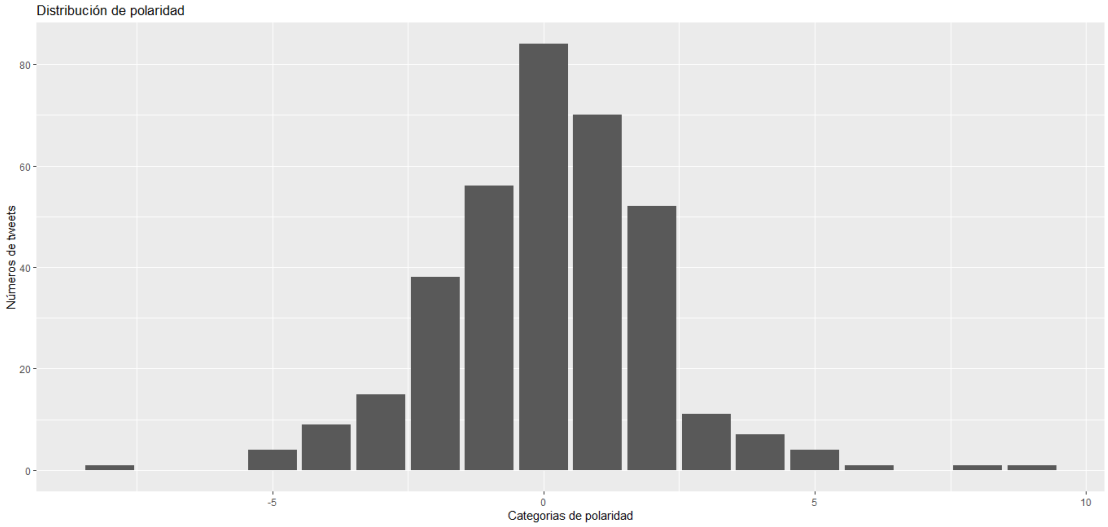


Figura 22: Distribución de polaridad grupo 4.

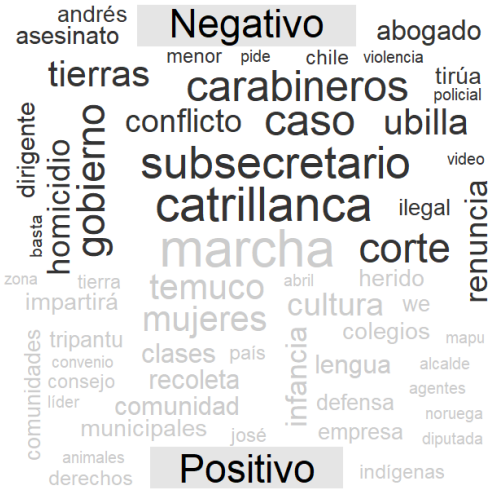


Figura 23: Palabras positivas y negativas más comunes del grupo 4.

El tamaño del texto de una palabra en la figura 23 es proporcional a su frecuencia dentro de su sentimiento. Podemos usar esta visualización para ver las palabras positivas y negativas más importantes.

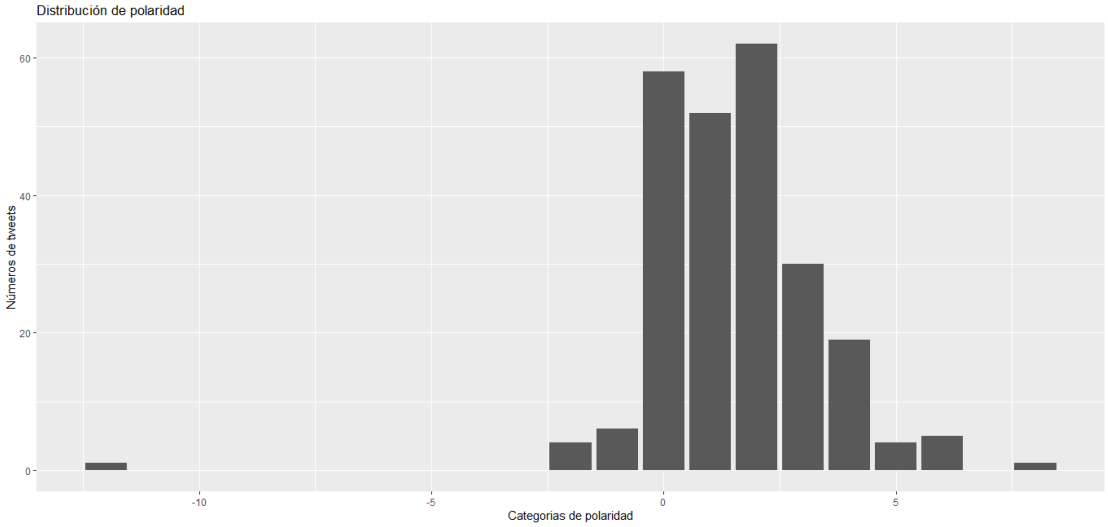


Figura 24: Distribución de polaridad grupo 5.

Figura 25: Palabras positivas y negativas más comunes del grupo 5.

El tamaño del texto de una palabra en la figura 25 es proporcional a su frecuencia dentro de su sentimiento. Podemos usar esta visualización para ver las palabras positivas y negativas más importantes.

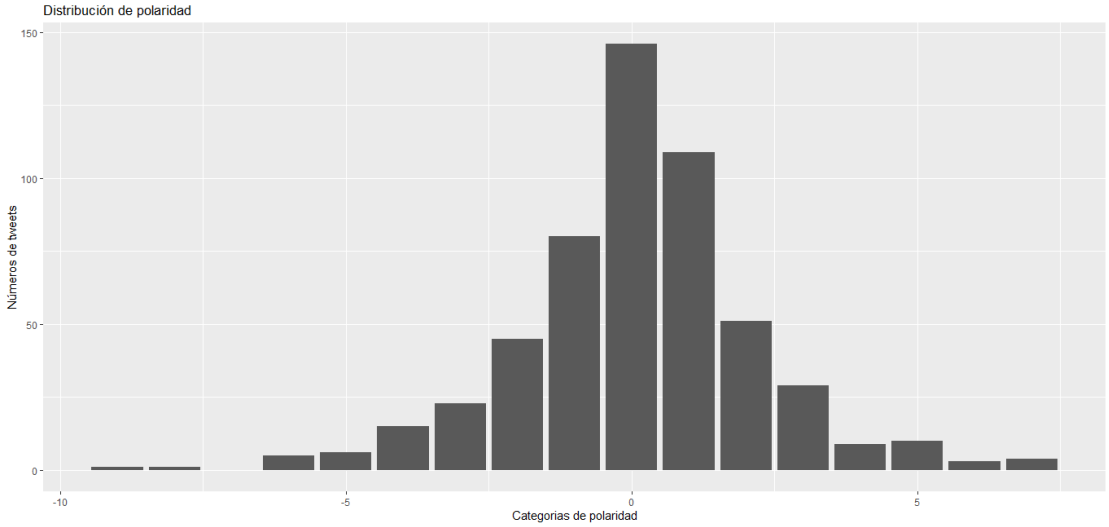


Figura 26: Distribución de polaridad grupo 6.

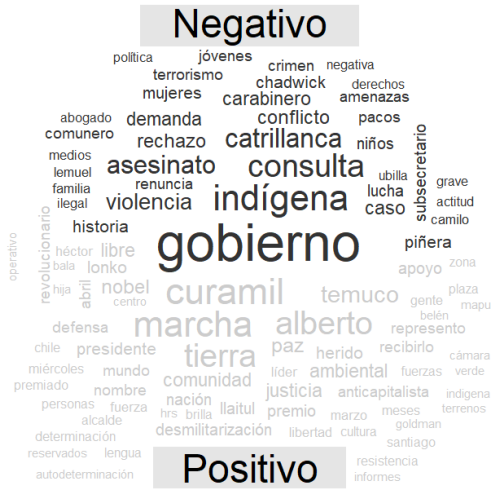


Figura 27: Palabras positivas y negativas más comunes del grupo 6.

El tamaño del texto de una palabra en la figura 27 es proporcional a su frecuencia dentro de su sentimiento. Podemos usar esta visualización para ver las palabras positivas y negativas más importantes.

5.10. Diferencia entre grupos mediante correlación y log of odds ratio de las frecuencias

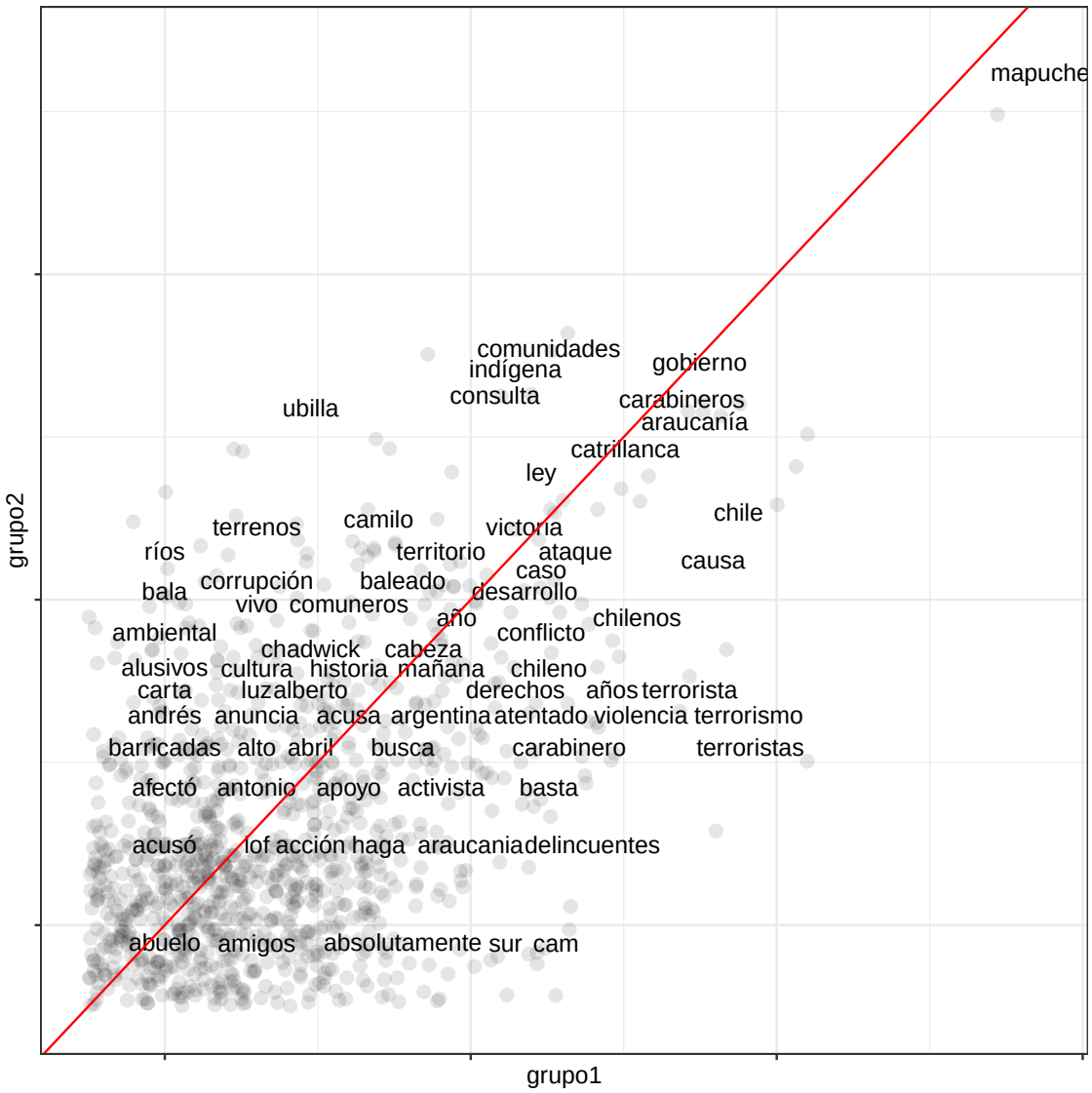


Figura 28: Comparación de las frecuencias de palabras entre grupo 1 y 2 mediante correlación.

Las palabras que están cerca de la línea en estas gráficas tienen frecuencias similares en ambos conjuntos de textos. Las palabras que están lejos de la línea son palabras que se encuentran más en un conjunto de textos que en otro.

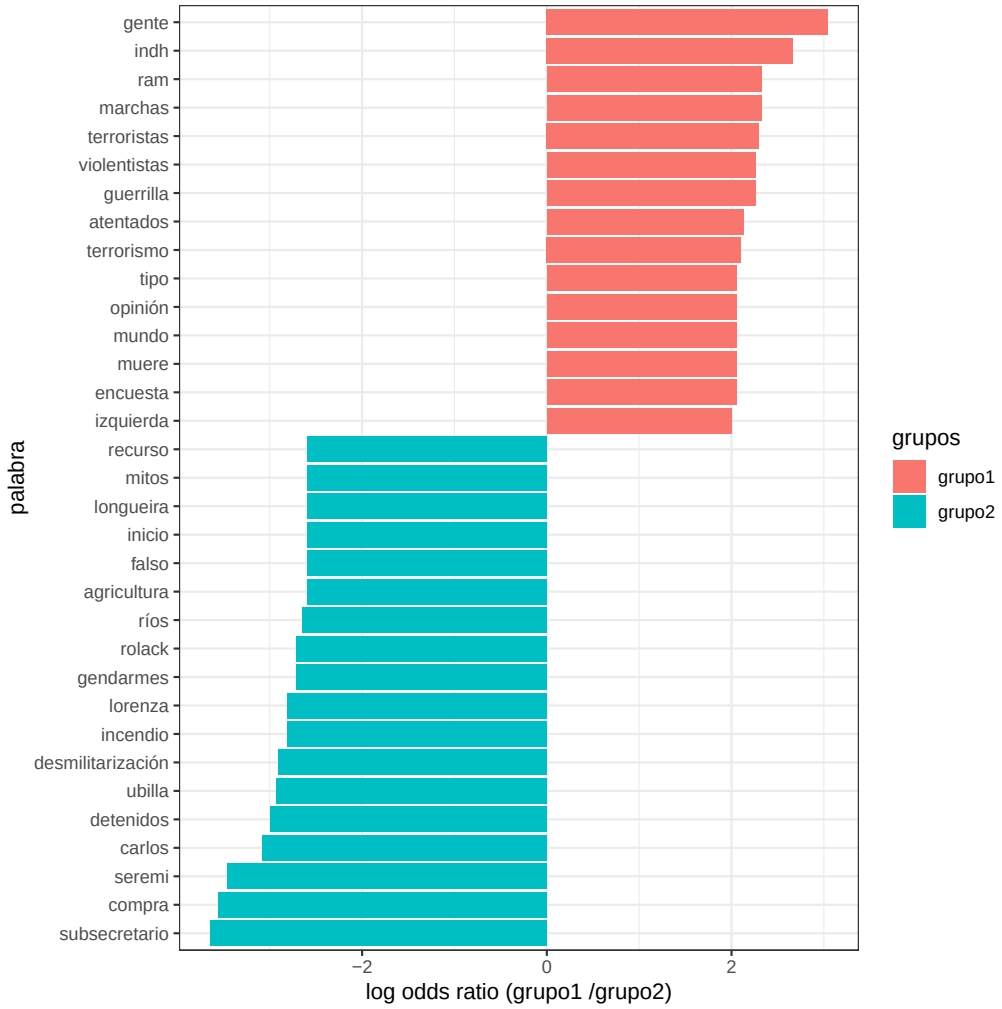


Figura 29: Comparación de las frecuencias de palabras entre grupo 1 y 2 mediante log of odds ratio de las frecuencias .

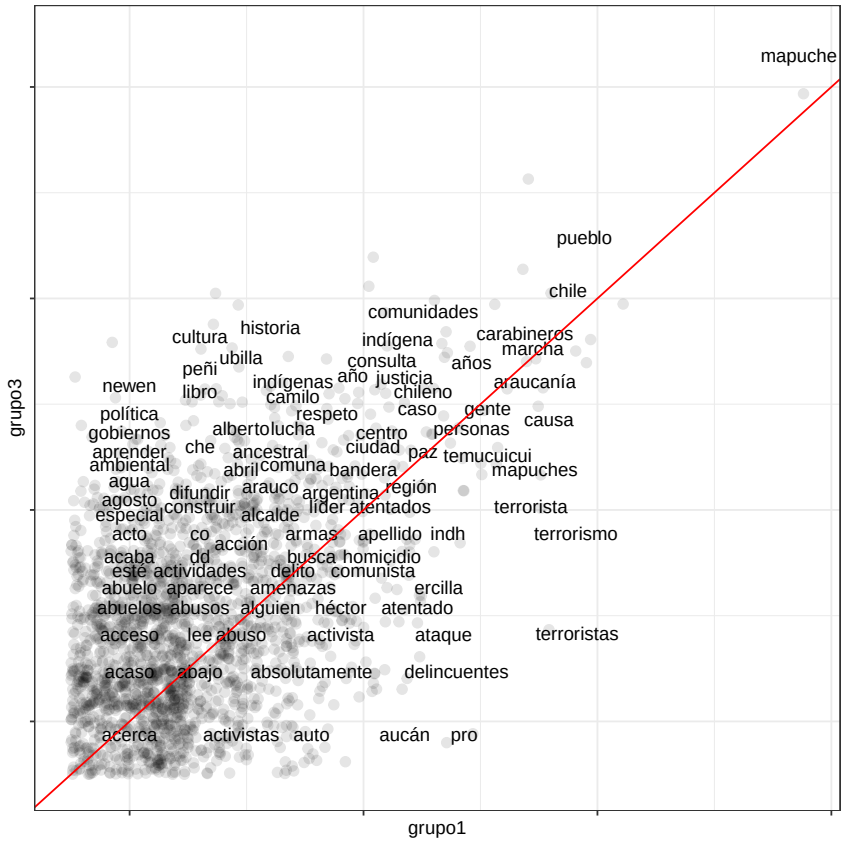


Figura 30: Comparación de las frecuencias de palabras entre grupo 1 y 3 mediante correlación.

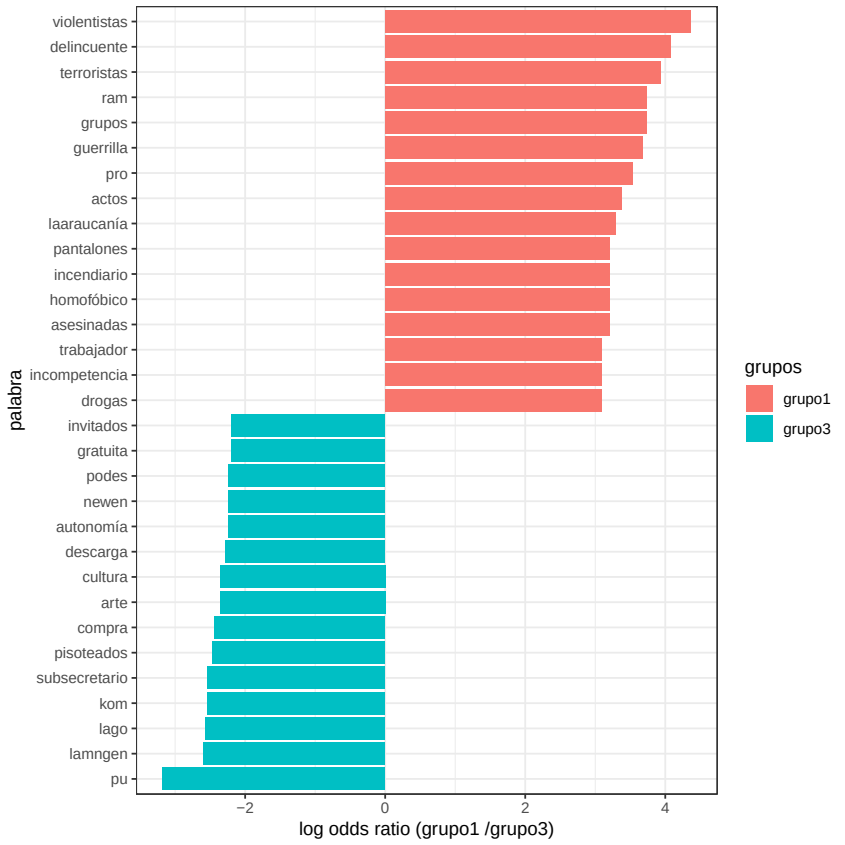


Figura 31: Comparación de las frecuencias de palabras entre grupo 1 y 3 mediante log of odds ratio de las frecuencias .

Figura 32: Comparación de las frecuencias de palabras entre grupo 1 y 4 mediante correlación.

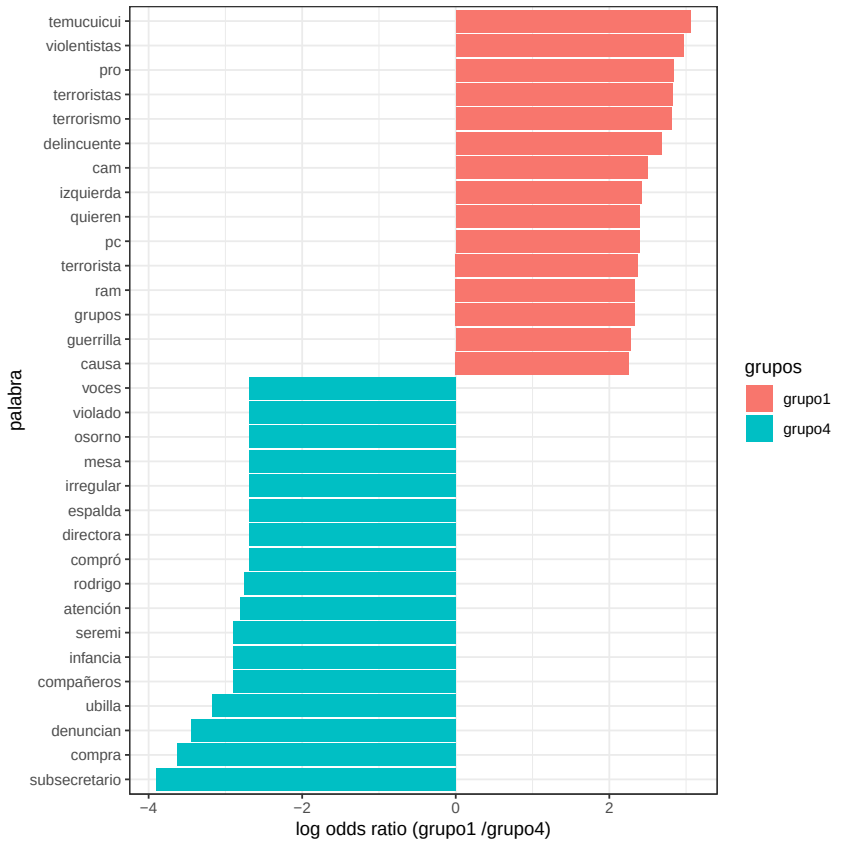


Figura 33: Comparación de las frecuencias de palabras entre grupo 1 y 4 mediante log of odds ratio de las frecuencias .

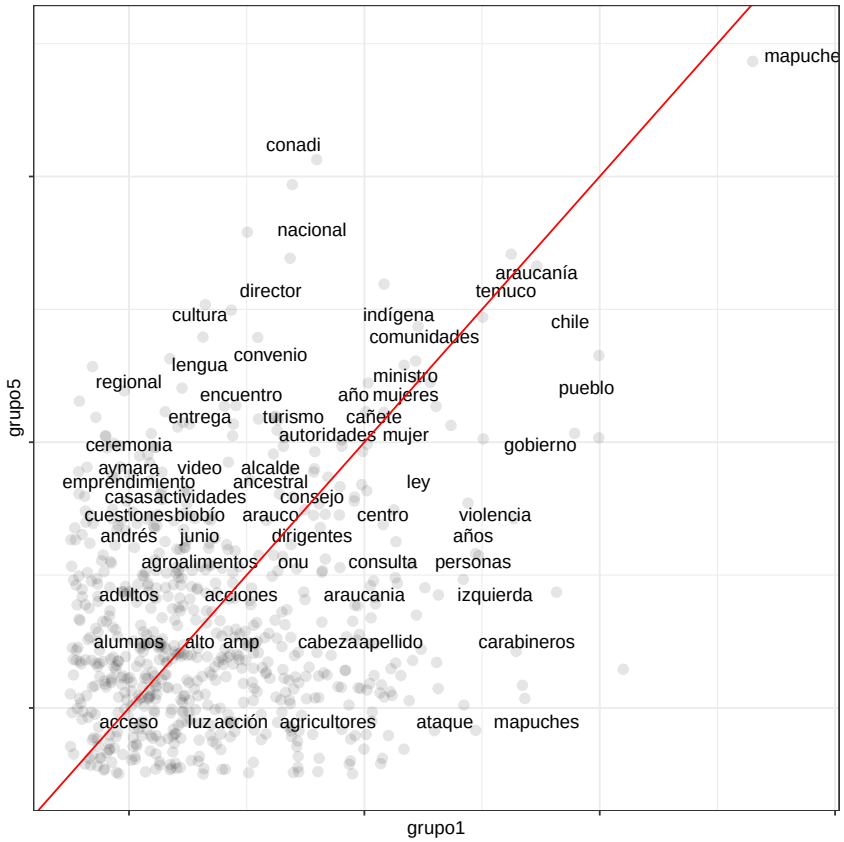


Figura 34: Comparación de las frecuencias de palabras entre grupo 1 y 5 mediante correlación.

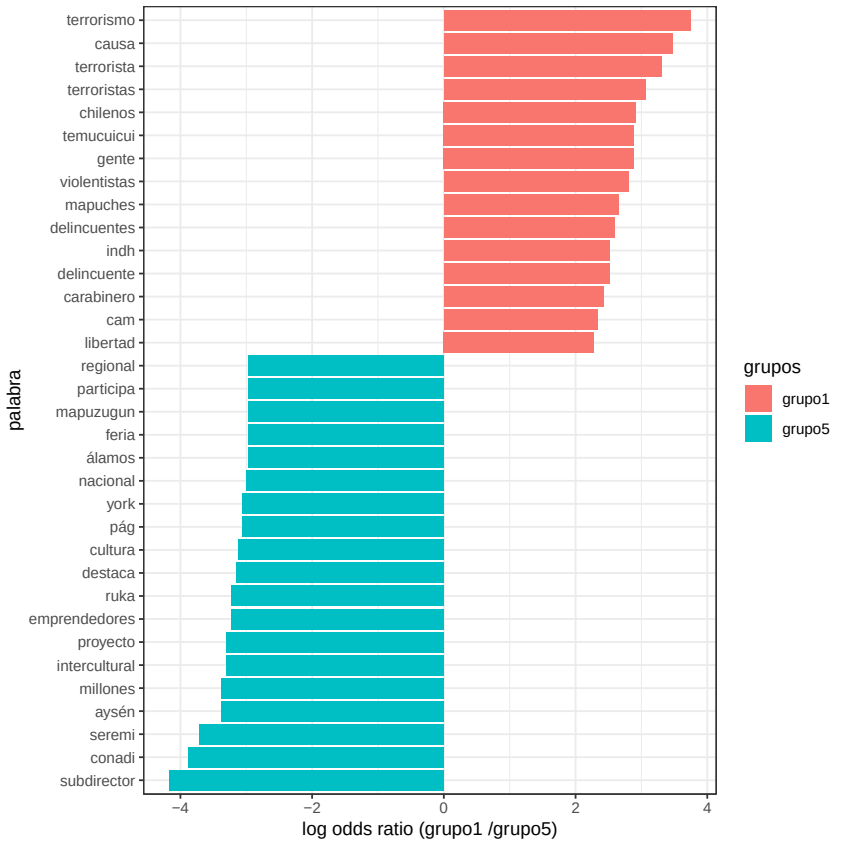


Figura 35: Comparación de las frecuencias de palabras entre grupo 1 y 5 mediante log of odds ratio de las frecuencias .

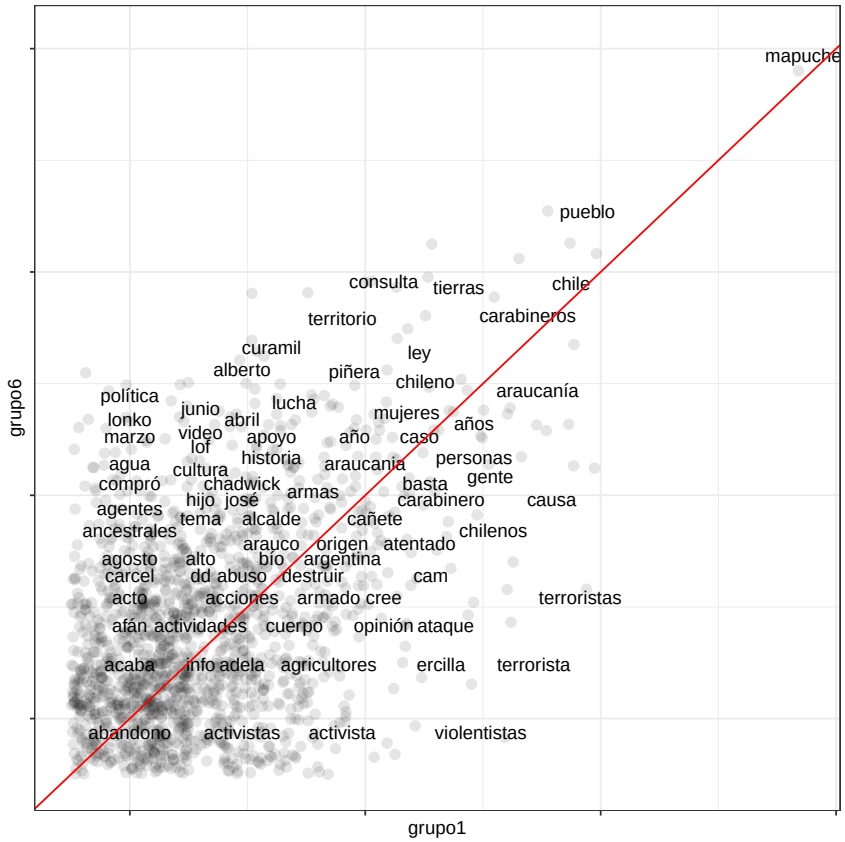


Figura 36: Comparación de las frecuencias de palabras entre grupo 1 y 6 mediante correlación.

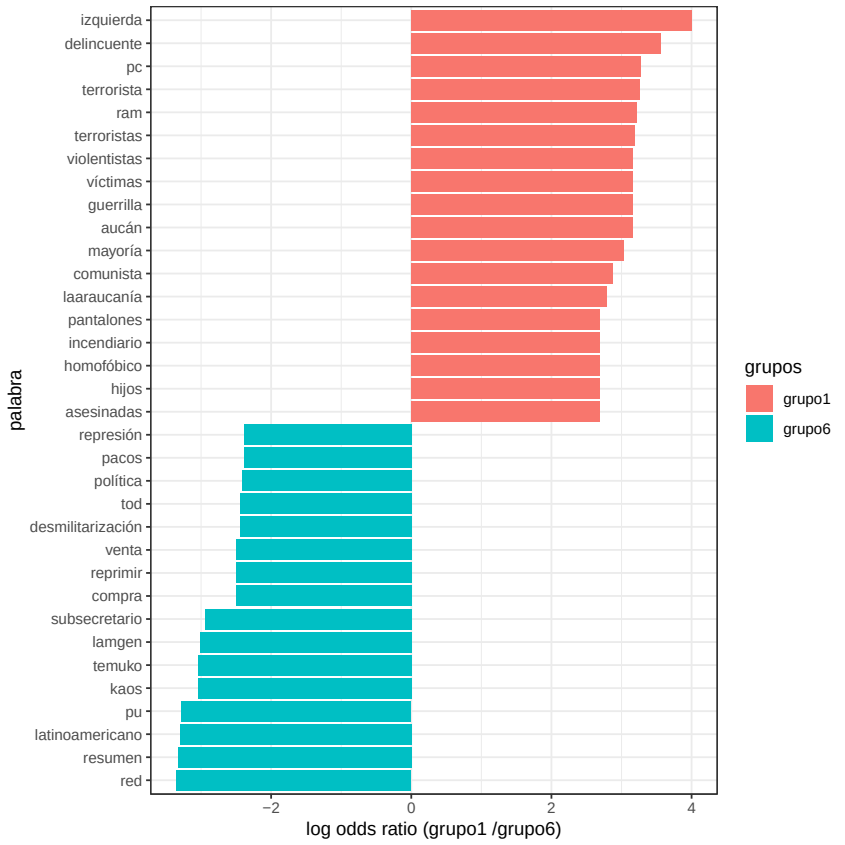


Figura 37: Comparación de las frecuencias de palabras entre grupo 1 y 6 mediante log of odds ratio de las frecuencias .

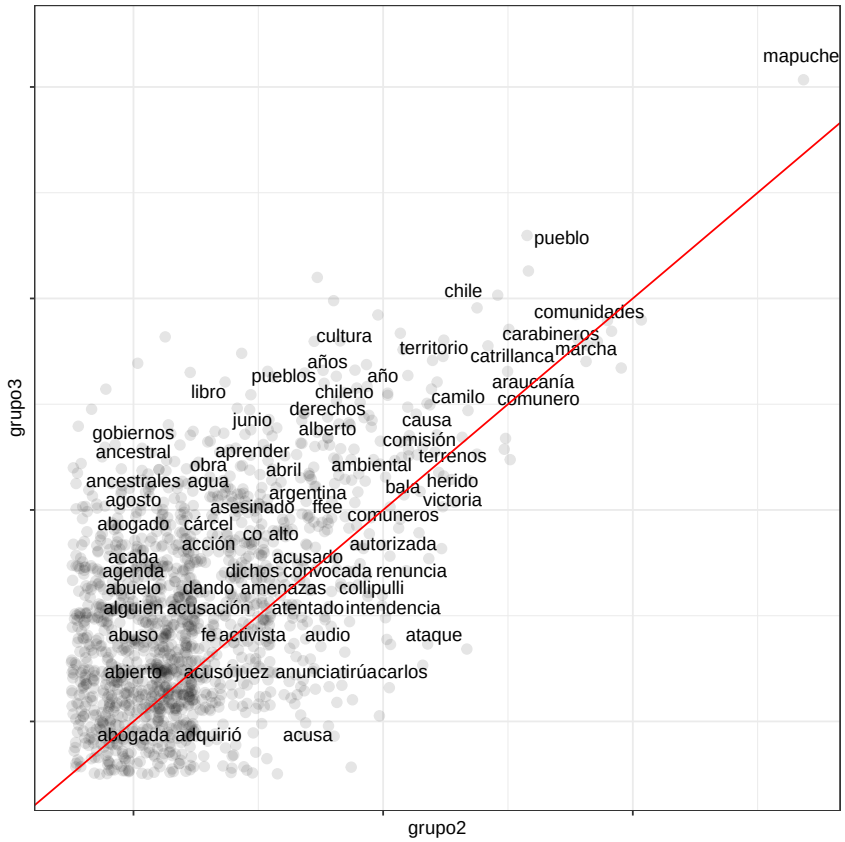


Figura 38: Comparación de las frecuencias de palabras entre grupo 2 y 3 mediante correlación.

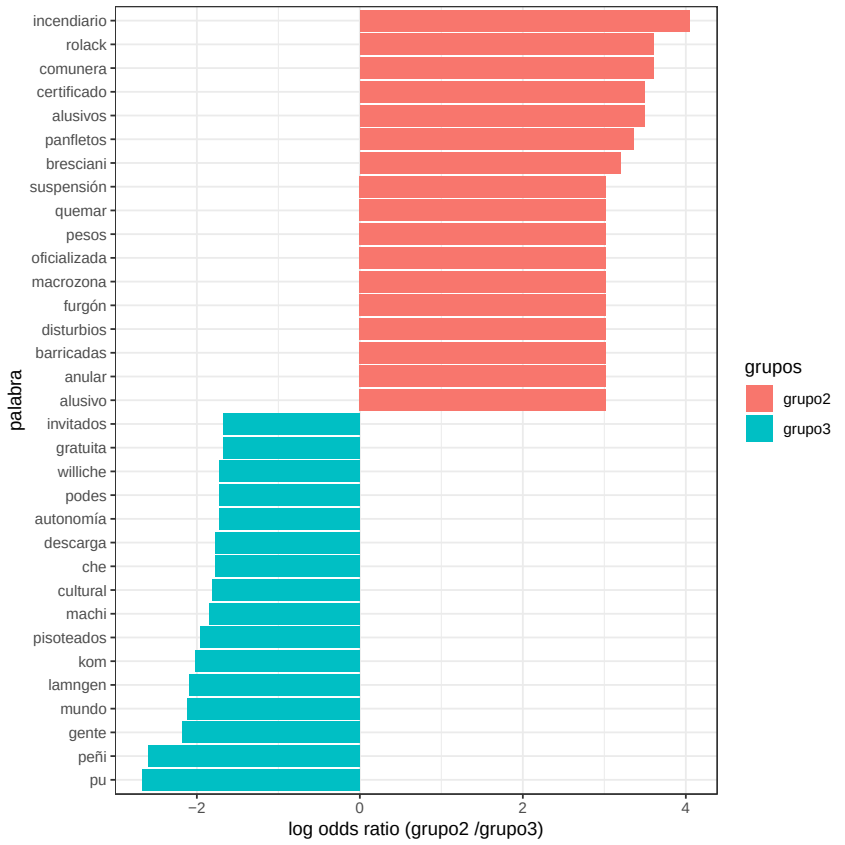


Figura 39: Comparación de las frecuencias de palabras entre grupo 2 y 3 mediante log of odds ratio de las frecuencias .

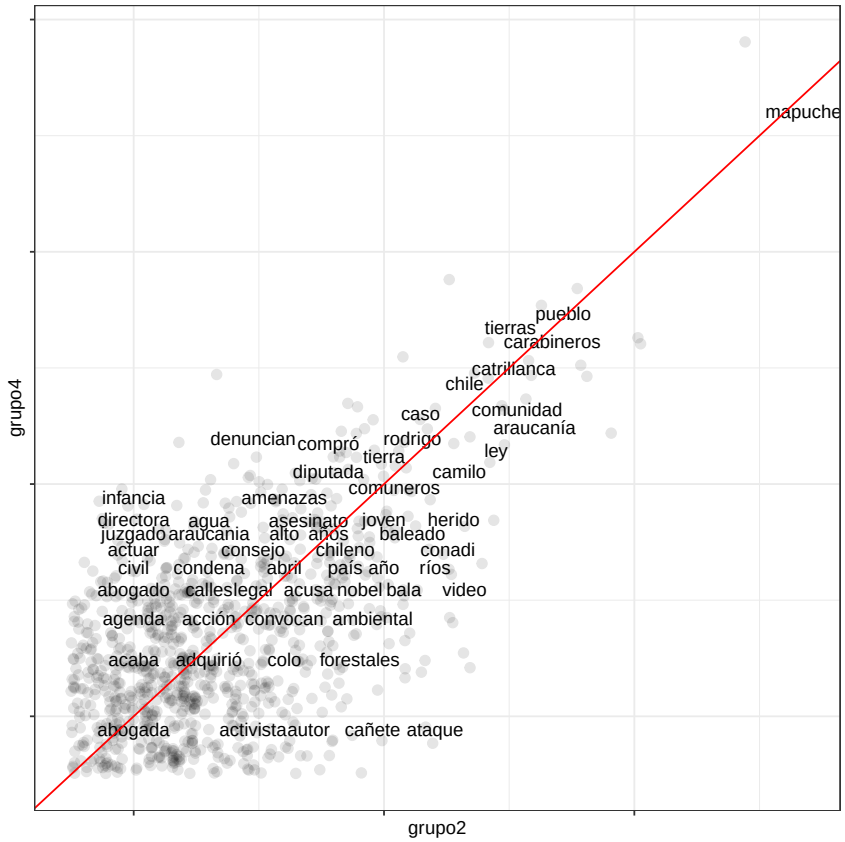


Figura 40: Comparación de las frecuencias de palabras entre grupo 2 y 4 mediante correlación.

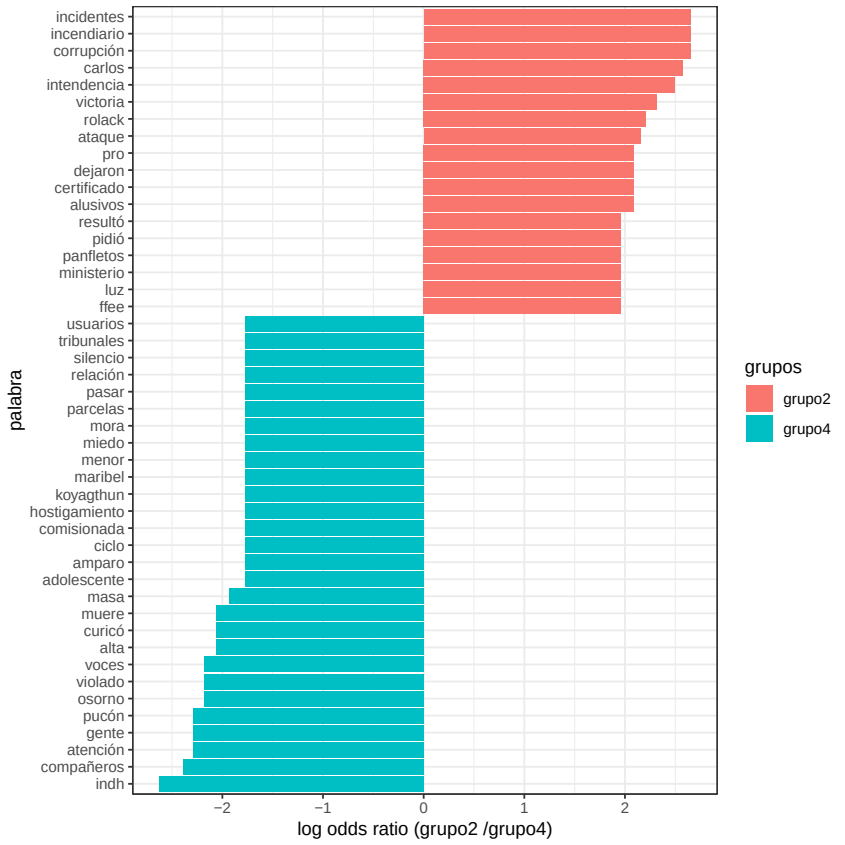


Figura 41: Comparación de las frecuencias de palabras entre grupo 2 y 4 mediante log of odds ratio de las frecuencias .

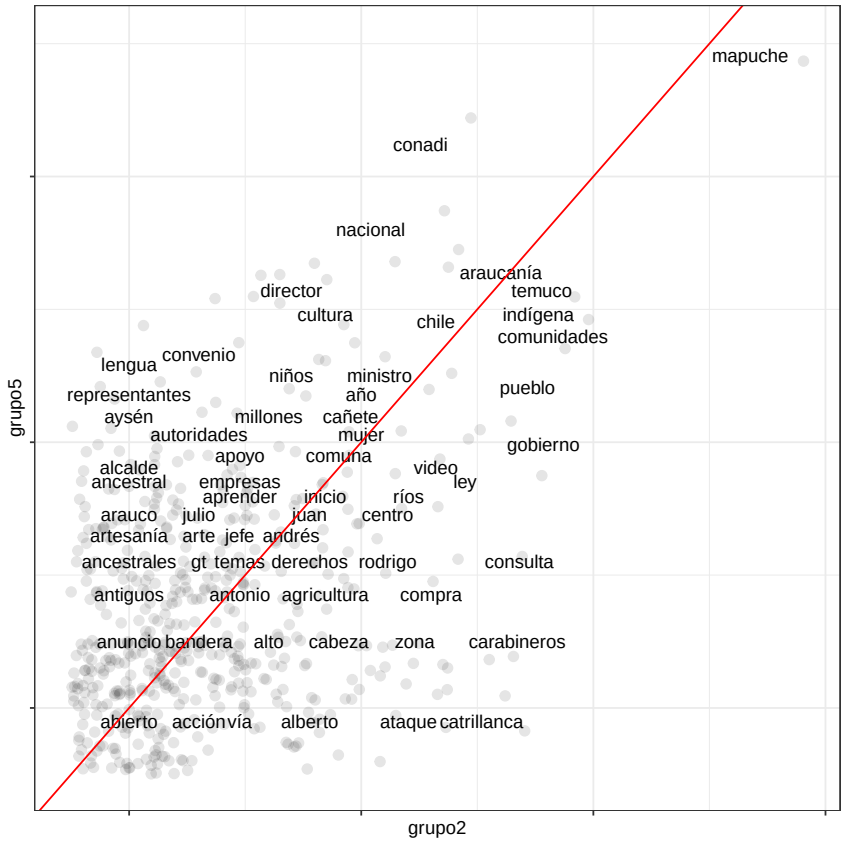


Figura 42: Comparación de las frecuencias de palabras entre grupo 2 y 5 mediante correlación.

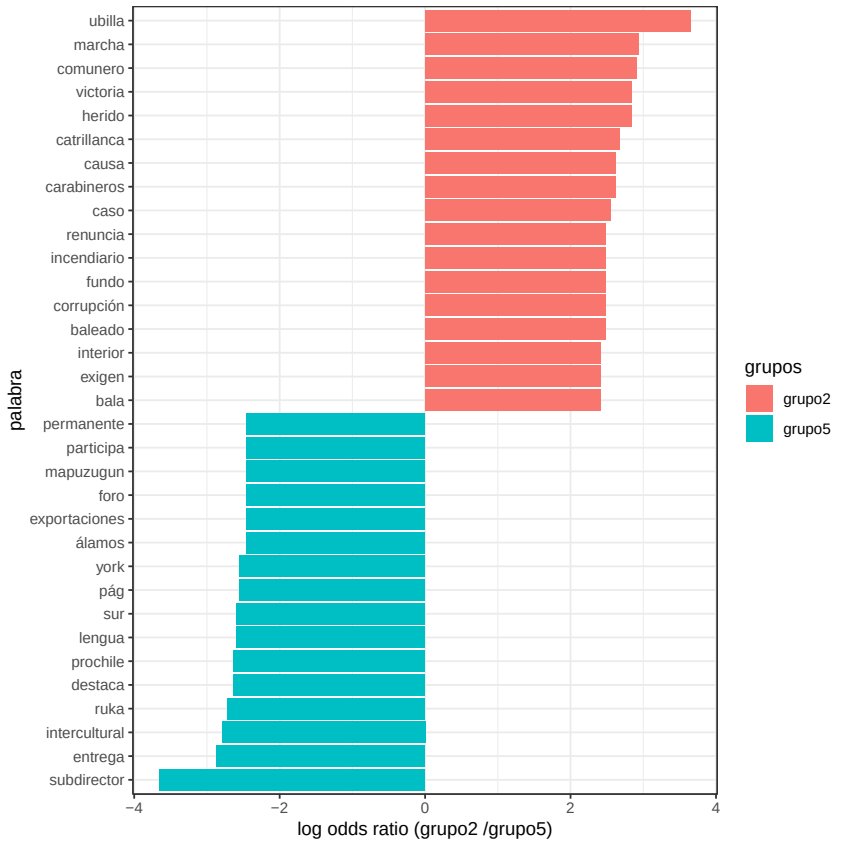


Figura 43: Comparación de las frecuencias de palabras entre grupo 2 y 5 mediante log of odds ratio de las frecuencias .

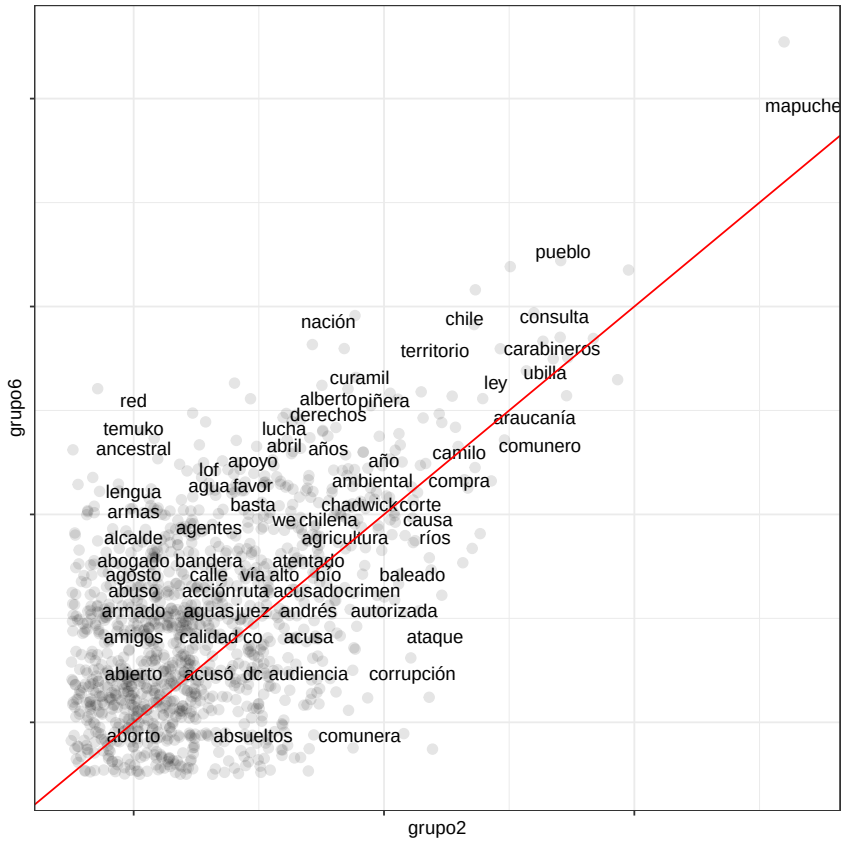


Figura 44: Comparación de las frecuencias de palabras entre grupo 2 y 6 mediante correlación.

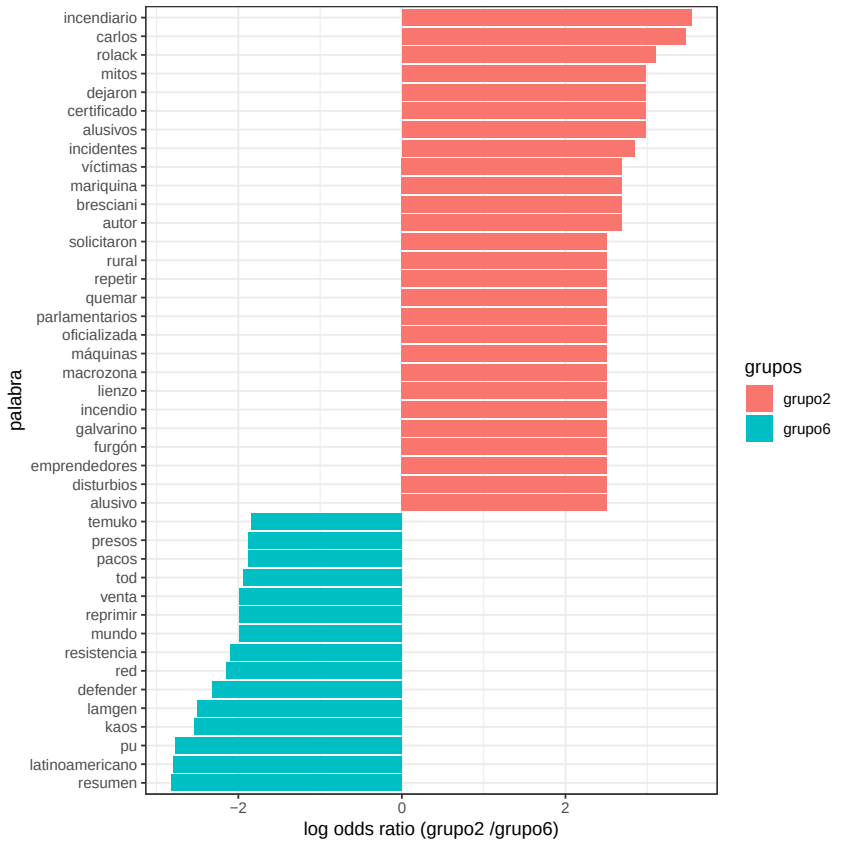


Figura 45: Comparación de las frecuencias de palabras entre grupo 2 y 6 mediante log of odds ratio de las frecuencias .

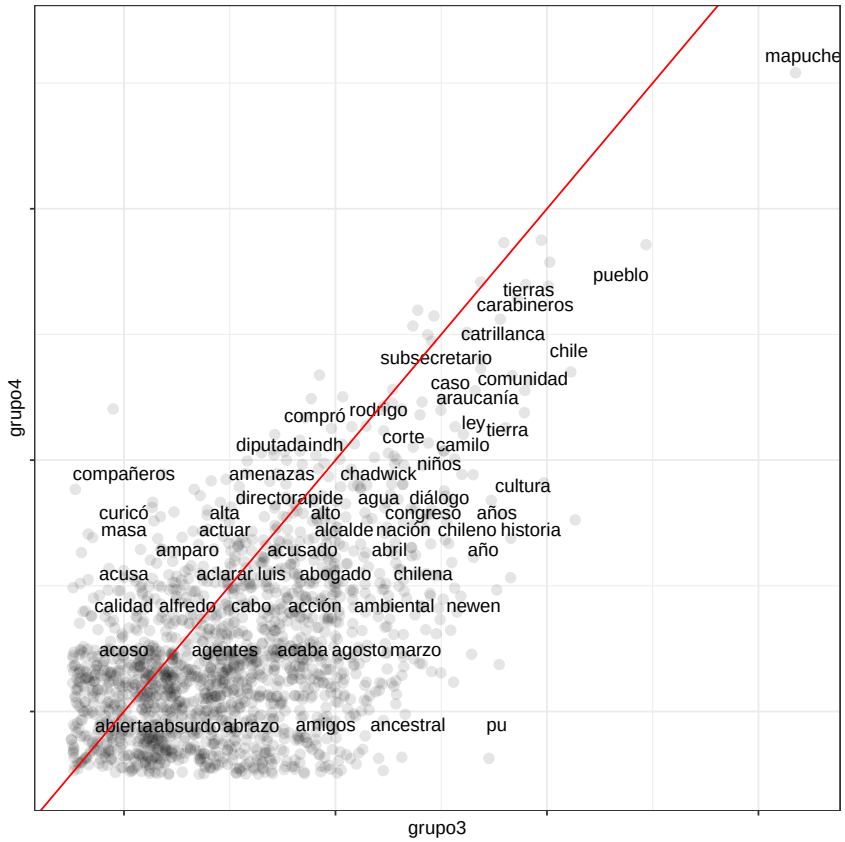


Figura 46: Comparación de las frecuencias de palabras entre grupo 3 y 4 mediante correlación.

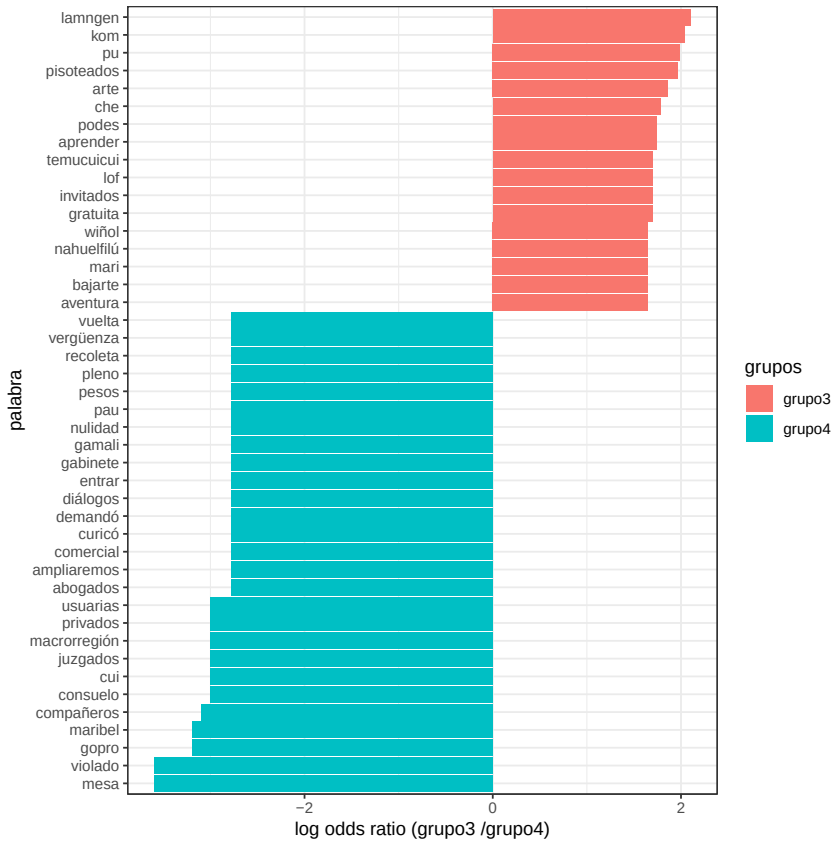


Figura 47: Comparación de las frecuencias de palabras entre grupo 3 y 4 mediante log of odds ratio de las frecuencias .

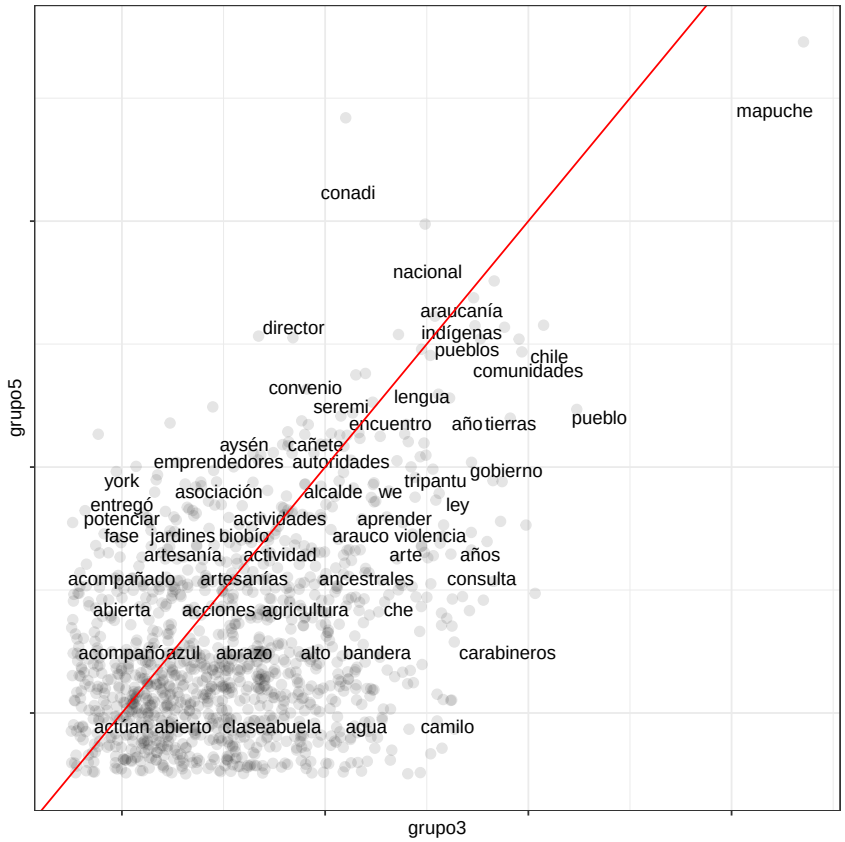


Figura 48: Comparación de las frecuencias de palabras entre grupo 3 y 5 mediante correlación.

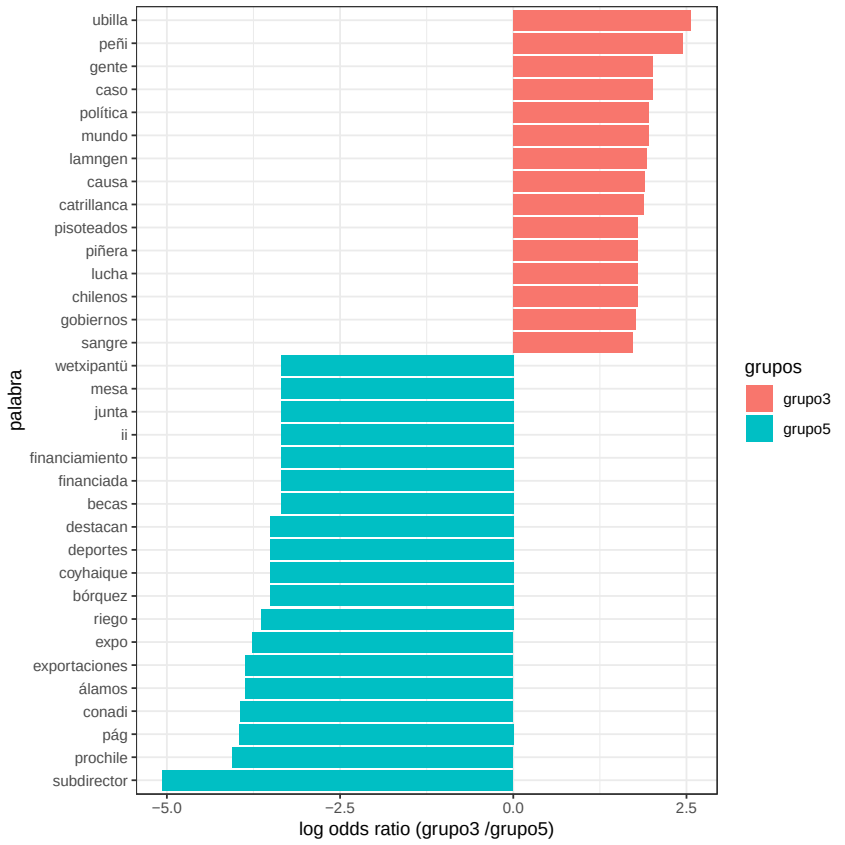


Figura 49: Comparación de las frecuencias de palabras entre grupo 3 y 5 mediante log of odds ratio de las frecuencias .

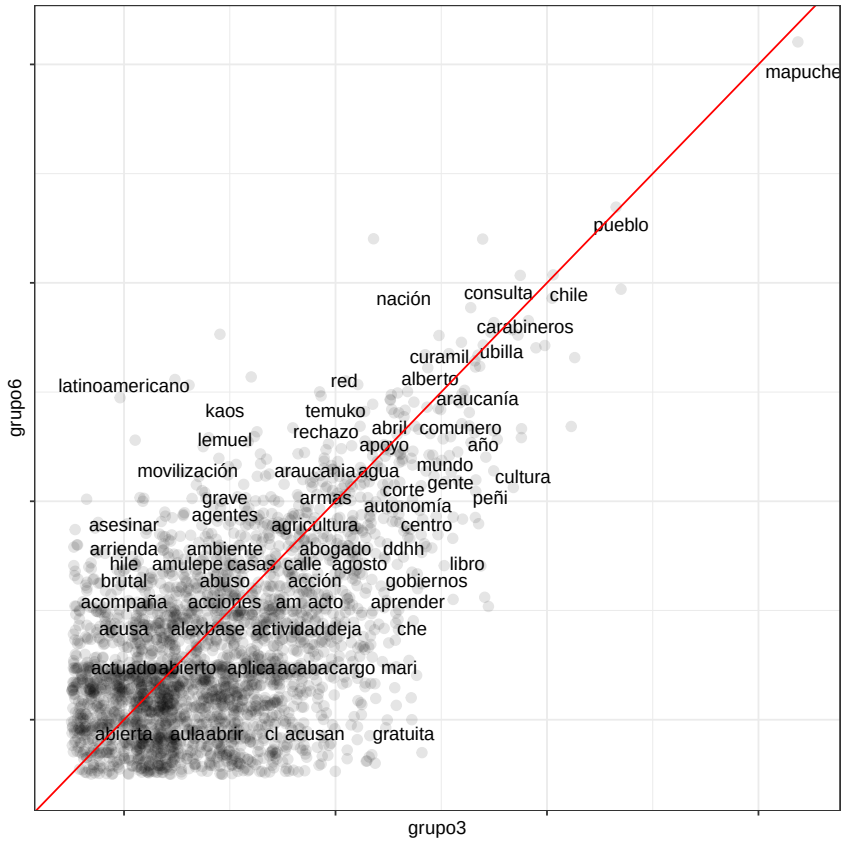


Figura 50: Comparación de las frecuencias de palabras entre grupo 3 y 6 mediante correlación.

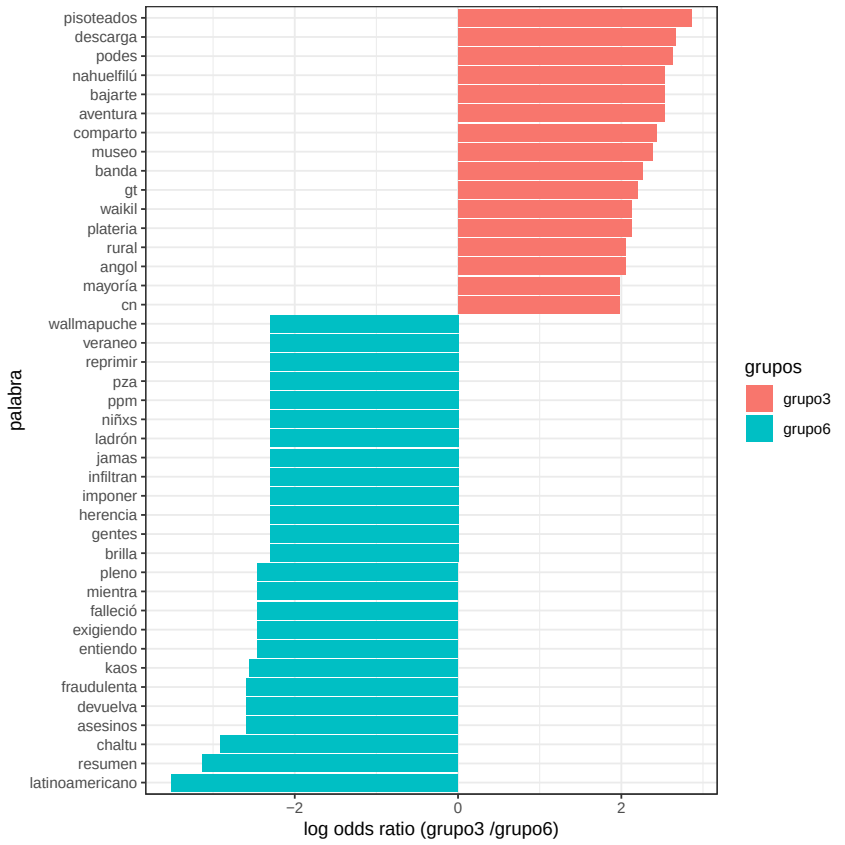


Figura 51: Comparación de las frecuencias de palabras entre grupo 3 y 6 mediante log of odds ratio de las frecuencias .

Figura 52: Comparación de las frecuencias de palabras entre grupo 4 y 5 mediante correlación.

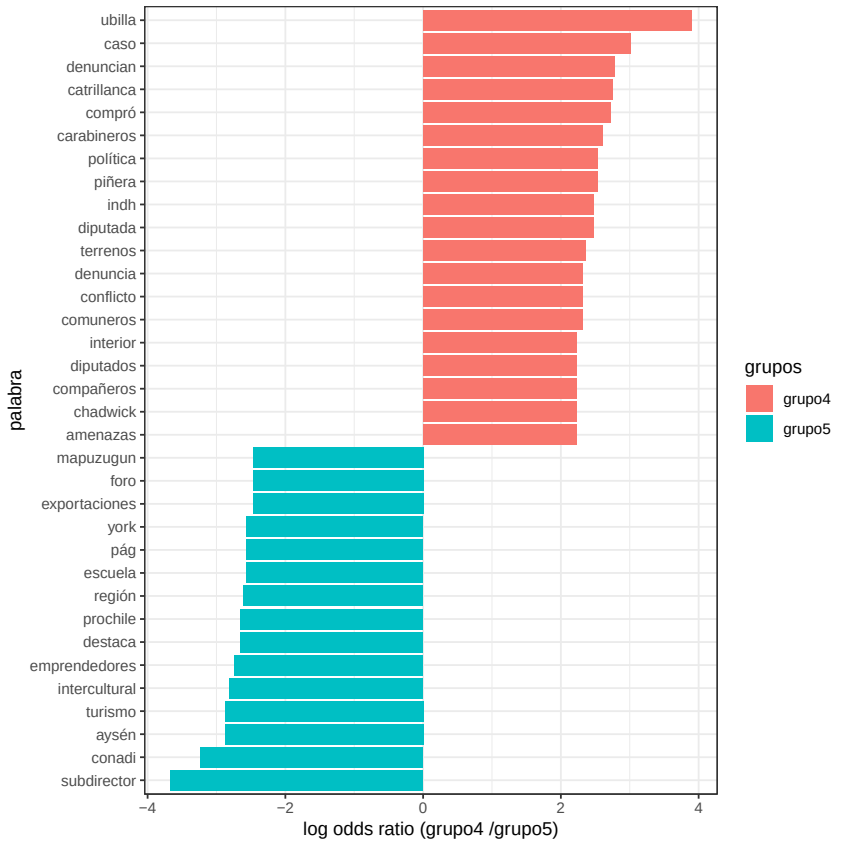


Figura 53: Comparación de las frecuencias de palabras entre grupo 4 y 5 mediante log of odds ratio de las frecuencias .

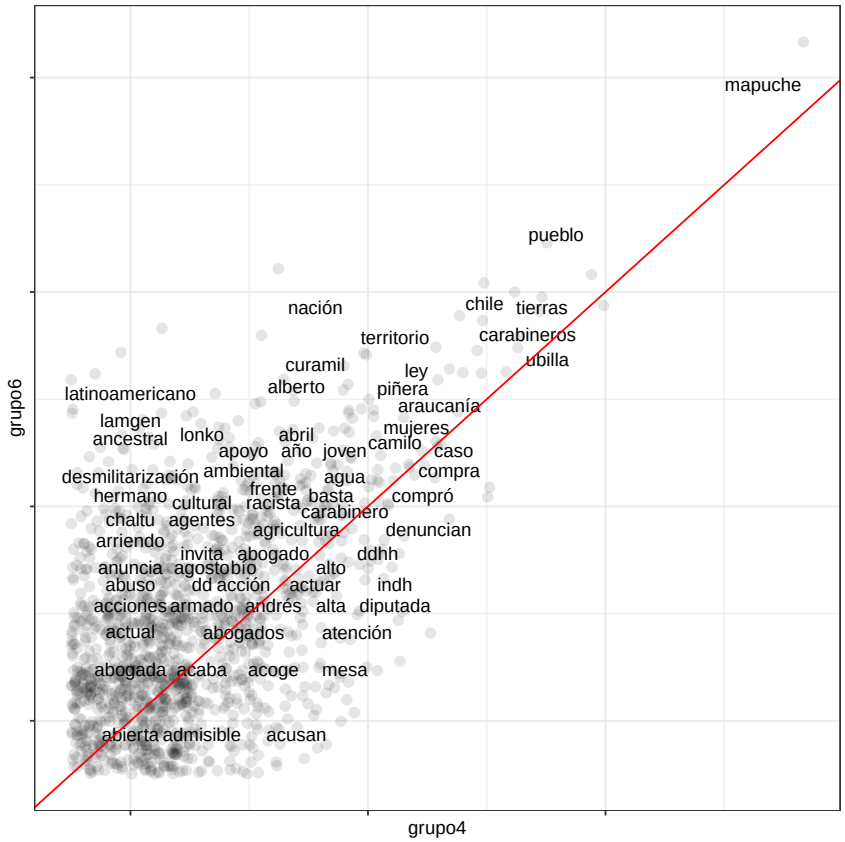


Figura 54: Comparación de las frecuencias de palabras entre grupo 4 y 6 mediante correlación.

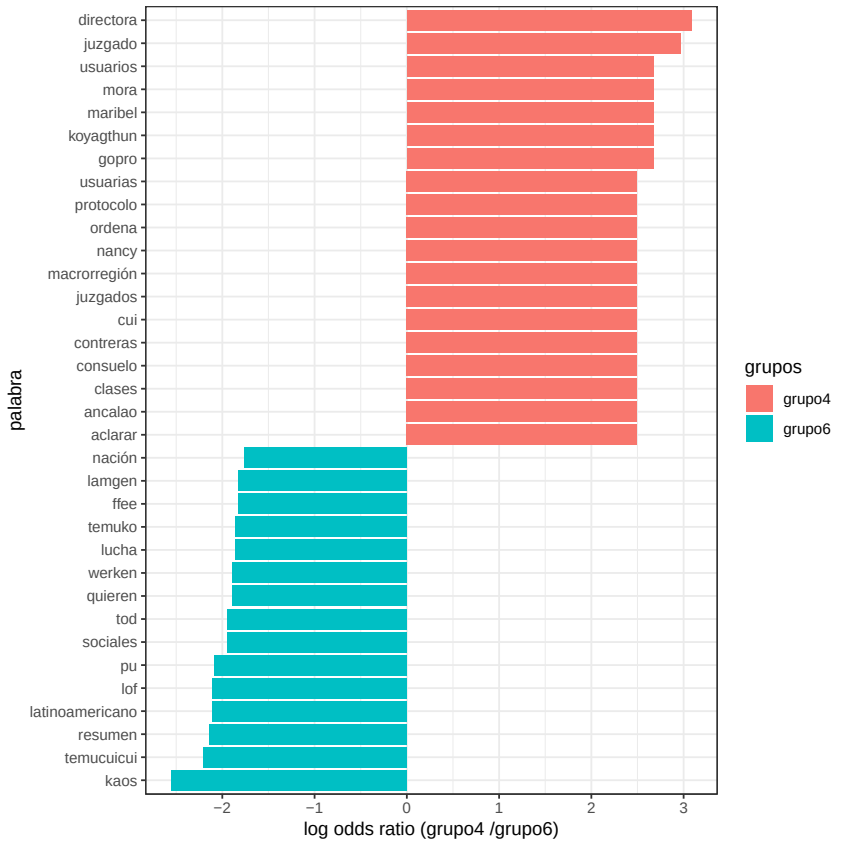


Figura 55: Comparación de las frecuencias de palabras entre grupo 4 y 6 mediante log of odds ratio de las frecuencias .

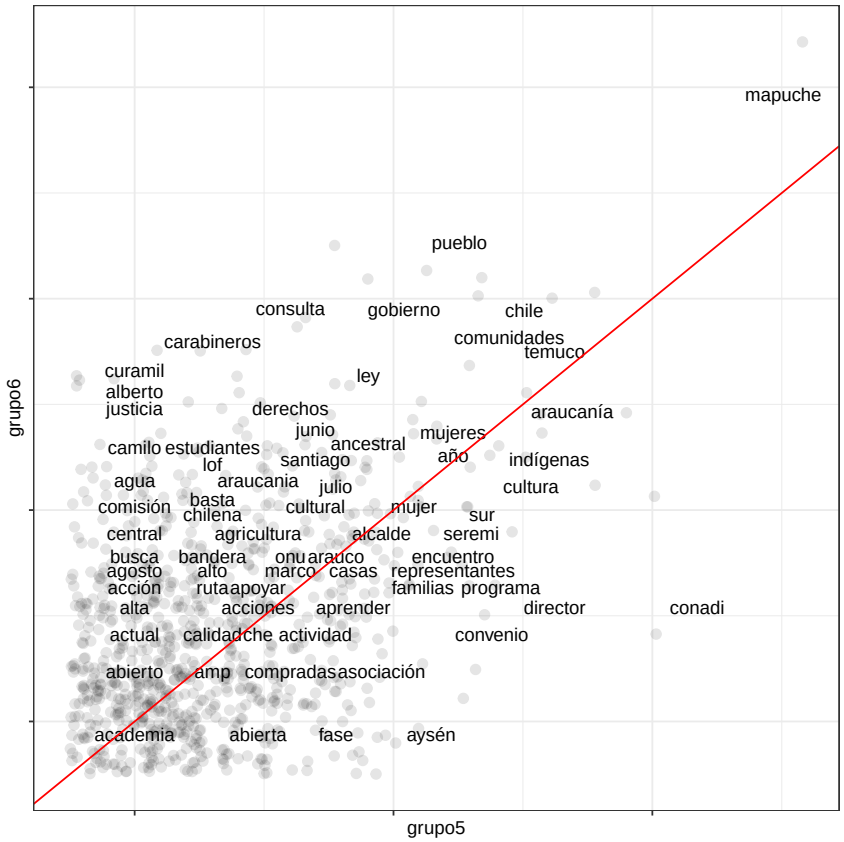


Figura 56: Comparación de las frecuencias de palabras entre grupo 5 y 6 mediante correlación.

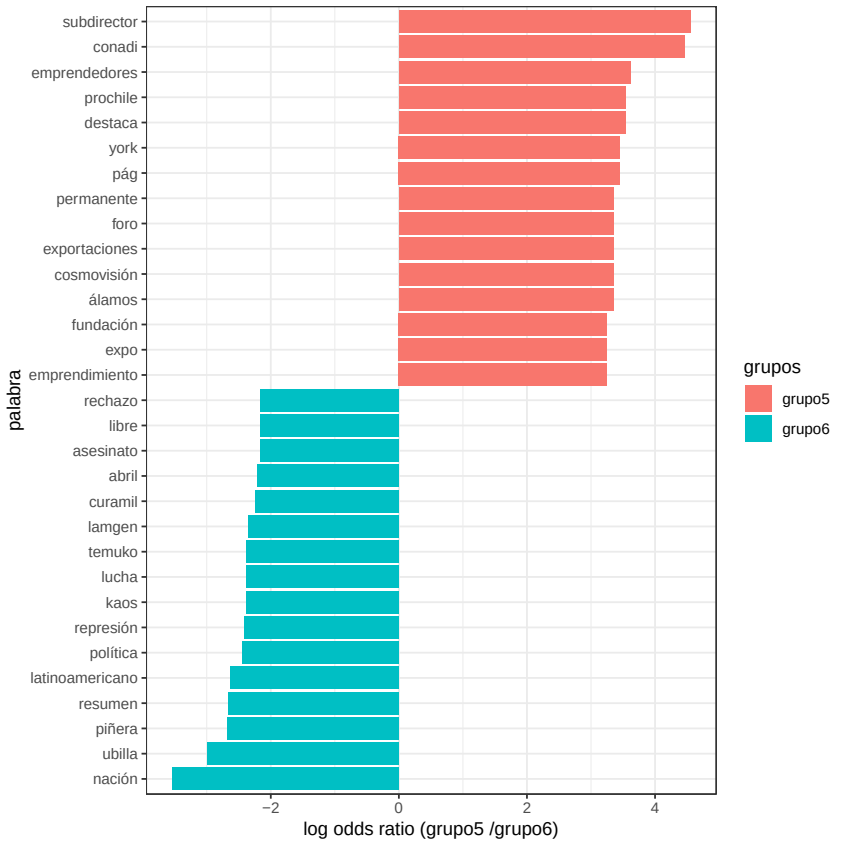


Figura 57: Comparación de las frecuencias de palabras entre grupo 5 y 6 mediante log of odds ratio de las frecuencias .

5.11. Análisis de sentimiento

Para obtener una visión general de los términos positivos y negativo se analizada mediante el análisis de sentimiento mediante el paquete syzhet y mediante un diccionario.

Figura 58: mediante el paquete syzhet.

Podemos observar que los tweets clasificado positivamente se encuentra entre los términos más frecuente justicia, derecho, apoyo, histórico, tierra, cultura sintetizan las demandas del pueblo mapuche. Por otra parte, los tweets clasificado negativamente el término más frecuente es gobierno, lo que llama mucho la atención . En este contexto podemos deducir que existe una opinión negativa hacia el gobierno de chile hacia las demandas del pueblo mapuche.

5.12. Clasificación mediante variables relacionada con los usuarios

nos interesa saber la opinión sobre el conflicto mapuche , para ello se consideran algunas variables relacionada con los usuarios para generan un criterio de clasificación . Se aplica la técnica de arbol de decisión , ya que , tiene una fácil interpretación Se consideran .

- Partido política
- edad
- sexo
- ubicación
- profesión

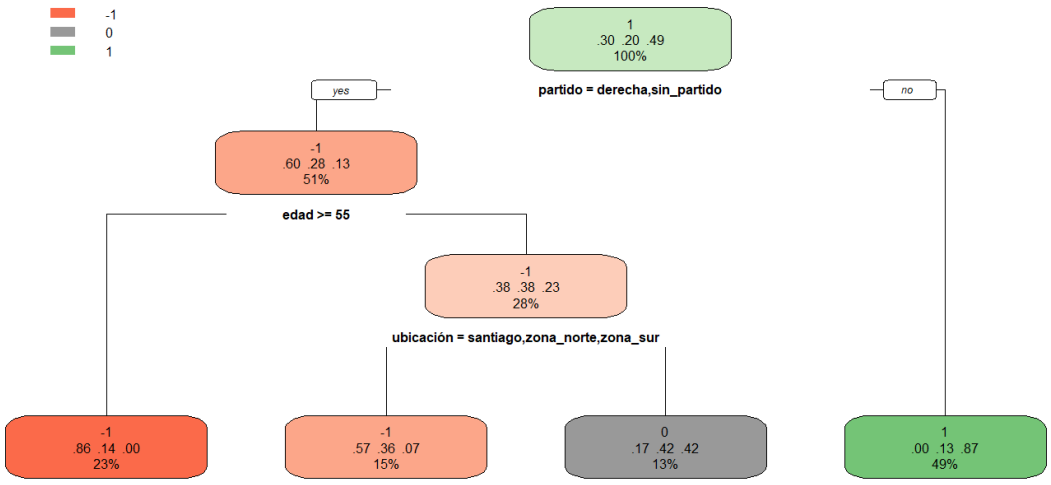


Figura 59: Árbol de decisión

Se visualiza que 51 % de los usuarios pertenecen a la derecha o no tiene definido su partido político,de ellos 23 %, con una edad superior a los 55 años , un 86 % opinan de forma negativa sobre el conflicto mapuche y un 14 % son neutro antes la problemática mapuche . También podemos observar otros usuarios que no pertenecen ni a la derecha o no tiene definido su partido político se encuentra un 49 % que apoya las demandas del pueblo mapuche.

6. Conclusión

Este trabajo ha abordado tareas del análisis de opinión expresado por los usuarios en las redes sociales, más concretamente sobre Twitter.

En cuanto al análisis descriptivo se observar una frecuencia considerable de palabras o combinaciones de palabras aludiendo a las demandas del pueblo mapuche. Por ejemplo, tierras, justicia, causa, territorio, derecho, ancestral, originario, entre otras. Debido que estos termino presenta una mayor frecuencia se puede entender que una gran porción de los usuarios de twitter apoya el conflicto mapuche.
También se encuentra otro términos o conjuntos de palabras (forma una oración). Por ejemplo, caso Camilo Catrillanca, subsecretario Rodrigo Ubilla compró, premio nobel ambiental verde, que aluden a temas actuales que afectan al pueblo mapuche. Por otra parte, encontramos justicia reconoció, agente encubierto, corte suprema, prisión preventiva , fallo histórico, ley anti terrorista indígena, fuerza especial, violeto operativo policial que aluden al gobierno.

En cuanto al análisis de sentimiento, las principales conclusiones que obtenemos tiene que ver con el gran número de tuits neutrales, en su mayoría se relacionado con las entidades gobernarles y políticos. Los tuis negativos y positivos difieren significativamente, debido a diferentes factores socio político o económico de los usuarios. La mayoría de los tuits son negativo, mediante el paquete de R utilizado o un diccionario codificado, los dos funcionan de forma igual (mediante comparación simple), el problema que presenta al momento de análisis el sentimiento de los usuarios es la orientación, es decir, el contexto. Por ejemplo, algunos usuarios apoyan las demandas del pueblo mapuche, pero en sus tuits se encuentran más términos negativos que positivos, por ende, el tuit se clasifica negativo. Para intentar solucionar este problema se intenta visualizar los términos más frecuentes, positivo versus negativo, observando una diferencia significativa.

Por último, se ha utilizado diferente técnicas de minería de texto; no se centra solo en la estadística descriptiva. Por ejemplo, se realizo un análisis de sentimiento con el paquete syzhet mediante con el léxico 'NRC' y otro mediante un diccionario codificado. Por otra parte, la creación de comunidades entre los usuarios mediante el algoritmo de Girvan y Newman que presenta una buena separación entre los grupos, es decir, su la Modularidad es 0.87. para analizar estas comunidades se aplicando algunas estadísticas descriptiva y análisis de sentimiento para encontrar diferencias. Se puede apreciar que existen un solo grupo que opina de forma negativa, en el sentido que no apoya las demandas mapuches, este grupo en su mayoría tienen una gran simpatía por algunos partidos políticos .
La separación de los grupos es buena, pero falta buscar otras medida para que términos o conjuntos de palabras puede rescatar lo esencial.

7. Bibliografía

Aguirre J.L.(2011). Introducción al Análisis de Redes Sociales.

Amolef, G.F. (2000). La alteridad en el discurso mediático: Los Mapuches y la prensa chilena.

Bengoa, J. (2002). Historia de un conflicto. el estado y los mapuches en el siglo XX. 2da. Edición. Editorial Planeta: Santiago.

Cordon Oscar (2007). Redes y Sistemas Complejos. https://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/RedesSistemasCompejos/Tema03-RedesSociales-13-14.pdf

Duran, J. (2016). Redes sociales y comunicación publicitaria. La metodología SETAL para la obtención de insights publicitarios.

El Mostrador. Consultado el 15 de diciembre de 2019. Redacción (9 de julio de 2019). «Caso Catrillanca vuelve a hacer sombra en La Moneda: comisión investigadora aprueba informe que establece responsabilidades políticas de Chadwick y Ubilla». https://www.elmostrador.cl/noticias/pais/2019/07/09/caso-catrillanca-vuelve-a-hacer-sombra-en-la-moneda-comision-investigadora-aprueba-informe-que-establece-responsabilidades-politicas-de-chadwick-y-ubilla/

Gil, F. (2010). Fuentes de análisis para el estudio de la prensa diaria.

Huenchumilla, F. (2002). Cómo los mapuches fueron despojados por el estado y las huincas. Recuperado de https://www.archivochile.com/Pueblos_originarios/hist_doc_gen/PODocgen0010.pdf

Landa Javier (2016). ¿Qué es KDD y Minería de Datos? http://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/

Liddy, E.D. (2001). Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

Loor, G. (2017). El “conflicto en Mapuche”: ejemplo de lucha, constancia y reivindicación. Recuperado de https://abpecuador.wixsite.com/ecua/single-post/2017/11/22/El-%E2%80%9Cconflicto-en-Mapuche%E2%80%9D-ejemplo-de-lucha-constancia-y-reivindicaci%C3%B3n.

Lorenzo Mateo, A. (2016). Detección de comunidades en redes sociales, estudio sobre las elecciones catalanas y elecciones generales de 2015 (Bachelor's thesis).

Kuz, Falco & Giandini(2016). Análisis de redes sociales: un caso práctico.

Luke, D. A. (2015). A user's guide to network analysis in R. Springer.

Maldonado, C. (2011). Narrativa hipertextual mapuche: emplazamiento y reivindicación cultural en youtube. Revista de Comunicación de la SEECI Año XV, no 26, pp. 62-70. Disponible en https://doi.org/10.15198/seeci.2011.26.62-70

Martínez, F. (2017). Análisis de sentimiento en twitter de las principales compañías del sector asegurador español.Tesis Universidad de Valencia para optar al grado de Magíster en ciencias actuariales y financieras.

Ornelas, E. L., Mena, R. A., & Hernández, S. Z. (2018). Detección y análisis de comunidades en redes sociales (#TodosSomosPolitécnico). Pistas Educativas, 36(112).

Rodríguez León, Ciro, & García Lorenzo, María Matilde. (2016). Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor. *Revista Universidad y Sociedad*, *8*(4), 43-53. Recuperado en 13 de enero de 2020, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218-36202016000400005&lng=es&tlng=es.

Silge, J. and Robinson, D. (2017). Text mining with R: A tidy approach. o'Reilly Media, Inc.".

Kwartler, Ted. (2017). *Text Mining in Practice with R.* Wiley Online Library.

Tascón, M. y Quintana, Y. (2012). Ciberactivismo. Las nuevas revoluciones de las multitudes conectadas. Madrid: Catarata.

Telesur (2017). Conflicto Mapuche en Chile: Razones de la lucha y sus demandas. Recuperado de : https://www.telesurtv.net/news/Conflicto-Mapuche-en-Chile-Razones-de-la-lucha-y-sus-demandas-20171004-0008.html#.

Wiebe, J. M. (1994). Tracking Point of View in Narrative. *Comput. Linguist., 20(2), 233– 287.*

Tunazzina Islam. (2019). Yoga-Veganism: Correlation Mining of Twitter Health Data. *In Proceedings of 8th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM@KDD'19), August 04–08, 2019, Anchorage, Alaska, USA. ACM, New York, NY, USA, 7 pages. Disponible en:* https://doi.org/10.1145/

Velázquez, A. y Aguilar, N. (2005). Manual introductorio al análisis de redes sociales. Disponible en: http://revista-redes.rediris.es/webredes/talleres/Manual_ARS.pdf