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

METODOLOGÍA PARA EL DISEÑO Y CONSTRUCCIÓN DE UN LEXICÓN DE OPINIÓN, BASADO EN COMENTARIOS DE TWITTER APLICADO AL PROYECTO "OPINIONZOOM"

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

NATALIA PAOLA HERNÁNDEZ MUÑOZ

PROFESOR GUÍA:
JUAN VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
FELIPE VILDOSO CASTILLO
PATRICIO MOYA MUÑOZ

Este trabajo ha sido parcialmente financiado por el proyecto CORFO 13IDL2-23170

SANTIAGO DE CHILE
2016

RESUMEN DE LA MEMORIA PARA OPTAR AL
TITULO DE: Ingeniera Civil Industrial
POR: Natalia Paola Hernández Muñoz
FECHA: 21/01/2016
PROFESOR GUÍA: Juan Velásquez Silva

METODOLOGÍA PARA EL DISEÑO Y CONSTRUCCIÓN DE UN LEXICÓN DE OPINIÓN, BASADO EN COMENTARIOS DE TWITTER APLICADO AL PROYECTO "OPINIONZOOM"

El presente trabajo tiene como objetivo diseñar y construir una metodología para la creación de un lexicón de opinión en el que se identifique su polaridad, considerando las características del español de Chile y basado en comentarios de Twitter, aplicado al proyecto "OpinionZoom". Es desarrollado dentro del proyecto "OpinionZoom, plataforma de análisis de sentimientos e ironía a partir de la información textual en redes sociales para la caracterización de la demanda de productos y servicios", donde se presenta la problemática de no tener un lexicón de opinión apropiado para el análisis de sentimientos que se realiza.

La hipótesis de investigación de este trabajo postula que la construcción de un lexicón de opinión que considere las particularidades del español de Chile en Twitter mejora el desempeño de la herramienta "OpinionZoom". Para comprobar esta hipótesis se ha utilizado una metodología basada en un corpus lingüístico para la generación de un lexicón de opinión.

Se construyó un corpus de tweets clasificados en positivos y negativos según los emoticones que presentan, luego se utilizó este corpus en la construcción del lexicón, utilizando la frecuencia de las palabras presentes en comentarios positivos y negativos y calculando su polaridad en base a la información mutua que se tiene, empleando el cálculo de PMI.

Para la validación del lexicón de opinión se midió el desempeño del sistema de análisis de opiniones con el lexicón de opinión actual, que presenta licencia sólo de uso académico, y con el lexicón de opinión construido en este trabajo. Comparando ambos desempeños, se observaron mejoras en cuanto a exactitud, precisión y exhaustividad para el sistema con el lexicón construido, por lo que la hipótesis planteada en este trabajo se comprueba.

En conclusión, la utilización de un lexicón de opinión que considere las características del español de Chile mejora el desempeño del sistema de análisis de opiniones de "OpinionZoom", la utilización de emoticones para identificar la polaridad representa un indicador representativo en comentarios de Twitter, por lo que se puede ampliar la investigación utilizando *emojis* para la identificación de polaridad.

"Un corazón ardiendo de amor es, necesariamente, un corazón lleno de alegría."

- Madre Teresa de Calcuta

Agradecimientos

Quisiera iniciar mis agradecimientos a quienes me han acompañado en todo mi proceso de formación como profesional y como persona, quisiera agradecer a mi familia; a mi mami por todo el apoyo incondicional, amor y enorme sabiduría que siempre me ha brindado, a mi papi, quien incentivó el camino hacia la ingeniería y a mi hermano, con quien he compartido durante gran parte de mi vida.

Continuar agradeciendo a mi compañero de este último tiempo, quien ha estado presente en mis últimos años de carrera, me ha apoyado y ayudado durante todo este proceso; gracias Fernando, gracias por todo el apoyo, cariño y amor que me has brindado.

Gracias a mis compañeras de vida, gracias a mis queridas amigas, que conocí en el colegio y que hasta el día de hoy han sido parte fundamental de mi formación como persona, gracias a: Javi, Pasi, Peque, Pía, Cata y a mi querida tía madrina, quien ha estado presente en gran parte de mis locuras, Memé. Agradecer por su puesto a Samuel, por su confianza y compañía.

Agradecer también a mis amigos, quienes han estado presente en gran parte de mi carrera universitaria, a quienes estuvieron desde sus inicios: Pedro, Ismael, Esteban, Felipe, Ramiro, a quienes fui conociendo en el camino, Iván, Paulina, Mauricio, Gustavo y a mi querido grupo cine: César, Dani, Jimbo y Oscar.

Muchas gracias al profesor Juan Velásquez, quien me dio la oportunidad de realizar este trabajo, gracias por toda la sabiduría entregada para llevar a cabo esta investigación. Muchas gracias a Felipe y Patricio por todo el apoyo y por la disposición a ayudarme en todos los momentos en que lo necesité. Extender mis agradecimientos a quienes estuvieron durante la gestación de este trabajo de título y ayudaron a sacar una buena memoria, a mis profesores del E y F: Alejando Muñoz, Daniel Varela y Jorge Aravena.

Agradezco a todo el equipo WIC que colaboró y apoyó este trabajo, un hermoso equipo humano del cual aprendí y mantengo un gran cariño. Gracias a la Salita Sur por toda su compañía, comprensión y ayuda, gracias a Felipe, Andrés, Rocío, Nicole, Rominna, Pangui; Gracias a los trabajos anteriores realizados en el centro, gracias a: Jorge, Víctor y a quienes colaboraron de una u otra forma a sacar adelante el proyecto: Yerko, Gaspar.

Agradecer a mis minions, a todas aquellas almas que colaboraron con la clasificación de Tweets.

Agradecer a todos y cada uno de los que ha sido parte de mi vida y de mi trabajo.

Finalmente concluir agradeciendo a Dios, por todo cuanto me ha brindado y amado.

Tabla de contenido

Índice de tablas

Índice de figuras

Capítulo 1

1 Introducción

Con el aumento de los blogs, foros, microblogs en la Web, las personas han empezado a expresar sus opiniones sobre diversos temas en plataformas sociales como Facebook, Twitter, Google+, LinkedIn, entre otras.

Actualmente el interés por la caracterización y comprensión de los clientes son conceptos clave para cualquier negocio. Es importante mencionar que la caracterización de la demanda es un proceso complejo que apunta a lograr una predicción de demanda de la forma más acertada posible, lo que se logra con el conocimiento de sus clientes y necesidades, logrando también una importante reducción de costos para empresas.

Aprovechando el crecimiento que ha tenido la Web, que inició en 1989 con el sistema propuesto por Tim Berners-Lee llamado World Wide Web [1], y la gran cantidad de información disponible actualmente, se ha hecho necesario la aplicación de técnicas de Inteligencia Web (*Web Intelligence*) para ser capaz de procesar textos e identificar temas importantes que no surgen con métodos tradicionales, y que presentan una solución a la compleja tarea de comprender el comportamiento de clientes, sus preferencias, además de lograr identificar estrategias que beneficien el negocio.

Las opiniones están siendo declaradas en diversas fuentes, desde la información interna de una organización con comentarios del cliente vía e-mail, call centers; en diarios y reportajes con comentarios sobre noticias y artículos, experiencias personales y opiniones plasmadas en blogs, foros, Twitter, Facebook, redes sociales en general.

Así, podemos ver diversas aplicaciones del análisis de opiniones, ya sea para organizaciones en temas de marketing, conocimiento de los consumidores, productos y servicios otorgados; para personas individuales a quienes les permita tomar una mejor decisión sobre productos o servicios, para conocer la opinión pública de candidatos políticos.

1.1 Descripción del proyecto y justificación

El siguiente proyecto se desarrolla dentro del proyecto OpinionZoom[1]. Este proyecto INNOVA CORFO 13IDL2-23170 es titulado como "OpinionZoom: Plataforma de análisis de sentimientos e ironía a partir de la información textual en redes sociales para la caracterización de la demanda de productos y servicios".

El proyecto se desarrolla dentro del centro de investigación de inteligencia web: Web Intelligence Centre [2] "WIC" del departamento de Ingeniería Civil Industrial de la Universidad de Chile.

1.1.1 Web Intelligence Centre

El centro de inteligencia web "Web Intelligence Centre" (WIC) es un centro de investigación, miembro del Web Intelligence Consortium. Este centro de investigación ha estado a cargo del académico Juan Velásquez desde sus inicios como director del centro. En la página web se declara su misión, visión y objetivos:

Misión del WIC:

"Desarrollar investigación de frontera en el campo de Tecnologías de Información creando nuevas soluciones para abordar problemas complejos de ingeniería utilizando herramientas basadas en la Web de las Cosas".

Visión:

"Ser un líder a nivel internacional en la investigación de tecnologías de información y comunicaciones aplicadas a la resolución de problemas del mundo real".

Objetivos del centro:

- Publicar en las principales revistas, conferencias y editoriales relacionadas con Web Intelligence.
- Proveer un servicio profesional, excelente y rápido para todos nuestros clientes.
- Dictar cursos de orientación práctica acerca de las Tecnologías de Información y su aplicación en los negocios.

[1] http://www.opinionzoom.cl/index, 11 de enero de 2016
[2] http://www.wic.cl/, 11 de enero de 2016

1.1.2 OpinionZoom

El proyecto OpinionZoom es una aplicación que se dedica a analizar la información de usuarios de redes sociales, principalmente en Twitter, quienes comentan sobre determinadas organizaciones, marcas, productos y servicios que a partir de lo anterior, se pueden patrones para estudiar el comportamiento y caracterizar demanda.

El objetivo general de este proyecto es:

"Desarrollar una plataforma de análisis de sentimientos e ironía a partir de la información textual en redes sociales para la caracterización de la demanda de productos y servicios".

Se contemplan los siguientes resultados específicos:

1. Construir un repositorio de palabras claves etiquetadas (corpus) en base al análisis lingüístico de los textos en una comunidad afín usuaria de redes sociales.

2. Adaptar e integrar algoritmos de data mining para extraer patrones que permitan interpretar los datos y generar modelos de caracterización de demanda de productos o servicios a partir de la información textual en redes sociales.

3. Diseñar, construir y evaluar un prototipo de plataforma de software que integre los algoritmos, los modelos y el repositorio para la caracterización de la demanda de productos o servicios a partir del análisis de sentimientos e ironía a partir de la información textual en redes sociales.

El proyecto OpinionZoom incluye dentro de sus lineamientos una misión, visión y valores [2] que se muestran a continuación:

Visión:

"Ser la empresa de más alta reputación en la industria, reconocida por entregar un servicio de análisis de opiniones preciso, confiable y rápido".

Misión:

"Entregar un servicio de extracción y análisis de opiniones, sentimientos y polaridades, que contribuya a acercar a las distintas organizaciones proveedoras de servicios y productos con los consumidores, propiciando así mejores resultados a las empresas y aporte a la calidad de vida de la comunidad en general a través de la mejor satisfacción de sus necesidades".

Valores:

"Respeto por la privacidad e identidad de las personas".

"Veracidad y transparencia en el tratamiento de los datos con el objeto que los resultados obtenidos sean confiable".

Para esto el proyecto se ha diseñado en varias etapas. En un inicio se plantea la realización de un desarrollo de repositorio de palabras y algoritmos para luego, en una segunda etapa realizar el prototipo tecnológico.

1.1.3 Motivación del trabajo

Actualmente, el análisis de opiniones que se realiza se hace en base a algoritmos de minería de opiniones utilizando un lexicón de opinión (repositorio de palabras etiquetadas). Éste es un lexicón exclusivamente de uso académico que no permite comercializar los servicios que brinda el proyecto.

Es importante mencionar que otro de los problemas detectados es que este lexicón ocupado fue realizado en Canadá, en base al idioma español utilizado en España y no aplicado a la realidad chilena; por lo que han sido añadidas de forma manual palabras y expresiones propias a la cultura chilena.

El análisis de opiniones planteado en el proyecto corresponde a un análisis en redes sociales en general, pero actualmente el centro se concentrado en realizar análisis de sentimientos en la red social "Twitter".

El presente trabajo pretende abordar el primer objetivo específico del proyecto y solucionar el problema que se da al querer comercializar los servicios que brinda "OpinionZoom", diseñando un lexicón de opinión para el proyecto y que entregue un recurso que facilite la investigación futura en temas de análisis de opinión en la cultura chilena.

1.2 Hipótesis de investigación

La expresión de opiniones en las redes sociales son una fuente de información abundante y análisis asociados a ellos se han tornado más y más relevantes en la toma de decisiones. Se observa gran cantidad de investigación y estudios de este fenómeno en comentarios de habla inglesa, pero estudios relevantes y recursos para el idioma español son escasos, considerando que el español es también uno de los idiomas que más se emplean dentro de Internet, posicionándose como el tercer lenguaje más utilizado durante el 2015 [3].

Dentro de este trabajo se ha propuesto como hipótesis de investigación que: *la construcción de un lexicón de opinión que considere las particularidades del español de Chile en Twitter mejora el desempeño de la herramienta "OpinionZoom".*

1.3 Objetivos

1.3.1 Objetivo general

Diseñar y construir una metodología para la creación de un lexicón de opinión en el que se identifique su polaridad, considerando las características del español de Chile y basado en comentarios de Twitter, aplicado al proyecto "OpinionZoom".

1.3.2 Objetivos específicos

1. Estudiar el estado del arte actual respecto a la generación de lexicones de opinión dentro del contexto de "Opinion Mining".

2. Implementar una metodología para la realización de lexicón de opinión basado en comentarios de Twitter.

3. Elaborar un lexicón de opinión basado en comentarios de Twitter que considere las características del español utilizado en Chile.

4. Validar el lexicón de opinión construido en el marco del proyecto "OpinionZoom".

1.4 Metodología

Este trabajo se inicia con una amplia revisión bibliográfica de técnicas de “Opinion Mining” utilizando investigaciones y trabajos previos en el área para construir un capítulo que conceptualice en detalle el marco teórico en que se trabaja.

Luego se propondrán una serie de alternativas metodológicas existentes para la generación de un lexicón de opinión que se ajuste al trabajo que se realiza en el proyecto OpinionZoom, cada una de ellas investigada y analizada para identificar aquella que se ajuste mejor a los requerimientos y funcionalidades esperadas del lexicón. En esta etapa se determinarán las propiedades del lexicón que permitan decidir la mejor forma de elaborarlo.

Se implementará una metodología para la construcción de lexicón de opinión que considere las características del español utilizado en Chile, para esto se realizará la recolección de datos necesarios, su categorización, para luego construir un corpus donde se apliquen las reglas para la construcción del lexicón, su distribución de palabras y la asignación de polaridad a cada una de ellas.

Finalmente, el resultado de este trabajo será validado en el sistema de análisis que presenta el proyecto “OpinionZoom”. La evaluación del lexicón de opinión elaborado se realizará identificando las métricas relevantes, analizando las mejoras que muestra con respecto al actual sistema utilizado en el centro de investigación WIC.

1.5 Contribuciones y alcances

El resultado de este proyecto es un lexicón de opinión utilizado en la aplicación de OpinionZoom. Este lexicón considera las características del español utilizado en Chile para ser integrando al sistema de análisis de sentimientos llevado a cabo dentro de la red social de Twitter.

Es un repositorio de palabras etiquetadas, que proporciona una asignación de nivel de polaridad. No proporciona reglas semánticas aplicadas sobre las palabras, así como tampoco presenta información adicional sobre la polaridad según contextos específicos en que puedan ser utilizadas.

Está orientado al lenguaje utilizado en la red social Twitter, utilizando principalmente comentarios que realicen cuentas chilenas. No se incluyen otras redes sociales dentro de este estudio.

1.6 Estructura del informe

La estructura del presente informe sigue la estructura que se detalla a continuación:

El presente capítulo, capítulo 1 trata de la introducción al trabajo realizado, identificando el proyecto, el problema, la hipótesis de investigación, los objetivos y el alcance del proyecto.

El segundo capítulo describe el marco teórico investigado en cuanto a investigaciones anteriores dentro del área, las aplicaciones y modelos que se han propuesto.

El tercer capítulo realiza una recopilación de los métodos más relevantes para la generación de lexicones que existen, identificando algunas de sus ventajas y desventajas.

Luego, en el capítulo 4 se muestra el diseño para la construcción de lexicón, donde se muestra la metodología propuesta, el proceso de construcción y los recursos utilizados para su elaboración.

En el capítulo 5 se detalla la construcción del lexicón de opinión junto al proceso de construcción del corpus lingüístico utilizado como base para la elaboración del lexicón.

En el sexto capítulo se evalúa y discuten los principales resultados del lexicón construido, para finalmente en el capítulo 7 presentar las conclusiones del trabajo realizado, junto a algunas recomendaciones y trabajos futuros propuestos.

Capítulo 2

2 Marco conceptual

Dentro de este capítulo se presenta el marco teórico necesario para el entendimiento del estudio realizado, se inicia con las tecnologías y servicios generados gracias a la Web, dando un énfasis en la red social Twitter para luego conceptualizar los procesos de Data Mining y Opinion Mining.

Se incluyen dentro de este capítulo las métricas de evaluación que se ocuparan en capítulos posteriores.

2.1 Web 2.0: Tecnologías y servicios generados

La “World Wide Web”, comúnmente conocida como “Web”, “www” o “w3” es una gran colección de documentos interconectados, disponibles a través de internet. Comenzó como un proyecto de información interconectada en CERN, durante los 90’s cuando se definió un cuerpo de software, hardware y un grupo de protocolos.

La Web 2.0 es la segunda fase en la evolución de la Web, conocida como la web de comunicación, web participativa, la web centrada en las personas, la web de lectura-escritura. Es una colección de tecnologías, estrategias de negocios y tendencias sociales.

A continuación se muestra la clasificación de [4] sobre los servicios generados gracias a las características de la Web 2.0:

Blogs: el término Web Log o simplemente blog fue propuesto en 1997 por Jorn Barger. Es un sitio web donde las personas exponen sus ideas, pensamientos, comentarios. Las entradas de los blog o post están basados en contenido propio expuesto en forma de revista o bitácora, mostrados usualmente en orden cronológico inverso. [5]

Las entradas de un blog pueden ser categorizadas según sus palabras claves para poder ordenar los contenidos de cada uno. Por ejemplo, cuando las entradas se vuelven antiguas, éstas pueden ser filtradas según el tema que tenga el menú de la página [6]. Utilizar links en los blogs es un aspecto importante, pues enlazar a otros blogs o páginas web amplía la información, permite citar fuentes, o bien, continuar un tema de otro blog.

RSS o Really Simple Syndication: es un formato XML utilizado para resumir la información y links de recursos para compartir el contenido de blogs o páginas web. Informa a usuarios sobre actualizaciones de blogs o sitios web en los que tienen interés.

Wikis: una wiki es una página web (o un conjunto de páginas web) que son fácilmente editables para cualquiera que acceda a ellas. A diferencia de los blogs, las versiones previas de las wikis pueden ser buscadas en su historial, además estas versiones pueden

ser restauradas. Algunas de las características de las Wikis son: su lenguaje, la estructura simple de su sitio y navegación, diseño simple, soporta múltiples usuarios, construido en la búsqueda y su simple seguimiento

Comunidades (contenido): dentro de los sitios web encontramos algunos que están especialmente organizados sobre un contenido particular, donde se comparten los contenidos. Las comunidades pueden compartir videos, fotos, enciclopedias públicas o social bookmarking.

Foros / Bulleting boards: sitios para intercambiar ideas e información comúnmente sobre intereses en específico [4].

Redes sociales: los servicios de redes sociales son esencialmente un grupo online de aplicaciones que conectan personas a través de la información sobre intereses que comparten. Éstos permiten a los usuarios tener enlazados amigos mutuos o conocidos, construir perfiles y actualizar libretas de direcciones. Algunos sitios entregan herramientas sociales para construir comunidades con el propósito de facilitar reuniones cara-a-cara a lo largo de ciudades alrededor del mundo [4].

2.1.1 Twitter

Twitter es un sitio web de microblogging nacido el 2006 [7], se ha vuelto uno de los sitios más populares mundialmente, tiene 316 millones de usuarios activos mensuales, 500 millones de tweets enviados por día, y se encuentra disponible en más de 35 idiomas compatibles [8].

Millones de usuarios comparten diferentes opiniones, comentarios por esta red cada día, en distintos ámbitos, así como también la audiencia de la plataforma varía desde usuarios regulares a celebridades, representantes de compañías, políticos, hasta presidentes de diversos países.

Dentro de esta red se permite enviar mensajes que contienen texto, con un máximo de 140 caracteres. Esta red ha contribuido a generar un lenguaje propio de las acciones y usuarios del sitio:

- **Tweet:** corresponde al mensaje que un usuario de la red puede escribir, comentar.
- **Seguidores o "followers":** la acción de un usuario en suscribirse a los tweets de otro usuario se llama "seguir", y por consiguiente el usuario que se ha suscrito a otro es llamado "seguidor" o "follower"

Actualmente la plataforma puede ser utilizada desde el sitio web o desde dispositivos inteligentes (smartphones). Además se pueden realizar mensajes o "tweets" desde el servicio de mensajería cortos (SMS) disponible en algunos países.

2.1.1.1 Comentarios en Twitter

Dentro de Twitter se tienen algunas convenciones propias en el lenguaje utilizado. A continuación se presentan algunos ejemplos de las convenciones de Twitter:

1. “RT” es el acrónimo de retweet, se coloca al frente de un comentario para indicar que el usuario está repitiendo y volviendo a comentar algún tweet que fue realizado por otro usuario.

2. “#” conocido como hashtag, es utilizado para marcar, organizar o filtrar tweets dependiendo del tema o categoría. Este tipo de términos es muy común dentro de las redes sociales para mostrar énfasis en el tema de discusión.

3. “@usuario1” representa que el mensaje es una respuesta al usuario cuyo nombre es “usuario1”.

4. Emoticones y expresiones coloquiales usadas frecuentemente en tweets, por ejemplo “:D”, “tqm”.

5. Links web externos (ejemplo: http://rbb.cl/cqa8) que se encuentran para referirse a fuentes externas, algunos de estos links pueden estar asociados a fotografías.

6. Largo: los tweets están limitados a 140 caracteres.

Es conveniente utilizar este tipo de comentarios para investigación ya que existe un gran número de mensajes, con gran cantidad de opiniones y obtenerlos ocupa una técnica bastante simple en comparación a la extracción de comentarios desde blogs en la Web [9].

2.2 Knowledge Discovery y Data Mining

Knowledge Discovery in Databases (KDD) o proceso de extracción de conocimiento es un proceso organizado para identificar grandes patrones y grupos complejos de sets [10].

El proceso empieza determinando los objetivos y termina con una implementación del conocimiento descubierto. Como resultado los cambios deben ser realizados dentro del dominio de aplicación. El núcleo del proceso KDD es el Data Mining, que involucra algoritmos de inferencia para explorar los datos, desarrollar el modelo y descubrir patrones previos desconocidos. El modelo es utilizado para entender fenómenos de los datos, analizar y predecir.

A continuación hay una breve descripción de las etapas del proceso de extracción de conocimiento caracterizados en [10]:

1. **Desarrollo y entendimiento del dominio de la aplicación**: en esta etapa se hace necesario entender y definir los objetivos que tendrá la realización de este proceso. Se entiende qué debería realizarse (sobre transformación, algoritmos, representación, entre otros) y se define el ámbito que el proceso de descubrimiento abarcará.

2. **Selección y creación del grupo de datos donde se trabajará**: luego de haber definido los objetivos, los datos que serán utilizados deben ser determinados. Esto incluye la búsqueda de los datos disponibles, la obtención de datos adicionales y luego la integración de todos los datos que se utilizarán en un solo conjunto de datos.

3. **Pre-procesamiento y limpieza:** en esta etapa se mejora la exactitud de los datos. Acá se incluye la limpieza de los datos en cuanto al manejo de los datos perdidos y la remoción de outliers.

4. **Transformación de los datos:** se realiza la generación de mejores datos para su análisis. En esta etapa preliminar se pueden incluir métodos de reducción de la dimensión, transformación de atributos. Es una etapa crucial para el proceso, pero es bastante específico a cada proyecto.

5. **Elección de la tarea apropiada de Data Mining**: de acuerdo a los objetivos del proceso de extracción de conocimiento, se decide qué tipo de Data Mining se utilizará, pueden ser por ejemplo: regresiones, métodos estocásticos o agrupamiento o $clustering$.

6. **Elección del algoritmo de Data Mining**: teniendo la estrategia, se elige la táctica, acá se selecciona el método específico que se utilizará para el reconocimiento de patrones. Por ejemplo, la elección de redes neuronales o arboles de decisión.

7. **Implementación del algoritmo de Data Mining**: luego de la elección del algoritmo, finalmente es implementado. En este paso puede ser necesario emplear el algoritmo varias veces hasta tener un resultado satisfactorio.

8. **Evaluación**: en esta etapa se realiza la evaluación e interpretación de los patrones analizados, éstos pueden ser reglas, confiabilidad, entre otros, con los respectivos objetivos definidos en el primer paso. También se realiza la documentación del conocimiento encontrado para futuros usos.

9. **Utilización del conocimiento descubierto:** ahora se puede incorporar el conocimiento dentro de otro sistema para próximas acciones. El conocimiento se vuelve activo en el sentido en que puede generar cambios al sistema y medir los efectos. En realidad, el éxito de esta etapa determina la efectividad del proceso completo de KKD.

A continuación en la Figura 1 se puede ver gráficamente un esquema de los pasos que componen el proceso de extracción de conocimiento a partir de bases de datos:

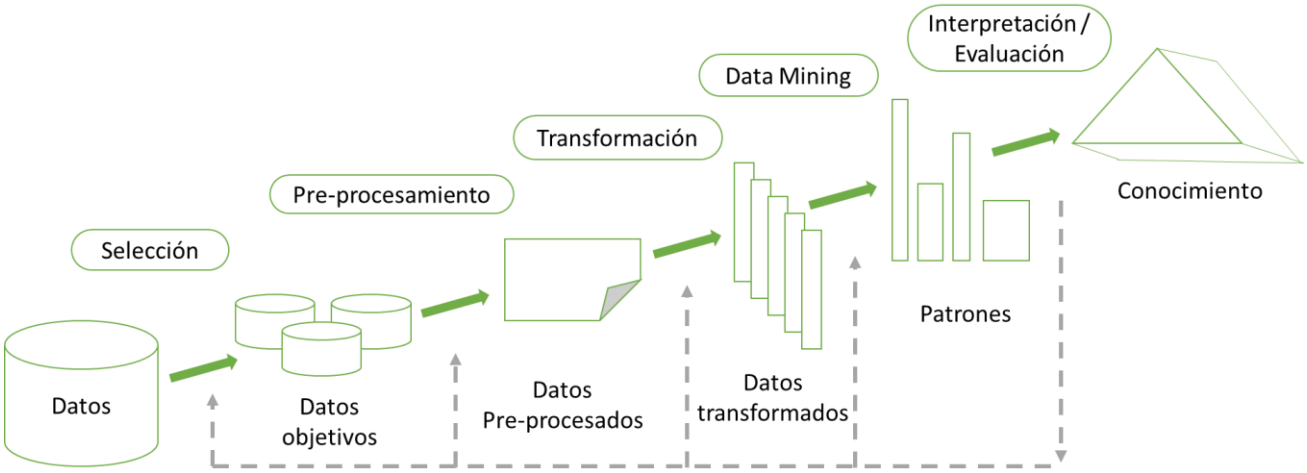


FIGURA 1: PROCESO KDD
FUENTE: ELABORACIÓN PROPIA

2.2.1 Data mining

Data mining, o minería de datos, es un campo de las ciencias de computación que se refiere al proceso de descubrir patrones en grandes volúmenes de datos [10]. Tiene como base los métodos: estadísticos, Inteligencia artificial (IA) que incluye Machine Learning y sistemas de bases de datos. El análisis de la información viene dado por el estudio de la información implícita que presentan los datos, previamente desconocidos y que pueden ser útiles para el usuario.

Dentro de las aplicaciones que tenemos de data mining, se puede hablar de análisis de sentimiento al estudio computacional de opiniones, evaluaciones, sentimientos, actitudes, emociones, subjetividad expresada en texto. Esto puede ser en reviews, blogs, comentarios, noticias, feedback, discusiones. Un ejemplo de esto es identificar preferencias de clientes para productos de turismo como en [11].

El uso típico que se le da a este tipo de análisis es extraer desde textos cómo se sienten las personas frente a diferentes productos, también puede ser utilizado para conocer el impacto de servicios, en áreas de salud, de finanzas, se pueden incluir eventos sociales y elecciones políticas.

Tenemos por ejemplo modelos de sentimiento utilizados para predecir las ventas, reviews en la web son utilizadas para categorizar productos y mercancías [12]. Se han estudiado los comentarios de Twitter con encuestas públicas de opinión. Predecir las elecciones también ha sido un área de estudio, así como también los comentarios de películas de cine, han sido utilizados para predecir los ingresos por películas. Vemos también como se ha estudiado en base al análisis de sentimientos las relaciones sociales.

Los algoritmos utilizados pueden ser clasificados en diversas categorías según los resultados esperados del análisis, entre ellos encontramos 2 importantes:

Supervisados: estos algoritmos generan una función matemática que mapea entradas en un set de salidas deseadas.

No supervisados: estos algoritmos no se preocupan de mapear las entradas en grupo de salidas conocidas, sino que simplemente modelar un grupo de entradas.

2.2.2 Text mining

Text mining o análisis de texto corresponde al descubrimiento de conocimiento a partir de grandes documentos de texto, este proceso también es conocido como “Knowledge Discovery from Text (KDT)”.

Text Mining aplica las mismas funcionalidades estadísticas de Data Mining pero también aplica análisis de Procesamiento Natural del Lenguaje (o PNL en sus siglas en inglés de “Natural Languaje Proccesing”) y técnicas de “Information Retrieval” [10].

Las herramientas de Text Mining son usadas para:

- Extraer información relevante desde un documento.
- Encontrar tendencias o relaciones entre personas, lugares, organizaciones, etc. Agregando y comparando información extraída desde los documentos.
- Clasificar y organizar documentos de acuerdo a su contenido.
- Documentos basados en varios tipos de información sobre el contenido del documento.
- Agrupar documentos de acuerdo a su contenido.

El sistema de Text Mining está compuesto por 3 grandes componentes: [10]

- **Extracción de información:** permite la conexión entre varias colecciones de texto y módulos de etiquetado. Este componente conecta a cualquier sitio web, fuente (como *news feed*), colecciones de documentos y cualquier tipo de colección de texto.
- **Etiquetado inteligente:** corresponde a la información relevante. Este componente puede ocupar cualquier tipo de etiquetado en documentos como etiquetado estadístico, etiquetado semántico y etiquetado estructural (extracción desde el diseño visual de los documentos).
- **Inteligencia de negocios:** componente que consolida la información desde diferentes fuentes, permitiendo análisis simultáneos dentro del espacio de información.

2.3 Opinion mining

La minería de opiniones u “opinión mining”, conocido también como análisis de sentimientos se refiere a la aplicación de técnicas de diferentes campos como lo son el procesamiento de lenguajes naturales “PLN”, lingüística computacional y análisis de textos, para identificar y extraer información subjetiva de un conjunto de datos que contengan texto.

Es necesario destacar las tareas más importantes de este análisis que son presentadas en [13]:

Clasificación de polaridad: se utiliza para detectar emociones dentro de opiniones en textos, identificando si una opinión tiene una connotación positiva o negativa con respecto al tema del cual se está hablando.

Clasificación de subjetividad: se utiliza para distinguir entre información objetiva y opiniones dentro de textos. Muchos de los trabajos en clasificación de polaridad se basan en que los documentos sobre los que trabajan contienen opiniones. Para muchas aplicaciones es necesario distinguir si un documento contiene información subjetiva o no, identificar qué porciones del documento contienen información subjetiva.

Análisis conjunto de temas/sentimiento: una simplificación realizada por la clasificación de sentimientos a nivel de documento es que cada documento se focaliza en el tema del cual se está interesado. Es posible que se encuentren interacciones entre los distintos tópicos y opiniones dentro del documento que se debieran considerar.

Puntos de vista y perspectivas: en general, muchos de los trabajos que están orientados en política se centran en las posturas generales que expresan a través de los textos que no están necesariamente focalizados en un tema en particular.

Otra información no objetiva en el texto: algunas investigaciones han considerado varios aspectos, como por ejemplo algunas emociones universales como: alegría, repulsión, enojo, miedo, sorpresa, tristeza. También se han estudiados problemas como la detección de lenguaje engañoso en el ámbito de la inteligencia y configuración de seguridad.

La clasificación manual de comentarios para la minería de opiniones es una tarea con una escala de esfuerzo inviable en escala humana, por lo que diversos métodos han sido propuestos para inferir automáticamente la opinión humana desde diferentes textos de lenguaje natural. Dada la subjetividad que presentan los datos, este tema aún es un problema abierto y existen esfuerzos en aumentar la investigación en este campo.

Además, existen diversas limitaciones para evaluar la opinión pública utilizando los diferentes métodos de minería de opinión aplicadas a las redes sociales. Una de las principales limitaciones es la población que utiliza estas plataformas, porque éstas no son necesariamente una muestra representativa de la población ya que reflejan la opinión de una fracción particular de la población.

Sin embargo, hay algunas ventajas en cuanto a que el procesamiento de grandes cantidades de datos puesto que ésta puede realizarse de forma automática, aumentando la cantidad de opiniones procesadas y cómo se van produciendo a través del tiempo, puede estudiarse la opinión pública a través del tiempo, en cuanto a tendencias, estacionalidad y volatilidad.

La literatura propone diversas técnicas en relación a Opinion Mining, los enfoques más conocidos son:

- **Aspect-Based Opinion Mining**: se dividen las entradas de texto en aspectos, también llamados tópicos o features. Estos aspectos son usualmente tópicos arbitrarios que son considerados relevantes o representativos del texto que es analizado. Este enfoque es bastante popular y variados autores han propuestos sus propios modelos y perspectivas.

- **Non-Aspect-Based Opinion Mining**: este enfoque agrupa todas las técnicas que no dividen el texto en sub-tópicos. En general estas técnicas consideran el texto como un gran objeto o bien, incrementan el análisis de granularidad por cada párrafo, oración o frase.

2.3.1 Aplicaciones

Dentro de la minería de opiniones o análisis de sentimientos se pueden encontrar innumerables aplicaciones dentro de diversos ámbitos, a continuación se muestran algunas de ellas de acuerdo a la clasificación propuesta en [13]

Aplicaciones a sitios relacionados con reseñas o reviews

Analizar las opiniones que realizan las personas sobre un producto, marca, objeto, persona en particular es un problema importante. Existen sitios donde se da espacio a las personas para valorizar en una escala la opinión que se tiene sobre un producto en particular, un ejemplo es el sitio Epinions.com que reúne variados comentarios sobre productos que van desde cámaras digitales, electrodomésticos, películas, computación, música, entre otros. Una alternativa a estos sitios pueden ser sitios que proactivamente reúnan esta información y no sea necesario pedir a los usuarios sus comentarios y opiniones con una escala en particular, donde los temas no estén restringidos a productos, sino que incluyan candidatos a cargos, asuntos políticos, entre otros.

Aplicaciones como un sub-componente de tecnología

Se tiene un potencial importante en cuanto a la habilitación de tecnologías para otros sistemas. Una posibilidad es mejorar los sistemas de recomendaciones, por ejemplo no recomendar ítems que reciban gran cantidad de comentarios negativos. Otra es la detección de lenguaje inapropiado en email u otro tipo de comunicación es un posible uso para la detección y clasificación de subjetividad.

Aplicaciones en negocios e inteligencia gubernamental

El campo de la minería de opinión se adecúa a varios tipos de aplicaciones de inteligencia, en efecto, la inteligencia de negocios parece ser uno de los factores fundamentales detrás del interés corporativo en esta materia.

Aplicaciones en diversos dominios.

Se sabe que las opiniones son una materia con la cual los políticos tienen que lidiar. Algunos trabajos se enfocan en entender lo que los votantes están pensando, si los proyectos están favoreciendo las posiciones de los políticos, como que figuras públicas pueden apoyar u oponerse a la causa, la calidad de información a la que los votantes tienen acceso.

2.3.2 Modelo

En esta sección se explica el modelo de Bing Liu [14] , que han servido como motivación y acercamientos en esta materia. Se presenta a continuación una visión general del modelo más utilizado, y la base de las investigaciones para el proyecto “OpinionZoom”.

Sea d un documento de opinión (por ejemplo el comentario sobre un producto en particular) compuesto por una lista de afirmaciones $s_1, \dots, s_n$. Como se presenta en [14], los componentes básicos de una opinión en d son:

- **Entidad:** puede ser un producto, persona, evento, organización o tópico en el cual la opinión es expresada. Una entidad es compuesta por una jerarquía de componentes y sub-componentes que pueden tener un conjunto de atributos. Por ejemplo un celular está compuesto por una pantalla, batería con sus componentes, donde los atributos pueden ser el tamaño y el peso. Por simplificación los componentes y atributos son nombrados como **aspectos.**

- **Titular de opinión**: la persona u organización que sostiene una opinión específica de una entidad particular. Mientras que en blogs, los titulares son usualmente los autores de los documentos, en los artículos de noticias son indicados explícitamente.

- **Opinión:** una visión, valoración de un objeto de una titular de opinión. Una opinión puede ser de orientación positiva, negativa o neutral, donde comúnmente es interpretada como no opinión. La orientación es conocida como orientación de sentimiento, orientación semántica o **polaridad**.

Teniendo en cuenta los componentes presentados anteriormente, se puede definir una opinión como una quíntupla (e_i,a_{ij},oo_{ijkl},h_k, t_l) donde e_i es la entidad, a_{ij} es un aspecto de e_i y oo_{ijkl} es la orientación de la opinión a_{ij} expresado por el titular de h_k en el periodo t_l . Los valores posibles para oo_{ijkl} son las categorías positivo, negativo y neutro o bien, diferentes niveles de fuerza o intensidad [15].

2.3.3 Niveles de análisis

A continuación se presenta la clasificación que expone Liu en [16] mostrando los diferentes niveles de análisis según su nivel de granularidad y principales problemas de investigación:

Nivel de documento: en este nivel, el análisis de la minería de opiniones se centra en clasificar los documentos en positivos o negativos, también conocida como clasificación de sentimientos. En este nivel de análisis se asume que cada documento expresa opiniones sobre una sola entidad (por ejemplo, un solo producto), por tanto no es aplicable a documentos que evalúan o comparan múltiples entidades. La aplicación de este nivel de análisis es limitado ya que reside en el contexto de la crítica analizada. [17]

Nivel de oración: este nivel es análogo al anterior considerando una oración como un documento corto. Sin embargo, presenta un paso adicional donde se separa el documento en oraciones. La tarea en este nivel es determinar en cada oración si se expresa una opinión positiva, negativa o neutral. Usualmente una opinión neutral no corresponde a una opinión. Este nivel de análisis está altamente relacionado con la clasificación de subjetividad donde se distinguen oraciones con información objetiva de oraciones con opiniones u oraciones subjetivas. Es necesario notar que subjetividad no es equivalente a sentimiento, ya que gran cantidad de oraciones objetivas pueden tener opiniones. [16] [17]

Nivel de entidad y aspectos: este nivel representa el nivel más pequeño donde la minería de opiniones es utilizada. Aquí, la tarea no es solo encontrar la polaridad de la opinión, sino que su objetivo (entidad, aspecto o ambos), por tanto la definición de la quíntupla para opiniones en la sección 2.3.2 aplica totalmente. En los niveles de documento y oración el análisis no descubre exactamente lo que a las personas les gusta o no, en el nivel de aspecto se realiza un análisis más fino. Este nivel de aspecto fue llamado nivel de características en [18]. A diferencia de buscar en la estructura del lenguaje (documentos, párrafos, oraciones o frases), el nivel de aspecto se centra en buscar una opinión por si sola. Se basa en la idea en que una opinión consiste en un sentimiento (positivo o negativo) y un objetivo (de opinión). Una opinión sin un objetivo siendo identificado es de uso limitado. Conociendo la importancia del objetivo de la opinión también ayuda a entender el problema del análisis de sentimientos mejor. [16] [17]

Además existen 2 tipos de opiniones: opiniones regulares y comparativas. Una opinión regular expresa un sentimiento sólo a una entidad particular o un aspecto de la entidad, por ejemplo “La Coca-Cola sabe bien “, que expresa un sentimiento positivo sobre el aspecto del sabor en la Coca-Cola. Una opinión comparativa ofrece una comparación entre múltiples entidades basados en algunos de sus aspectos en común, por ejemplo “La Coca-Cola sabe mejor que la Pepsi”, donde se compara la Coca-Cola y la Pepsi basados en su sabor (aspecto) expresando una preferencia por la Coca-Cola. [16]

2.3.4 Etapas del proceso de Opinion Mining

Los algoritmos para el procesamiento de texto en general abarcan las siguientes etapas del proceso de Opinion Mining:

1. **Recolección de datos:** en esta primera etapa se realiza la recolección de datos o corpus que consiste en la obtención del corpus que será analizado por opiniones. Básicamente existen 2 formas de realizarlo para la recolección de datos de la web, una es a través de la API (interfaz de programación de aplicaciones) del sitio o a través de un “web crawler” que rastrea las páginas web en busca de información relevante.

2. **Pre-procesamiento de texto:** en una segunda etapa se realiza el pre-procesamiento de texto para realizar el análisis correspondiente. Esta tarea tiene asociado el procesamiento de lenguajes naturales y análisis léxico.

3. **Núcleo del proceso de Opinion Mining:** se tienen principales enfoques para el análisis de opiniones, uno corresponde al enfoque no supervisado basado en un lexicón (unsupervised lexicon-based) donde el proceso se basa en reglas heurísticas obtenidas del conocimiento lingüístico, y el otro enfoque corresponde a técnicas de aprendizaje automático supervisado (supervised machine learning).

4. **Agregación y síntesis de los resultados:** el objetivo de esta etapa es agregar y representar el análisis desde los análisis individuales de opinión a un análisis global. Se ayuda al usuario a entender de una forma simple la variedad de opiniones que presentan los textos. El enfoque tradicional de sintetizar los resultados es construir nuevas oraciones desde los documentos con opiniones con el objetivo de extraer los principales tópicos y capturar la esencia de las opiniones: su objetivo y sentimientos asociados.

5. **Visualización:** se finaliza el proceso de Opinion Mining con la visualización de los resultados. Esta etapa depende del objetivo de análisis de los datos y su usabilidad para ser presentada de forma acorde a los objetivos.

Se detallarán a continuación los 3 primeros pasos del proceso de Opinion Mining, ya que son las etapas relevantes para el objetivo del trabajo que se realiza.

2.3.5 Pre-procesamiento del texto

El segundo paso, luego de la adquisición de los datos, es su pre-procesamiento. Dentro de las técnicas más utilizadas están:

- **Tokenización:** tarea de separar el texto completo en una lista de palabras separadas. Esto es simple para lenguajes que están delimitados por espacios como lo son el inglés, español o francés, pero se torna más difícil en lenguajes en que las palabras no están delimitados por espacios como los son el japonés y chino.

- **Identificación de la raíz:** proceso heurístico en que se eliminan los sufijos, prefijos, dejando la raíz de la palabra. Por ejemplo, persona, personificación, personas se transforman a persona cuando se ha sacado la raíz.

- **Lematización:** algoritmo que lleva la palabra a su forma no-flexionada. Es análogo a la identificación de la raíz, pero más rigurosa porque incorpora un análisis morfológico para cada palabra.

- **Eliminación de palabras no utilizadas:** tarea en la cual se eliminan las palabras que son utilizadas para estructurar el lenguaje pero que no contribuyen en el contenido. Ejemplo de estas palabras son: una, unas, el.

- **Segmentación de oraciones:** proceso en el cual se separan los párrafos en oraciones. Este paso presenta sus propias dificultados puesto que los puntos son utilizados para terminar una oración así como también para abreviaciones y números decimales.

- **Etiquetado de palabras:** en inglés “Part-of-Speech (POS) Tagging”, es la etapa en que se etiqueta cada palabra de una oración según su categoría gramatical como adjetivo, sustantivo, verbo, adverbio o preposición.

Notar que no todos los pasos descritos son utilizados para cada aplicación de minería de opiniones, pero se han presentado los más relevantes y comunes dentro de los algoritmos que se utilizan [19].

2.3.6 Enfoques para el análisis de opinión

Existen 2 enfoques establecidos en la literatura para la realización del núcleo del proceso de Opinon Mining, que son el enfoque no-supervisado basado en un lexicón y un enfoque supervisado de aprendizaje automático. Existen también estudios que utilizan ambas técnicas combinadas obteniendo buenos resultados, existen además nuevos acercamientos en el área con métodos en que se utiliza la ontología dentro del problema de Opinion Mining, este tipo de enfoque se conoce como análisis de opinión basado en los conceptos (en inglés, concept-based Opinion Mining). A continuación se presenta la clasificación realizada por [17] en cuanto a los 3 enfoques actuales dentro del Análisis de Opiniones:

Enfoque no supervisado basado en un lexicón (en inglés, unsupervised lexicon-based approaches): también es llamado enfoque semántico, o en inglés, semantic-based approaches. Intenta determinar la polaridad de un texto utilizando reglas y heurísticas obtenidas del conocimiento lingüístico. En general este proceso se inicia etiquetando cada frase o palabra con su correspondiente polaridad con ayuda de un lexicón, luego, se incorporan al análisis palabras de cambio y su alcance (intensificadores y negaciones), finalmente se manejan las conjunciones adversativas (“mas”, “pero”, “aunque”).

Enfoque supervisado de aprendizaje automático (en inglés, supervised machine learning): también es conocido como enfoque basado en el aprendizaje, en inglés, supervised learning-based approaches, o como métodos estadísticos para la clasificación de sentimientos.

Enfoque de análisis de opinión basado en conceptos (en inglés, concept-based approaches): este enfoque es relativamente nuevo y consiste en utilizar ontologías que ayuden las tareas de la minería de opinión. Se entiende por ontología como un modelo que conceptualiza el conocimiento de un determinado dominio de forma que las personas y los computadores entiendan. Usualmente, las ontologías son presentadas como grafos que mapean nodos enlazados identificando relaciones.

En particular el enfoque ocupado en el proyecto “OpinionZoom” es el enfoque basado en el lexicón que depende de palabras de opinión (o sentimiento), palabras que expresan sentimientos positivos o negativos. Estas palabras codifican un estado deseable, por ejemplo, “excelente” y “bueno" tienen una polaridad positiva, mientras que las palabras que codifican un estado no deseado tienen una polaridad negativa, por ejemplo, “malo" y “horrible". Aunque la polaridad normalmente se aplica a los adjetivos y adverbios, hay también verbos y sustantivos como palabras de opinión [20].

2.4 Métricas de evaluación

La evaluación más frecuente de algoritmos de clasificación son las medidas de $Precisión$ y $Exhaustividad$ ($Precision$ y $Recall$ en inglés respectivamente). Para la evaluación, los resultados se presentan como una matriz de confusión que ayuda al análisis de la clasificación realizada. Esta matriz para 2 clases se presenta en [21] como:

		Clases predichas	
		Positivo	**Negativo**
Clases reales	**Positivo**	VP	FP
	Negativo	FN	VN

TABLA 1: MATRIZ DE CONFUSIÓN
FUENTE: ELABORACIÓN PROPIA

Teniendo en consideración que se presentan 2 clases, los datos pueden ser agrupados en 4 conceptos detallados a continuación:

- $Verdadero\ positivo\ (VP)$ = Cantidad de clasificaciones correctas para el caso de datos positivos.
- $Verdadero\ negativo\ (VN)$ = Cantidad de clasificaciones correctas para el caso de datos negativos.
- $Falso\ positivo\ (FP)$ = Cantidad de clasificaciones incorrectas para el caso de datos positivos.
- $Falso\ negativo\ (FN)$ = Cantidad de clasificaciones incorrectas para el caso de datos negativos.

Al considerar estos conceptos, podemos calcular las métricas de desempeño relevantes para la clasificación realizada, ocupando las medidas de precisión y exhaustividad [22].

$Precisión\ (Precision)$: Corresponde a la proporción de casos de una clase predicha que eran efectivamente de la clase. Se puede calcular por tanto la precisión para la clase positiva, como para la clase negativa como sigue:

$$P_{pos} = \frac{VP}{VP + FP} \quad \text{y,} \quad P_{neg} = \frac{VN}{VN + FN}$$

$Exhaustividad\ (Recall)$: Corresponde a la proporción de casos reales de la clase que fueron efectivamente clasificados como tal. Se puede calcular la exhaustividad para las distintas clases como sigue:

$$R_{pos} = \frac{VP}{VP + FN} \quad \text{y,} \quad R_{neg} = \frac{VN}{VN + FP}$$

Otra medida que refleja el desempeño de un clasificador, es la medida de $Exactitud, o\ Accuracy$ en inglés, que corresponde al porcentaje de datos clasificados correctamente por el clasificador.

$$A = \frac{VP + VN}{VP + FP + VN + FN}$$

Si se analizan las métricas de precisión y exhaustividad podemos notar que existe un trade-off entre ambas métricas, si queremos aumentar la exhaustividad se pueden rescatar, por ejemplo, todos los posibles datos para todas las consultas, mientras que si se aumenta mucho la cantidad de datos la precisión puede disminuir considerablemente. Una medida simple ocupada para analizar la compensación entre ambas métricas es el estadístico F. definido como la media armónica ponderada entre $Precisión$ y $Exhaustividad$.

$$F = \frac{(\beta + 1)PR}{\beta^2 P + R}$$

Valores de $\beta < 1$ enfatizan $precisión$, mientras que valores $\beta > 1$ enfatizan la $exhaustividad$. Si se considera que ambas medidas tienen el mismo peso, el estadístico F queda como sigue:

$$F = \frac{2PR}{P + R}$$

Para determinar el grado de concordancia que tienen 2 observadores una métrica bastante utilizada y que representa un ajuste producto de la aleatoriedad es el coeficiente $Kappa$ que muestra el grado de concordancia que tienen 2 observadores. El cálculo se basa en la diferencia entre el acuerdo presentado o “acuerdo observado” comparado con el “acuerdo esperado” o acuerdo presente gracias al azar [23].

El $coeficiente\ kappa$ es calculado de la siguiente forma:

$$\kappa = \frac{P_O - P_E}{1 - P_E}$$

Donde:

$P_o = \ concordancias\ observadas$

$P_E = \ concordancias\ atribuibles\ al\ azar$

El rango de valores de $Kappa$ va desde $-1\ a\ 1$, donde 0 corresponde a un acuerdo solamente atribuible al azar, mientras que 1 considera un acuerdo perfecto. Si se tiene un valor de -1 corresponde a un perfecto desacuerdo, atribuible en algunos casos a errores sistemáticos.

A continuación en la **¡Error! No se encuentra el origen de la referencia.** se presenta la nterpretación del coeficiente kappa según [24].

Kappa	Fuerza de concordancia
0.00	Pobre
0.01 – 0.20	Leve
0.21 – 0.40	Aceptable
0.41 – 0.60	Moderada
0.61 – 0.80	Considerable
0.81 – 1.00	Casi perfecta

TABLA 2: INTERPRETACIÓN KAPPA
FUENTE: ELABORACIÓN PROPIA SEGÚN [24]

Capítulo 3

3 Generación lexicones

Muchas de las soluciones que se emplean para la detección de textos subjetivos, clasificación de polaridad de opiniones (positiva, negativa), o extracción de opiniones individuales, se apoyan en lexicones de opinión o lexicones de polaridad (opinion lexicon o sentiment lexicon en inglés).

Para el problema de generación de lexicones existen 3 grandes enfoques para abordarlo, que se detallan en este capítulo y algunos ejemplos de lexicones construidos.

3.1 Definición

Entendemos por un lexicón de opinión a un recurso lingüístico que contiene la orientación semántica de una palabra, consistente en un valor numérico que identifica la polaridad, ya sea positiva o negativa.

La importancia en la utilización de lexicones de opinión en el análisis de opiniones, es que ayuda a mejorar la exhaustividad ("Recall") en la identificación de las expresiones de opinión y que con ayuda de reglas lingüísticas pueden generarse nuevas expresiones de opinión. [25]

Dentro del problema de la identificación de polaridad (orientación semántica) en términos y palabras, los lexicones de opinión tienen un papel fundamental documentando algunos términos ya conocidos con sus orientaciones semánticas en determinados dominios.

Finalmente entenderemos dentro de este trabajo a un lexicón de opinión como un listado, diccionario, repositorio de palabras en el que se le asocie un grado de polaridad. Un ejemplo de lexicón de opinión se presenta en la Tabla 3 que se muestra a continuación:

Término	Polaridad
Excelente	5.4
Adorable	2.6
Castigo	-1
Pobre	-6

TABLA 3: EJEMPLO LEXICÓN
FUENTE: ELABORACIÓN PROPIA

3.2 Enfoque manual

Esta forma manualmente se realiza una lista de palabras con orientaciones conocidas, identificando la polaridad de cada palabra. Esta alternativa requiere una gran cantidad de tiempo invertido, por lo que suelen estar combinadas con otros métodos automáticos de generación.

Algunos de los lexicones realizados con éste método son:

- **Bing Liu's Opinion Lexicon:** contiene una lista de palabras de opinión positivas y negativas en inglés, alrededor de 6.800 palabras. Fue construida inicialmente de forma automática, pero esta lista de palabras ha sido recopilada por varios años desde 2004. Está formada por palabras flexionadas, que incluyen faltas de ortografía y slangs (expresiones informales comúnmente encontradas en internet). Este es uno de los lexicones más utilizados, ya que presenta un gran número de citas.

- **MPQA Subjetivity Lexicon:** MPQA es la abreviación de Multi-Perspective Question Anwering. Este lexicón cuenta con cerca de 8.000 palabras singulares subjetivas, donde cada una de ellas es clasificada como positiva o negativa. Este lexicón es mantenido por Theresa Wilson, Janyce Wiebe y Paul Hoffman.

VENTAJAS: este enfoque permite clasificar palabras con una baja cantidad de error.

DESVENTAJAS: requiere una alta cantidad de tiempo y personal capacitado con gran conocimiento del idioma. Para caracterizar según el dominio es necesario además que se esté siendo capacitado dentro del dominio en que se desea aplicar. El conocimiento de palabras de un idioma en particular de una persona es menor a la cantidad existente palabras, dejando fuera del lexicón palabras que no son utilizadas.

3.3 Enfoque basado en diccionario

Este método se inicia con una lista de palabras con su polaridad conocida, típicamente se utiliza bootstrap para identificar sinónimos, antónimos; para idioma inglés existe WordNet, base de datos que agrupa las palabras en sinónimos, almacenando las relaciones semánticas entre los conjuntos de sinónimos.

- **Kamps** en [26] el 2004 propone un método basado en las distancias en WordNet que determina la orientación de sentimientos de un adjetivo dado.

  La distancia $d(t_1, t_2)$ entre los términos t_1 y t_2 es el largo de la parte más corta que conecta t_1 y t_2 en WordNet.

  La orientación de un adjetivo t es determinado por su distancia relativa entre 2 referencias (o semillas) de términos bueno y malo, i.e.,

  $$SO(t) = \frac{d(t, bad)) - d(t, good)}{d(good, bad)}$$

  t es positivo ssi $SO(t) > 0$, y negativo en el caso contrario. El valor absoluto de $SO(t)$ muestra la fuerza del sentimiento.

- En [27] utilizan un método semi-supervisado donde entregan 2 grupos de palabras, positivas y negativas, que luego son expandidas en un diccionario online de relaciones de sinónimos y antónimos para generar 2 nuevos grupos expandidos que pasan a ser los grupos de entrenamiento utilizando finalmente los significados dentro de un diccionario de términos.

- Luego, en [28] se realiza un boostraping básico con grupos iniciales de adjetivos dados. Primeramente, se comienza expandiendo los grupos con los sinónimos, antónimos relaciones especificadas en WordNet. Luego se expande por el diccionario de significado e identifica las entidades que contienen en sus definiciones las palabras de sentimiento de los grupos extendidos, añadiendo la entrada a la correspondiente categoría.

  Finalmente el método de [28] es limpiado utilizando un etiquetado gramatical de palabras (POS tagger) para verificar que las palabras sean adjetivos y así remover contradicciones.

- **SentiWordNet** [29] es un lexicón que presenta 2 versiones 1.0 y 3.0, que ha sido referenciado por más de 300 investigaciones y disponible para propósitos de estudios en la materia. La versión 3.0 es una mejora realizada a la primera versión de este lexicón presentando grados de positividad y negatividad, incluyendo además la neutralidad de las palabras. La principal diferencia de estas versiones es el recurso utilizado, ya que este lexicón se basa en WORDNET en

un inicio en la versión 2.0, para luego ser mejorada a la versión 3.0. Este lexicón utiliza principalmente un algoritmo semi-supervisado para la generación de este lexicón.

- **ML-Senticon** [30] es un conjunto de lexicones de opinión para los idiomas: inglés, español, catalán, gallego y euskera que se presenta a nivel de lemas. Estos lexicones se han generado a partir de SentiWordNet incluyendo algunas mejoras. Se utilizó el recurso de WORDNET para la creación de este lexicón en inglés y *Multilingual Central Repository 3.0 (MCR 3.0)* para la obtención de los lexicones en otros idiomas. Este lexicón propone un modelo en capas, cuya distribución de polaridad se encuentra entre -1 y 1, siendo -1 de polaridad negativa y 1 con polaridad positiva.

VENTAJAS: la generación de un lexicón demanda una cantidad menor de tiempo que el enfoque anterior, incluye palabras que se encuentran normadas dentro del idioma.

DESVENTAJAS: al utilizar un diccionario ya establecido, se restringe a las palabras que estar normadas por el idioma, dejando de lado palabras que toman connotación según el dominio que se utilizan, este enfoque no muestra la polaridad de la palabra según el dominio en que se utiliza, palabras que corresponden a modismos u otras propias del lenguaje informal no se capturan con este método.

3.4 Enfoque basado en un corpus lingüístico

Con este enfoque, la generación de lexicón se inicia con una pequeña lista de palabras, el patrón de frecuencias de palabras puede utilizarse inicialmente para hacer crecer la lista. A menudo los algoritmos utilizan una doble propagación entre las palabras de opinión y los ítems que modifican. Requiere un corpus de tamaño grande para tener una cobertura adecuada.

Uno de los métodos más utilizados es el método de doble propagación que utiliza la dependencia de opiniones y aspectos para extraer nuevas palabras de opinión. Se basa en relaciones de dependencias en que el aspecto puede encontrar palabras de opinión que las modifique y que algunas palabras de opinión pueden ayudar a encontrar más palabras de opinión.

Algunos trabajos relevantes que se han basado en este método son:

- [31] donde se presenta una separación de adjetivos donde obtiene 2 clúster de palabras, positivas y negativas, que utiliza convenciones con los conectores: “and”, “or”, “but”, “either-or”, y “neither-nor” para lograrlo. Los autores usaron un corpus y algunos adjetivos de sentimiento semillas para encontrar adjetivos adicionales en el corpus. Su técnica explota un conjunto de reglas lingüísticas o convenciones de conectores para identificar más palabras y sus orientaciones desde el corpus.

- [32] se ocupa un algoritmo de aprendizaje no-supervisado para clasificar reseñas como recomendadas y no recomendadas. La clasificación de reseñas se realiza en base al promedio de la orientación semántica (Semantic Orientation o “SO”) de las frases en la reseña que contienen adjetivos o adverbios. La orientación semántica de una frase es calculada como:

  $$SO(frase) = PMI(frase,"excellent") - PMI(frase,"poor")$$

  Donde Pointwise Mutual Information o PMI entre 2 palabras, $word_1$ y $word_2$, es definida según [33] como:

  $$PMI(word_1, word_2) = log_2\left[\frac{p(word_1 \,\&\, word_2)}{p(word_1)p(word_2)}\right]$$

- En [34] se presenta un clasificador de oraciones de opinión ya sean positivas o negativas en términos del principal tema de opinión expresada ocupando un ratio de máxima verosimilitud modificado.

- Luego, en [35] se extendió la idea de introducir el concepto coherencia del contexto intraoracional (dentro de una oración) e interoracional (entre oraciones vecinas), llamada “context coherency”. Tendencia de que aparezca la misma polaridad sucesivamente en los mismos contextos.

- En [36] se encontraron orientaciones positivas, negativas y neutrales, asignándolas a palabras o frases, basados en la idea de ocupar palabras de opinión cercanas para determinar la orientación de opinión sobre una característica de un producto. Ocuparon no sólo adjetivos y adverbios como palabras de opinión, partieron de la base del método en [18] para ir agregando verbos y sustantivos por medio bootstrapping ocupando WordNet.

- **Hashtag Sentiment Lexicon** es un lexicón basado en comentarios de twitter que utilizan palabras que han sido marcadas con hashtag (#) que pueden indicar el tema o sentimiento del tweet. En [37] se mostró que etiquetar palabras que representen emociones como #joy #sad #angry y #surprised son buenos indicadores de que el tweet quiere expresar la misma emoción (incluso sin la palabra de emoción con hashtag).

  $$Puntaje\ de\ polaridad\ (w) = PMI\ (w, positive) - PMI\ (w, negative)$$

  Donde PMI corresponde a "pointwise mutual information"

  $$PMI(w, positive) = log_2 \frac{freq(w, positive) * N}{freq(w) * freq(positive)}$$

  Donde $freq(w, positive)$ es el número de veces que aparece un término w en tweets positivos, $freq(w)$ es la frecuencia total que aparece el término w en el corpus, $freq(positive)$ es el número total de tokens en tweets positivos y N corresponde al total de tokens dentro del corpus. $PMI(w, negative)$ es calculado de forma similar.

  Finalmente se puede simplificar la ecuación de polaridad a:

  $$Puntaje\ de\ polaridad\ (w) = log_2 \frac{freq(w, positive) * freq(negative)}{freq(w, negative) * freq(positive)}$$

  Como PMI es conocido por ser un estimador pobre para asociación de eventos con poca frecuencia, la construcción de este lexicón ignora los términos ocurridos menos de 5 veces en cada grupo de tweets (positivos y negativos).

  Este lexicón "Hashtag Sentiment Base Lexicon" tiene 39.413 unigrams y 178.851 bigrams.

- **Sentiment140 lexicon** [38]**:** este lexicón fue realizado en base al corpus "Sentiment140 Corpus" que consiste en una colección de 1.6 millones de tweets que contienen emoticones. Estos tweets están etiquetados como positivos o negativos de acuerdo al emoticón. Se generó el "Sentiment140 Base Lexicon (S140 Base)" de la misma forma que el léxico descrito anteriormente utilizando la ecuación de PMI. Se obtuvo un lexicón con 65.361 unigramas, 255.510 bigramas y 266.510 non-contiguos pairs.

VENTAJAS: puede mostrar la polaridad respecto al dominio en el que se realiza, su elaboración no requiere gran cantidad de tiempo.

DESVENTAJAS: necesita un corpus de gran tamaño, requiere mayor pre-procesamiento de textos, acotado al dominio del corpus, se necesita conocimiento de la estructura del corpus para la aplicación de algoritmos adecuados. Lexicones generados específicos al dominio de la aplicación. No abarca el total de palabras que pueden ser utilizadas para un determinado idioma, sólo las usadas en el corpus.

Finalmente se presenta la Tabla 4 que muestra las principales características de cada uno de los enfoques metodológicos presentados:

Enfoque	Manual	Diccionario	Corpus lingüístico
Tiempo de generación	Alto	Bajo	Medio
Cantidad palabras del idioma	6.000 ~ 9.000	10.000 ~ 25.000	12.000 ~ 25.000
Expresiones propias del dominio	Media	Baja	Alta
Características propias de la región	Media	Baja	Alta

TABLA 4: COMPARACIÓN ENFOQUES
FUENTE: ELABORACIÓN PROPIA

Donde:

- **Tiempo de generación:** tiempo de construcción del lexicón.
- **Cantidad palabras:** corresponde a la cantidad de palabras que presentan estos lexicones.
- **Expresiones propias del dominio:** capacidad de captar expresiones propias del dominio de aplicación del lexicón.
- **Características propias de la región:** capacidad de captar las características del idioma propias de una región en particular.

3.5 Lexicón de opinión y sus problemas

Uno de los más importantes indicadores de sentimientos son las llamadas palabras de sentimiento o “sentiment words”, también conocidas como palabras de opinión u “opinion words”. Estas palabras son utilizadas comúnmente para expresar sentimientos positivos o negativos; por ejemplo, bueno, maravilloso, excelente son palabras positivas, mientras que malo, pobre, horrible son palabras de opinión negativas. Así mismo, hay frases y modismos, como por ejemplo: “costar un ojo de la cara”. Una lista de estas palabras y frases es llamado lexicón de opinión o en inglés “opinion lexicon (sentiment lexicon)”. [16]

Dentro del análisis de sentimientos las palabras y frases de sentimientos son importante, pero no son suficientes. El problema es aún más complejo, se puede decir que el lexicón de opinión es necesario pero no suficiente para el análisis de sentimientos. A continuación se enumeran algunos de los problemas:

1. Una palabra de opinión positiva o negativa puede tener orientaciones opuestas según el dominio de aplicación.

2. Existen oraciones que pueden contener palabras de opinión pero que no expresen ninguna opinión.

3. La ironía presente en oraciones con o sin palabras de sentimientos puede ser bastante complejo de trabajar.

4. Muchas oraciones sin palabras de opinión pueden implicar opiniones. Muchas de estas oraciones son oraciones objetivas que se utilizan para expresar información real.

Capítulo 4

4 Diseño lexicón de opinión

En el presente capítulo se realiza el diseño del lexicón que posteriormente será construido, para ello se realiza una breve descripción de las metodologías con enfoques en corpus lingüísticos y los principales requerimientos del lexicón para luego proponer la metodología a trabajar y el proceso que se debe adoptar para la realización del lexicón de opinión.

4.1 Metodología utilizada

El enfoque metodológico basado en un corpus se caracteriza como se vio en la sección 3.4 por utilizar un corpus lingüístico al cual se le aplican reglas y algoritmos sobre éste, donde se puede notar que las principales técnicas utilizadas en este enfoque son:

- **Utilización de algoritmos de doble propagación o expansión de lexicones**: utiliza la dependencia de opiniones y aspectos para extraer nuevas palabras de opinión. Se basa en relaciones de dependencias dentro del corpus. Algunos de estos algoritmos emplean un listado de palabras iniciales en los que posteriormente por reglas lingüísticas de asociación se agregan palabras contenidas en el corpus utilizado. Como por ejemplo los presentados en [31] y [35].

- **Aplicación de técnicas estadísticas o algoritmos de aprendizaje no-supervisado**: principalmente en la literatura encontramos la aplicación de la medida PMI o “Poinwise Mutual Information”, se observan además distintas variaciones de ésta como el PMI-IR que es utilizado para clasificación de frases y no sólo a nivel de palabra. Como por ejemplo en lo realizado en [32], o en los trabajos de Mohammad en [39], [37], [38].

4.1.1 Requerimientos

Para realizar la propuesta metodológica y el posterior diseño del lexicón es necesario tener en consideración los requerimientos que se tienen para el lexicón de opinión:

- **Para ser utilizado en proyecto "OpinionZoom":** el lexicón que se construya tiene como finalidad ser utilizado en el proyecto por lo que debe estar adaptado a las funcionalidades del proyecto, sin perder el potencial de ser ocupado para futuras investigaciones.

- **Debe contener características propias del lenguaje utilizado en Chile:** como el análisis de opinión que se realiza en el centro corresponde a las opiniones que se realizan dentro de Chile, aplicadas a organizaciones, marcas, productos y servicios desde comentarios que son realizados por personas chilenas, el lexicón de opinión debe poder captar el uso propio de las palabras en el español que se ocupa en Chile.

- **Lexicón de opinión para identificación de polaridad:** un lexicón de opinión como lo vimos en la Sección 3.1 se define como un repositorio de palabras que se ocupan para expresar opiniones catalogadas como positivas o negativas. La categorización de las palabras puede ser sólo entre palabras positivas o negativas, se pueden incluir además palabras neutras así como también niveles de polaridad en diferentes escalas, por ejemplo una escala entre [-1 1] o la que se utiliza en el centro que es una escala entre [-5 5]. Como el análisis que se realiza en el centro de investigación determina niveles de polaridad, el lexicón que se necesita debe tener identificados niveles de polaridad.

- **Basado en los comentarios que se realizan en Twitter:** como dentro del proyecto OpinionZoom se realiza el análisis de opinión desde comentarios que se realizan en Twitter, el lexicón debe poder ser usado para el análisis de futuros comentarios en esta red social, caracterizando términos propios que puedan ser utilizados dentro del dominio de Twitter.

4.1.2 Metodología propuesta

Luego de haber identificado los principales requerimientos que se tienen para la construcción del lexicón y para la usabilidad de éste se plantean los siguientes pasos metodológicos:

1. **Definición y construcción de corpus lingüístico:**
   Estudiar las diferentes alternativas de corpus a utilizar y elegir aquel que mejor se ajuste a las características que se necesitan rescatar dentro del lexicón de opinión.

2. **Determinación de la medida de polaridad a utilizar:**
   Dado que el lexicón debe ser con niveles de polaridad, se necesita identificar una medida de polaridad que pueda ser utilizada sobre el corpus lingüístico.

3. **Construcción del lexicón de opinión:**
   Con los recursos identificados, armar el lexicón de opinión.

Finalmente con un corpus lingüístico y reglas para medir la polaridad sobre él, se puede obtener un lexicón de opinión orientado al dominio del corpus utilizado. En la Figura 2 se puede apreciar la metodología de construcción del lexicón propuesta:

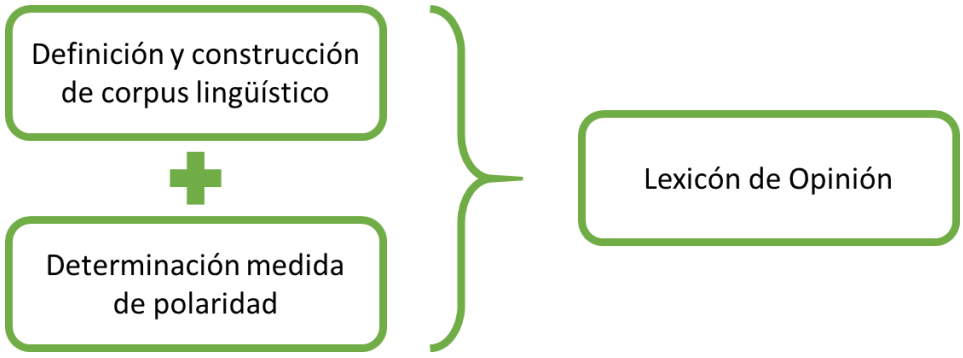


FIGURA 2: METODOLOGÍA PROPUESTA
FUENTE: ELABORACIÓN PROPIA

4.2 Proceso de construcción

El proceso de construcción que se presenta a continuación corresponde a las etapas de clasificación del corpus, en tweets negativos y positivos, para luego establecer las etapas de normalización y limpieza de los datos obtenidos.

Con los datos procesados, se realiza la asignación de polaridad correspondiente considerando una corrección simple de la negación.

4.2.1 Clasificación de corpus

Para la clasificación de corpus existen diversas técnicas, de estudios previos en el área. Teniendo en cuenta que el corpus proviene de comentarios en Twitter, éstos presentan características particulares que entregan información adicional a los métodos tradicionales de análisis de polaridad, existen diversas técnicas entre las que destacan técnicas que ocupan la cantidad de RT (retweet) para identificar polaridad, la utilización de los hashtag (#), o bien la utilización de emoticones para identificación de polaridad.

Considerando que la finalidad de este trabajo es lograr un buen lexicón, la construcción del corpus tiene un papel relevante en su desempeño, pero es necesario entregar un modelo simple, que también pueda ser fácilmente replicable en caso de tener que acotar el dominio del lexicón.

Se observa que la utilización de emoticones para el análisis de polaridad en Twitter entrega buenos resultados en comparación con otras redes sociales. Este fenómeno fue revisado en [40], en que dentro de los distintos sistemas estudiados para identificar la polaridad, el uso de emoticones presenta uno de los niveles más altos de desempeño, en que las métricas de evaluación se encuentran sobre el 80%.

Para la identificación de polaridad para diferentes redes sociales, en [40] el sistema que mejores resultados presenta dentro de Twitter es aquel que utiliza emoticones para determinar su polaridad, llegando a niveles de desempeño del 93%. Por lo cual se propone la utilización de este recurso siguiendo el método realizado en la construcción del Corpus Base Sentiment 140 en [41], obteniendo así un corpus que pueda estar clasificado en comentarios positivos y negativos.

En la Figura 3 se muestran los pasos propuestos para la clasificación del corpus:

FIGURA 3: PROCESO CLASIFICACIÓN CORPUS
FUENTE: ELABORACIÓN PROPIA

Por tanto los siguientes pasos son propuestos para la construcción del corpus etiquetado:

1. En primera instancia se debe realizar la recolección de tweets que se encuentren dentro del dominio que se requiere para el lexicón. Ej.: Si se requiere un lexicón orientado al turismo, se debe hacer una recolección de tweet dentro de este dominio como en [42].

2. Luego, se debe realizar un filtrado de tweets de todos aquellos que contienen emoticones. Se utilizarán sólo aquellos comentarios que utilicen emoticones, de forma que puedan expresar polaridad.

3. Finalmente se debe realizar la asignación de polaridad a cada tweet con las siguientes reglas de clasificación:

   - **Tweet positivo:** el tweet contiene sólo emoticones positivos y ningún emoticón negativo es clasificado como positivo.
   - **Tweet negativo:** tweet que contiene sólo emoticones negativos y ninguno positivo es considerado negativo.

     Estas reglas de clasificación fueron ocupadas en [43], para un sistema de clasificación de tweets presentada en la competición internacional de la tarea compartida: “SemEval 2015 Task 10” que tiene como objetivo tiene promover la investigación en el análisis de los sentimientos de los tweets.

Dado que el sistema propuesto en [43] se ocuparon los emoticones presentados en la Tabla 5 para la detección de polaridad en comentarios de twitter, y además representa un sistema desarrollado en el último año (2015) con un alto porcentaje de clasificación, cercano al 72%. La regla de clasificación por emoticones implementada utilizará, por tanto, los emoticones de la Tabla 5.

Emoticones Positivos	Emoticones Negativos
:) , (: , ;) , :-) , (-: , :D , :-D , :P , :-P	:(,): , ;(, :-(,)-: , D: ,D-: , :’(, :’-(,)’: ,)-’:

TABLA 5: EMOTICONES POSITIVOS Y NEGATIVOS
FUENTE: ELABORACIÓN PROPIA

4.2.2 Normalización y limpieza de palabras

Como paso previo para la asignación de polaridad, todos los tweets deben ser sometidos a un preprocesado ad-hoc, para tratar el uso particular que se hace del lenguaje en Twitter, se propone realizar un normalizado según [44] ya que se ajusta a los datos que se tienen.

- **Tratamiento de emoticonos:** existe una gran variedad de símbolos que se emplean para reflejar un estado de ánimo. Como estos términos no son incluidos dentro del lexicón, cada emoticón presente es eliminado del comentario porque ya no representa valor.

- **Eliminación de URL's**: las direcciones web presentes en un tweet son eliminadas. .
- **Corrección de abreviaturas más frecuentes:** se sustituyen algunos de los vocablos no gramaticales más habituales (por ejemplo: “q" “xq",...) por su forma reconocida.

- **Normalización de risas:** las expresiones típicas que permiten reflejar este fenómeno (por ejemplo: “Jajajaja”, “JEJEJEJE",...), son normalizadas como jaja siguiendo el patrón utilizado en trabajos anteriores dentro del centro, específicamente la normalización ocupada en [45].

- **Tratamiento de elementos específicos de Twitter** (“@" y “#"): las menciones a usuarios se eliminan ya que no representan palabras de utilidad para el lexicón. Respecto a los hashtags (por ejemplo: “#Bienvenidos”, “#CopaAmerica”), si aparecen al principio o al final del tweet se elimina del mismo. En caso contrario se suprime solamente el “#" (por ejemplo: “#Bienvenidos” pasa a ser “Bienvenidos") por ser considerado como parte de la frase.

4.2.3 Asignación de polaridad

Para la asignación de polaridad de las palabras dentro del corpus se ha decidido utilizar la medida del PMI, esto, pues es una medida utilizada con resultados positivos en varios estudios, representa niveles de polaridad de las palabras, no sólo una separación entre negativos y positivos.

Como esta medida de polaridad trabaja en base a frecuencias, no discrimina en gran medida la estructura lingüística que pueda tener la oración. Es además una medida que puede ser aplicada a diferentes idiomas, como en inglés, francés, portugués, consiguiendo resultados favorables.

Uno de los supuestos de esta medida es que palabras positivas pueden ser encontradas en mayor frecuencia en comentarios positivos, mientras que palabras negativas pueden ser encontradas en comentarios negativos, es así como se le asigna un valor de polaridad de las palabras como sigue:

$$Medida\ Polaridad\ (w) = \ PMI\ (w, positivo) - \ PMI\ (w, negativo)$$

En que PMI corresponde a “Pointwise Mutual Information” que es definida en [32] como:

$$PMI(w, positivo) = log_2 \frac{freq(w, positivo) * N}{freq(w) * freq(positivo)}$$

Donde:

- $freq(w, positivo)$ = número de veces que aparece w en tweets positivos.
- $freq(w)$ = frecuencia total que aparece el término w en el corpus.
- $freq(positivo)$ = número total de tokens en tweets positivos.
- N = total de tokens dentro del corpus.

Para $PMI(w, negativo)$, el cálculo es análogo.

Finalmente se puede simplificar la ecuación de polaridad a:

$$Medida\ Polaridad\ (w) = \ log_2 \frac{freq(w, positivo) * freq(negativo)}{freq(w, negativo) * freq(positivo)}$$

4.2.4 Corrección de la negación

Un problema importante a considerar en el desarrollo de técnicas como la propuesta es la incorporación de un método para el procesamiento de la negación en oraciones. La corrección de la negación ha sido propuesto como trabajo futuro en diversas investigaciones del área de construcción de lexicones en español como en [42] y se ha resuelto en trabajos más complejos de sistemas para minería de opiniones en español como en [44] y [46].

Para la solución de este problema existen diversos modelos planteados, muchos de los cuales se encuentran en lengua inglesa; dado que considerar modelos para lengua inglesa con en textos en español no conducirían a resultados confiables, se ha decidido ocupar los trabajos en sistemas de minería de opiniones en textos en español como se realiza en [47] y que fueron ocupados en trabajos anteriores dentro del WIC, como en los trabajos de Balasz en [17] y [45] de una forma simplificada.

Se han tomado términos los “no”, “nunca” y “sin” para el tratamiento de la negación en comentarios de Twitter, ya que son estos términos los que se han considerado en análisis de opiniones en corpus en español, como en [47] . Considerando los comentarios dentro del corpus contienen en promedio 11 palabras, esto deja poco espacio para la elaboración de frases y párrafos muy complejos, y que necesiten un análisis más exhaustivo.

El modelo más sencillo que se ha utilizado, por ejemplo en los trabajos de Sauri en [48] es utilizar una la técnica llamada “switch negation” que cambia la polaridad del ítem que se encuentra inmediatamente a continuación del término.

Por ejemplo en el siguiente comentario:

$$@JORGEGOLERO\ Hace\ frío\ y\ ese\ poncho\ no\ abriga\ mucho\ :($$

En que en los términos: “no abriga”, se pueden descomponer por la palabra que está siendo negada “abriga” y en la palabra “no”. Si consideramos este comentario como un comentario negativo y aplicamos la regla de corrección de negación simple, la palabra “abriga” que inicialmente sería considerada como negativa, con la corrección realizada se consideraría como palabra positiva.

Para el caso del corpus lingüístico utilizado, el término “no” aparece en su mayoría seguido de formas de los verbos: “ser”, “estar” considerados en algunos análisis de opinión como verbos neutros o bien, *stopwords*, por lo que la corrección de esta negación no conduciría a grandes resultados erróneos y ayudaría a no tener grandes cantidades de palabras etiquetadas con su polaridad contraria.

Finalmente, considerando que se ocupa la herramienta FreeLing[3] para la identificación de palabras, sólo se ha considerado el término “no” para corregir la negación, que es el adverbio de negación capaz de identificar esta herramienta.

[3] http://nlp.lsi.upc.edu/freeling/, 11 de enero de 2016.

4.3 Recursos

Dentro del diseño del lexicón de opinión se han establecido algunos recursos necesarios para su construcción. En específico para el enfoque metodológico basado en un corpus lingüístico, es necesario tener un corpus que pueda ser procesado.

El corpus utilizado en este proyecto, es un corpus que ha sido refinado desde los datos obtenidos por el proyecto OpinionZoom en el seguimiento de redes sociales, en particular de Twitter.

4.3.1 Corpus lingüístico

Actualmente, la lingüística de corpus se encuentra asociado con la búsqueda de concordancia entre líneas y listas de palabras generados por programas informáticos, en un intento de dar sentido a los fenómenos asociados a grandes textos o grandes colecciones de textos pequeños [49].

El conjunto de textos o corpus ocupados suelen ser de un tamaño que desafía el análisis humano que sea realizado por si solo en un periodo de tiempo razonable. Es la gran escala de datos utilizados lo que explica el uso técnicas automáticas para leer los textos. A menos que se utilice un computador para leer, buscar y manipular los datos, trabajar con grandes conjuntos de datos no es factible debido al tiempo que necesitaría un analista humano o un conjunto de analistas, para buscar a través del texto [50].

Para la realización del lexicón de opinión con un enfoque basado en un corpus lingüístico es necesario primeramente, la correcta elaboración del corpus. Como el lexicón generado será utilizado para análisis de sentimientos en Twitter, el corpus utilizado son comentarios en Twitter de cuentas que presentan las características que debe tener finalmente el lexicón de opinión.

El corpus corresponde a los tweets de usuarios que son utilizados por el centro para realizar el análisis de polaridad, teniendo así una coherencia entre las características del corpus del cual se extraerán las palabras para formar el lexicón y el corpus donde se utilizará este recurso.

4.3.1.1 Base de tweets

A continuación se detalla la base con que trabaja OpinionZoom, que será utilizada para caracterizar el corpus que finalmente será utilizado.

La base de tweets con que trabaja OpinionZoom se inició con el trabajo de título de Felipe Vera, actual ingeniero parte del proyecto OpinionZoom. Se formó con la recolección de cuentas chilenas de alta popularidad, utilizando primeramente 235 cuentas chilenas con más de cien mil seguidores, junto a los seguidores de éstas, para tal cantidad de usuarios fue necesario realizar filtros de cuentas de habla hispana, más de 90 seguidores, más de 20 tweets y que sea una cuenta pública como se detalla en [51].

A esta versión preliminar se le agregaron medios de comunicación chilenos y algunos periodistas asociados a modo de incluir cuentas relacionadas con contenido chileno. Con este trabajo, se incluyeron aproximadamente 8 millones de cuentas, finalmente se eliminaron cuentas repetidas obteniéndose una base de datos de 5 millones de cuentas para la base de datos llamada “GORDA”.

Análisis posteriores del tráfico que producían estas cuentas concluyeron que el 10% de los usuarios generaban el aproximadamente el 80% del tráfico de tweets chilenos. En la Figura 4 se aprecia que el 80% de los usuarios no alcanzan a generar el 20% de los tweets que se generan en un día, lo que refuerza la idea que un grupo de usuarios sólo consume información y no está expresando su opinión.

Gracias a este análisis se eliminaron cuentas que estuvieran inactivas, cuentas con un índice menor de tweets al promedio diario, obteniéndose cerca de 500.000 cuentas chilenas.

Finalmente, se eliminaron manualmente cuentas no chilenas y algunas cuentas no representativas obteniéndose 170.000 usuarios que actualmente se siguen sus comentarios y las menciones que se realizan sobre ellos.

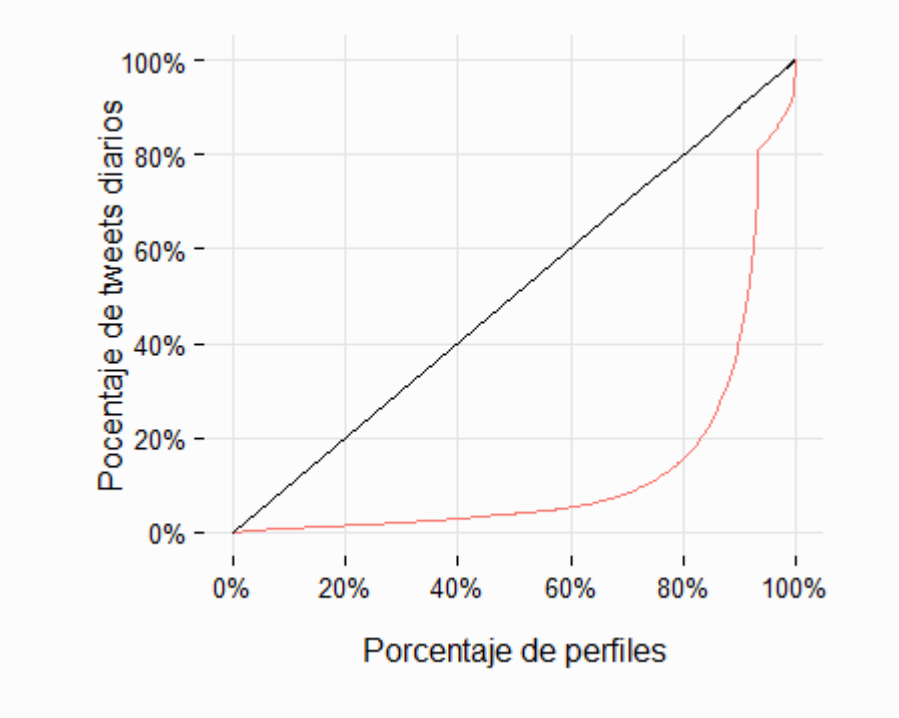


FIGURA 4: CANTIDAD DE TWEETS DIARIOS GENERADOS
FUENTE: EXTRAÍDO DE [51].

La base de tweets "GORDA" estuvo en funcionamiento durante aproximadamente 4 meses, entre 2 de Mayo de 2015 y 14 de Septiembre de 2015, en el Web Intelligence Centre (WIC) para uso del proyecto "OpinionZoom" y las investigaciones asociadas. Para esta base de datos se utilizaba postgres guardando los tweets en formato JSON.

Actualmente por requerimientos de tiempo y almacenamiento se ocupa la base de datos "NEW GORDA" que funciona desde 1 de Septiembre, de acuerdo al mismo sistema anterior pero que ha migrado a SOLR para disminuir el tiempo de consulta.

NEW GORDA

Base de tweets que se utiliza en el Web Intelligence Centre WIC para proyecto "OpinionZoom". Sigue a 170.000 cuentas chilenas y sus menciones en Twitter, a partir del primero de Septiembre de 2015. Contiene tweets de al menos 23 millones de usuarios.

Hasta el 8 de Octubre a las 17.00 "New Gorda" contiene 60.504.777 tweets.

Para entender qué tipo de usuarios contiene la base de datos, se agruparon los 235 usuarios iniciales (anexo 5) que se ingresaron a la base de datos, en 12 categorías, que se muestran a continuación:

- **Medios de comunicación:** como @biobio, @T13, @TVN, @CNNChile, @Cooperativa, @lacuarta. Se incluyen medios masivos de comunicación como televisión, radio, periódicos.
- **Periodistas:** dentro de esta categoría se encuentran las cuentas de periodistas principalmente de programas informativos de televisión, como @consuelosaav, @tv_mauricio, @SoledadOnetto.
- **Políticos:** se encuentran las cuentas de personas asociadas a la política chilena, como @camila_vallejo, @mxperez, @lgolborne, @marcoporchile.
- **Actores**: dentro de esta categoría se han incluido cuentas de actores chilenos como: @benjavicunaMORI, @iamdelafuente, @zabaletachile.
- **Rostros de televisión:** se presentan cuentas de personas que aparecen en programas de televisión abierta nacional, dentro de esta categoría se incluyen: @sergiolagos, @KarendTV, @RafaAraneda, @rubionatural.
- **Modelos:** dentro de esta categoría se encuentran modelos que han participado en programas de televisión o tienen presencia en medios masivos de comunicación como: @vale_ortega, @Lucilavit, @lucialopezchile.
- **Programas de televisión:** se encuentran dentro de esta categoría aquellos programas de televisión asociados principalmente a programas de entretención que son parte de canales de transmisión abierta en chile como: @asisomosoficial, @SV_CHV, @Enportada, @13AR, @YINGO_oficial, @buenosdiatodos.
- **Cantantes:** cuentas de destacados músicos nacionales como: @DjMendezmusic, @franciscamusic, @Los_Bunkers, @anatijoux.
- **Deporte:** se encuentran asociados deportistas chilenos, clubes de futbol y cuentas de periodistas ligados al deporte como: @elfergonzalez, @tomasgonzalez1, @el_mago_oficial, @udechile, @ColoColo.
- **Utilidades:** en esta categoría se han incluido cuentas que presentan información útil en caso de emergencias. Se presentan las cuentas de: @PDI_CHILE, @reddeemergencia, @sismos_chile, @onemichile, @metrodesantiago, @Carabdechile, @CruzRojaInforma.
- **Comunicadores:** se encuentran cuentas asociadas a comunicadores de radio y televisión, que no representan un status de periodista. Se incluyen cuentas de: @jasalfate, @renenaranjo, @Rumpy1000, @RinconSalfate.
- **Otros:** esta categoría contiene usuarios que no han sido clasificados en las categorías anteriores, se encuentran youtubers, algunas marcas, entre otros. Por ejemplo: @GermanGarmendia, @hoytschile, @PatricioDelSol, @MovistarChile.

Teniendo en cuenta las categorías anteriores para entender su distribución dentro del corpus se presenta la Figura 2 que muestra gráficamente la agrupación de usuarios iniciales realizado:

FIGURA 5: DISTRIBUCIÓN USUARIOS INICIALES
FUENTE: ELABORACIÓN PROPIA

Notar que las categorías realizadas fueron hechas en base al perfil de las cuentas iniciales y que el corpus se expandió con los seguidores de estas cuentas teniendo un universo más grande de usuarios, no incluidos en la categorización realizada.

Capítulo 5

5 Construcción

Para la construcción del lexicón de opinión que se propuso en la sección 4.1.2 es necesario definir y construir un corpus lingüístico adecuado, para luego elaborar el lexicón de opinión de acuerdo a los recursos que se tienen. En este capítulo se detalla el proceso de construcción del corpus, recurso clave para la correcta elaboración del lexicón de opinión, para luego ahondar en el proceso de construcción del lexicón propiamente tal.

5.1 Construcción corpus lingüístico

Para la construcción del corpus que será utilizado para la elaboración del lexicón de opinión se utilizó la base de tweets con que contaba el centro de investigación WIC, llamada “GORDA”. Dado que la construcción del corpus requería consultas a la base de datos de acuerdo a emoticones, se ocuparon tweets almacenados en la base de datos *PostgreSQL.*

PostgreSQL es un sistema relacional de bases de datos orientado a objetos, de código libre (open source), bajo la licencia de libre uso: *PostgreSQL License*. Este sistema de base de datos puede ser utilizado en los sistemas operativos de LINUX, UNIX o Windows, brindando desarrollo activo durante más de 15 años [52]. Presenta interfaces de programación para C/C++, Java, .Net, Perl, Python, entre otros.

> **GORDA**
>
> Primera base de tweets utilizada en el Web Intelligence Centre WIC para proyecto “OpinionZoom”. Contiene 242.464.520 tweets, entre 2 de Mayo de 2015 y 14 de Septiembre de 2015, correspondientes a 1.013.005 de usuarios diferentes. Diariamente existen 1.707.497 tweets.

Considerado que la base de datos “GORDA” contenía fechas en que no se almacenaban tweets, los meses en que mejor comportamiento (pocas caídas del sistema, pérdida de datos) fueron los meses de mayo y junio, por lo que se utilizaron estos meses para la recopilación de comentarios.

5.1.1 Recolección de datos

En primera instancia se rescataron los comentarios desde “GORDA” que contenían los emoticones propuestos en la sección 4.2.1 donde se puede apreciar en la Tabla 5 los emoticones a utilizar para la clasificación.

El primer filtro que consideraba sólo tweets con emoticones positivos y negativos deja aproximadamente 40,7 mil comentarios diarios. En la Figura 6 se puede apreciar la cantidad de tweets con emoticones durante la semana del 12 de mayo al 25 de mayo.

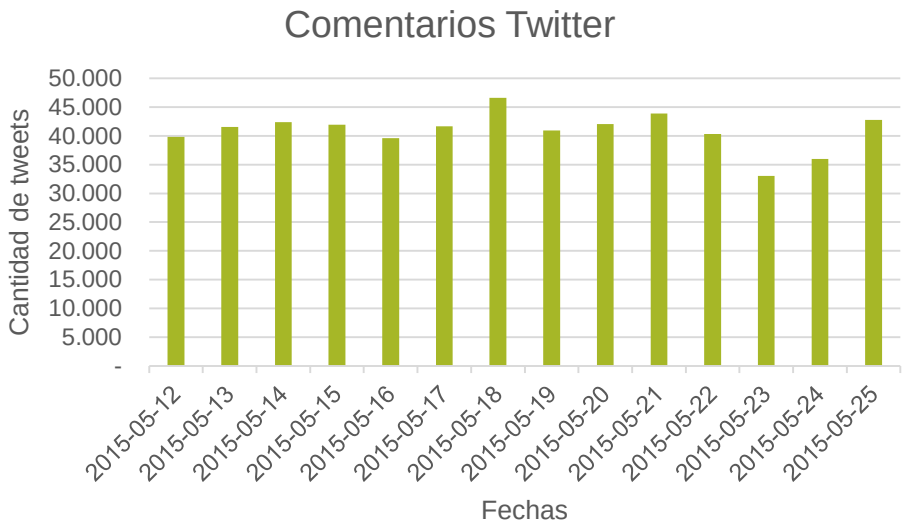


FIGURA 6: COMENTARIOS DOS SEMANAS TWITTER
FUENTE: ELABORACIÓN PROPIA

La consideración inicial de tweets necesarios para la construcción del corpus fue de 126 mil tweets mínimo, teniendo en cuenta que existen aproximadamente 76.000 combinaciones Pos – Lemmas en la herramienta de análisis de lenguaje utilizada para el posterior normalizado a idioma español (FreeLing[4]) de los comentarios extraídos, observando que en promedio se utilizan 11 palabras en cada comentario de las cuales 6 de ellas serán utilizadas para formar el lexicón de opinión por eliminación de stopwords. Finalmente en el modelo propuesto es necesario tener una frecuencia de la palabra mayor a 5 en cada set de tweets positivos y negativos, lo que da el número mínimo calculado anteriormente de 126.000 comentarios.

El tamaño mínimo sugerido del corpus es logrado los primeros 3 días, pero se ha considerado que se tendrá un tamaño prudente para el corpus cuando el tamaño del lexicón no aumente considerablemente en cantidad de palabras, ni en frecuencia; como se detalla en la siguiente sección 5.2.3 Tamaño lexicón.

[4] http://nlp.lsi.upc.edu/freeling/, 18 marzo de 2016

5.1.2 Clasificación de polaridad en comentarios

Una dificultad dentro de este trabajo es la tokenización realizada para las consultas dentro del campo de texto de los tweets, ya que dejaba algunos errores en la detección de emoticones; por ejemplo, se muestra a continuación un comentario que en primera instancia fue clasificado como negativo, por presentar el emoticón “D:” pero que realmente era parte del nombre de un usuario de twitter:

Comentario original:

$$RT\ @CrichardGD\mathbf{:}\ Q\ decidor\ q\ @Savantdija\ se\ confunda\ y\ se\ le\ venga$$

$$\#davidarellano\ a\ la\ memoria,\#savantdijacolocolino\ \#clubdelmal$$

Emoticón erróneamente encontrado:

$$“\ \ldots @CrichardG\boldsymbol{D}\mathbf{:}\ Q\ decidor\ \ \ldots”$$

Otra dificultad es que la consulta realizada a la base de datos no discriminaba en comentarios que podían presentar emoticones con distinta polaridad. A continuación se presenta un ejemplo de comentario que presenta ambos tipos de emoticones:

Comentario original:

$$@nikiValentine\_\ //\ \#FelizMartes\ para\ ti\ también\ (\mathbf{:}\ saludos,\ buenos\ días\ D\mathbf{:}$$

Presencia emoticón positivo:

$$“\ \ldots para\ ti\ también\ (\mathbf{:}\ \ldots”$$

Y emoticón negativo:

$$“\ \ldots buenos\ días\ \boldsymbol{D}\mathbf{:}\ \ \ldots”$$

Existe, como se mencionaba anteriormente, una gran cantidad de problemas asociados a la clasificación de polaridad en comentarios, uno de ellos es la ironía presente, pero que este trabajo no se hará cargo de ello. Otro problema asociado es el error al tipear o escribir los emoticones. Por tanto, para mejorar levemente el corpus construido se ha decidido eliminar comentarios por incongruencias con palabras conocidas como positivas o negativas.

Este pequeño set de palabras conocidas como positivas y negativas se extrajo del tesauro de Roget (en inglés, Roget's Thesaurus) [53], estas palabras pueden ser observadas en el anexo 1. Las palabras fueron extraídas por estar asociadas a las palabras "positive" y "negative", traducidas al español, eliminando palabras repetidas y aquellas que no representaban considerablemente su polaridad para el idioma.

Finalmente al grupo de comentarios clasificados como positivos o negativos según sus emoticones presentes, se le hizo una limpieza por incongruencias donde se eliminaron aquellos comentarios que presentaban palabras con connotación negativa dentro de comentarios clasificados como positivos según los emoticones presentes. Análogamente se eliminaron de comentarios clasificados por sus emoticones como negativos, aquellos que tengan presentes palabras positivas.

A continuación se muestran un par de comentarios que presentan este tipo de incongruencias:

Comentario original:

$$@AngelaCaripan\ ES\ HORRIBLE :)\ xDDDDD$$

Comentario original:

$$@miclaro\_cl\ Han\ venido\ 2\ veces\ "técnicos", pero\ NADA, el\ wifi\ tiene\ una$$

$$intermitencia\ horrible, pero\ no\ se\ preocupen, volveré\ a\ movistar :)$$

Ambos presentan la palabra "horrible" conocida por su polaridad negativa, etiquetada dentro de un comentario positivo, por la polaridad del emoticón ":)".

Este procedimiento de eliminación de incongruencias fue de ayuda para la limpieza del corpus, pero no ayuda considerablemente en su precisión ya que la eliminación de estos comentarios representa bajo el 1% de los comentarios dentro del corpus.

El caso particular de la ironía presente en los comentarios no es parte de los alcances de este trabajo, por lo que este tipo de comentarios no han sido analizados.

5.1.3 Arquitectura construcción corpus etiquetado

Para la extracción de comentarios y posterior clasificación se utilizó en primer lugar un computador con sistema operativo Windows 10 de 64 bits, para luego realizar el procesamiento de tweets y normalización en un computador con sistema operativo Ubuntu 14.04 de 64 bits.

El lenguaje de programación utilizado para el etiquetado de comentarios fue Java SE 8, dado que para el posterior procesamiento de lenguaje se utilizó la herramienta FreeLing [54] con su API para JAVA, teniendo un sistema programado en la misma plataforma durante todo el proceso de construcción.

FreeLing es una herramienta desarrollada por "TALP Research Center" en la Universidad Politécnica de Cataluña bajo la licencia *GNU General Public License* [54]. Este paquete provee una librería con servicios de análisis de lenguaje, soporta lenguajes como español, catalán, italiano, inglés, ruso, entre otros. Algunos de los servicios de análisis de textos ofrecidos son: tokenización, segmentación de oraciones, análisis morfológico, PoS tagging.

Java es un lenguaje de programación que desde su lanzamiento en 1995 por Sun Microsystems de alto nivel y plataforma de software que se encuentra en más de 50 millones de computadores y en miles de millones de aparatos a lo largo del mundo. Existen cerca de 9 millones de desarrolladores que han realizado aplicaciones de Java en diferentes industrias [55]. Las cualidades de concurrencia, la orientación a objetos y que sea de propósito general, hacen que este lenguaje sea suficiente para los requerimientos del sistema.

Los pasos que se siguieron para la construcción del corpus clasificado fueron las siguientes:

1. Extracción comentarios en español con emoticones positivos y negativos desde base de datos "GORDA" almacenada en *PostgreSQL.*
2. Tokenización de los comentarios
3. Eliminación de comentarios que no presentaban emoticón o que presentaban emoticones con distintas polaridades.
4. Lematización de comentarios para la normalización y posterior limpieza de éstos.
5. De los comentarios clasificados como positivos por la presencia de emoticones positivos sin ningún emoticón negativo y que presentaban alguna palabra conocida como negativa de las presentes en el anexo 1 fueron eliminados.
6. De los comentarios clasificados como negativos por sus emoticones y que presentaban incongruencias con palabras positivas fueron eliminados.

A continuación, en la Figura 7 se puede observar un diagrama del proceso de construcción del corpus:

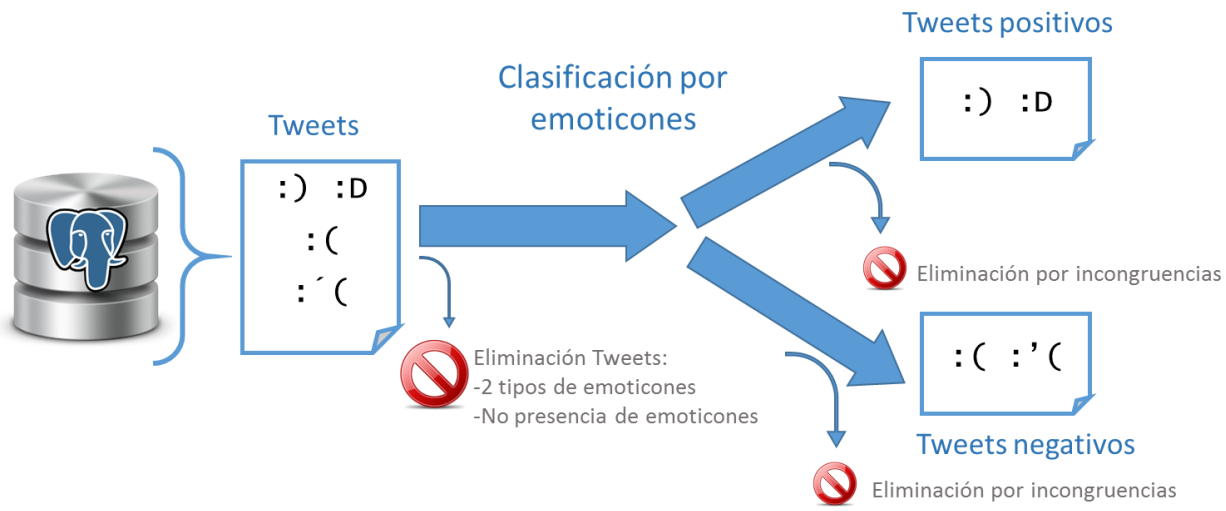


FIGURA 7: PROCESO CONSTRUCCIÓN CORPUS
FUENTE: ELABORACIÓN PROPIA

De este procedimiento, aproximadamente el 46% de los tweets que fueron recolectados de la base de datos “GORDA” fueron eliminados por presencia de emoticones con distinto tipo de polaridad, o bien, porque no presentaban emoticones propiamente tal. Quedando un 69% de comentarios etiquetados como positivos y un 31% como negativo, notar que cerca del 0,01% de estos comentarios fueron eliminados por incongruencias.

Para los días entre 12 de mayo de 2015 y 17 de junio de 2015, exceptuando los días comprendidos entre 4 y 9 de junio por falta de almacenamiento de tweets en la base de datos, se obtienen los resultados presentados en la Tabla 6, donde se incluye el promedio diario de comentarios recolectados, etiquetados y eliminados.

	Promedio diario
Tweets recolectados de “GORDA”	40.744
Eliminados por presencia emoticones 2 polaridades o no presencia de emoticones	18.874
Comentarios positivos	15.020
Comentarios negativos	6.851
Eliminados por incongruencias en comentarios positivos	115
Eliminados por incongruencias en comentarios negativos	100
Tweets etiquetados	21.656

TABLA 6: PROMEDIO TWEETS PROCESADOS
FUENTE: ELABORACIÓN PROPIA

5.1.4 Características del corpus

Para entender qué tipo de información contienen los mensajes dentro del corpus se ha decidido realizar un agrupamiento de los mensajes en diferentes conjuntos (clusters). Este procedimiento de análisis de grupos, también conocido como análisis de conglomerados o bien, del inglés “cluster analysis”, consiste en agruparlos según grupos lo más homogéneos entre sí y heterogéneos entre ellos.

Considerando que este proceso pretende entender a grandes rasgos el contenido de los mensajes y no realizar mayor análisis sobre éstos, se ha determinado la utilización del algoritmo “K-means” para agrupar los comentarios y se ha calculado el error cuadrado para determinar el número de grupos.

El algoritmo de agrupación K-means es un método que dispone N observaciones en un vector de I dimensiones en K grupos. Cada grupo es parametrizado por un vector $m^{(k)}$ llamado su media y las observaciones son asignadas al grupo con su media más cercana. La implementación utilizada se basa en el algoritmo propuesto en [56].

Para determinar el número de grupos para el algoritmo K-means se compararon las sumas de errores cuadrados para diferentes números de soluciones de grupos. Una solución apropiada para la determinación del número adecuado de grupos es el número tal que la suma de errores cuadrados decrece considerablemente.

De la observación gráfica de la Figura 8, en cuanto al comportamiento de la suma de errores cuadrados (WSS) versus el número de grupos (# de clusters), se observa que con 15 clústeres el modelo comienza una disminución considerable de los errores, por lo que se ha determinado agrupar los mensajes contenidos en el corpus en 15 grupos.

La implementación de este algoritmo fue realizada en R, un lenguaje y entorno utilizado para computación estadística y gráfica de libre uso por existir implementaciones de estas características para trabajos anteriores. Se utilizó R versión 3.2.3 (64 bits) y por limitaciones de procesamiento del computador se utilizó una muestra aleatoria del corpus correspondiente a 18.000 comentarios.

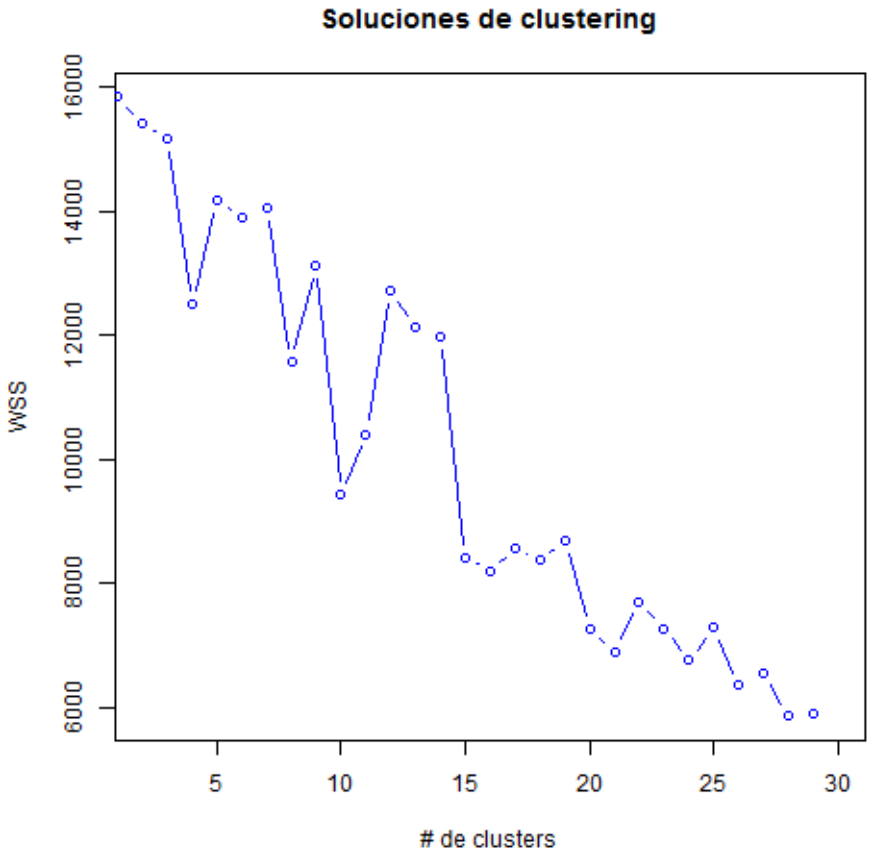


FIGURA 8: NÚMERO DE CLÚSTERES
FUENTE: ELABORACIÓN PROPIA

El análisis con más detalle de los grupos realizados se puede ver en el anexo 6, donde se muestra el gráfico de frecuencias de palabras dentro de cada grupo, junto a una breve descripción y comentarios asociados a cada grupo y en el anexo 7 el dendograma que organiza los grupos realizados.

A continuación en la Tabla 7 se puede ver una breve descripción de los comentarios presentes en cada uno de los grupos analizados, donde se aprecia que existe una gran cantidad de comentarios repetidos, ya sea por la gran cantidad de retweets, o bien por la gran cantidad de spam presente en el corpus. En la Tabla 8 se puede observar el tipo de comentario asociado a cada uno de los grupos.

Grupo 1	Comentarios que expresan tristeza, decepción
Grupo 2	Comentarios que expresan de deseos, voluntad de lograr algo
Grupo 3	Saludos, buenos deseos para el día
Grupo 4	Comparaciones, mejoras
Grupo 5	Acciones para mañana y comentarios asociados a Matías Ferrario
Grupo 6	Referencias al clima y a la temporalidad de situaciones
Grupo 7	Comentarios asociados a German Garmendia
Grupo 8	Agradecimientos, comentarios buscando seguidores para Ana Gabriel
Grupo 9	Comentario de German Garmendia
Grupo 10	Situaciones del día y eventos asociados al día lunes
Grupo 11	Comentario de agradecimiento de @tuitutil
Grupo 12	Situaciones del instante y comentarios de German Garmendia
Grupo 13	Deseos de buenos días y comentarios asociados a Jorge Vilches
Grupo 14	Comentarios con presencia de risas o burlas
Grupo 15	Saludos y comentarios en busca de seguidores para Ana Gabriel

TABLA 7: RESUMEN GRUPOS DE COMENTARIOS DE CORPUS LINGÜÍSTICO UTILIZADO
FUENTE: ELABORACIÓN PROPIA

Con este análisis se puede observar la gran cantidad de retweets de comentarios de "Matt Ferrario"[5], modelo y actor argentino con gran cantidad de seguidoras. Se puede notar que el corpus no sólo contiene cuentas chilenas, sino que también tiene presente cuentas de otros países como por ejemplo Argentina.

Vemos que existen 2 grupos de comentarios asociados a Germán Garmendia, un youtuber chileno, reconocido en el territorio y a nivel internacional. El twitter de German Garmendia[6] es la cuenta chilena con mayor cantidad de seguidores, llegando a 7.181.515 y el tercer twittero más influyente en la región según Adsocia [57]. Por tanto el corpus construido presenta una alta influencia de comentarios asociados a Germán Garmendia.

Otro hallazgo de la realización de grupos asociados al corpus es la gran influencia que tienen los retweets que a simple vista no se aprecia y sólo al agregar la información y procesarla por medio de herramientas de análisis de datos se puede notar el ruido que provocan dentro del corpus. Es el ejemplo del comentario que realiza la cuenta @TuitUtil[7] en agradecimiento a cada uno de sus seguidores, que representa un grupo completo dentro del análisis de grupos realizado, otro ejemplo es la gran cantidad de retweets asociado a JorgeVilchesV[8] y que representa parte importante de uno de los grupos analizados.

[5] https://twitter.com/matiasdferrario
[6] https://twitter.com/germangarmendia
[7] https://twitter.com/tuitutil
[8] https://twitter.com/jorgevilchesv

	Comentario asociado
Grupo 1	pucha, tu bio es la tristeza máxima, la vida sin palta y ajo no vale la pena vivirla po! :(
Grupo 2	Quiero dormir 24 horas nada mas :(
Grupo 3	Buenos días gente guapa. Por fin es viernes. A pasar buen día!!!! Se agradecen los RT:)
Grupo 4	al menos no se es mejor ver series k pasan piola pero son igual de buenas :D
Grupo 5	Y hoy fue mi último día en la pega desde mañana soy cesante :)
Grupo 6	Que frío hace, a ponerle onda que es viernes :)
Grupo 7	@GermanGarmendia a que hora sale el video…? Si podes saludame en unos de tus videos soy fanatico tuyo:D
Grupo 8	@PilarOsso muchas gracias!!! Nos vamos contentos a la cama ;)
Grupo 9	RT @GermanGarmendia: No importa lo que haga,siempre llego 10 minutos tarde a todo! :(
Grupo 10	Buen dia mucha energia y animo para Hoy Lunes que tengan una Buena Semana :)
Grupo 11	@GorkaAhal Gracias por seguirme,en breve te devuelvo follow :) #TuitUtil http://t.co/PAHlypOQ9j
Grupo 12	FULL ESTUDIO AHORA MAÑANA PRUEBA Y NO E TENIDO TIEMPO PARA ESTUDIAR ….VAMOS A VER QUE SE PUEDE LOGRARA A MENOS DE 20 HORA DE LA PRUEBA :p
Grupo 13	@JorgeVilchesV Estamos esperando el informe OFICIAL por TV :(#QEPDJorgeVilches #FuerzaJorge #JorgeVilches #VuelaAltoJorge ?
Grupo 14	adnradiochile jajaja conozco una que ahora se hace hincha del colo. Jajaja cambia la unión por colo colo. Menos mal no cb mi equipo :-)
Grupo 15	@yanka18m Podrías seguir a @ANAGABRIELRL y dar RT para que llegue a 500 mil seguidores? Gracias :) http://t.co/CIjsXEFB7p

TABLA 8: COMENTARIOS ASOCIADOS A GRUPOS DE CORPUS LINGÜÍSTICO
FUENTE: ELABORACIÓN PROPIA

5.2 Construcción lexicón

Para la construcción del lexicón, se utilizó como base el corpus construido anteriormente. Teniendo en cuenta que el corpus construido contiene en un 69% comentarios positivos y en un 31% comentarios negativos, se utilizaron diariamente la misma cantidad de comentarios de modo de tener un corpus equilibrado y que no se tuviera una mayor ponderación en polaridad positiva por tener mayor cantidad de comentarios positivos. Para esto, se fueron incluyendo alternadamente comentarios de distintas polaridades para la construcción del lexicón.

Se consideraron por tanto, 13,5 mil comentarios de twitter en promedio diariamente.

5.2.1 Arquitectura construcción lexicón

Para el desarrollo de la construcción, se utilizó como fue mencionado anteriormente el lenguaje de programación Java, y para el procesamiento de lenguaje se utilizó la herramienta FreeLing.

La elección de FreeLing fue realizada tomando en cuenta que la "Sociedad española para el procesamiento del lenguaje natural" (SEPLN) utiliza esta herramienta para la normalización de tweets en español en [58], y que diversos investigadores que han realizado estudios en textos en español han utilizado esta herramienta obteniendo buenos resultados.

El proceso de construcción en un principio consiste en la lectura de comentarios de twitter ya clasificados, según positivo o negativo del corpus para luego realizar el procesamiento del texto en cuanto a normalización y limpieza de datos.

Para cada comentario se sigue el siguiente proceso de normalización y limpieza:

1. **Tokenización del comentario**: proceso en el cual se divide el texto, considerando el comentario como una secuencia de caracteres, la tokenización separa estos caracteres en tokens (usualmente palabras). Un ejemplo de este proceso es la oración "Javier come mucho pan" es separada en los siguientes tokens: "Javier", "come", "mucho", "pan", ".". En esta primera instancia se consideró una tokenización simple, sólo por caracteres de espaciado para lograr capturar de mejor forma los caracteres especiales de Twitter como "@" o "#".

2. **Eliminación de URLs:** considerando que las URLs no presentan información para el lexicón de opinión se han eliminado todas las URLs presentes en el comentario, considerando las direcciones web que tienen la forma "http:.." o "www….".
3. **Eliminación de usuarios:** en algunos sistemas de análisis de opinión en Twitter se considera la eliminación del @ del usuario y capitalizar la primera letra del nombre. Considerando la gran cantidad de usuarios diferentes, y que los nombres propios no son un aporte relevante para la construcción del lexicón de opinión, los usuarios han sido eliminado. Un ejemplo de token eliminado es "@canal13".

4. **Eliminación de hashtags:** como se vio en la sección 4.2.2 se eliminan solamente los hashtag del principio y del final.

5. **Corrección de abreviaciones:** en esta etapa se corrigen algunas de las abreviaciones más frecuentes utilizadas en esta red social. Para esto se utilizó el listado elaborado por Jorge Balazs en [45] que se encuentra en el anexo 2.

6. **Normalización de risas:** se consideraron las risas más frecuentes y fueron normalizadas al término “jaja”. En el anexo 3 se encuentra el listado de risas más frecuentes utilizadas para este trabajo.

7. **Tokenización mediante FreeLing:** en esta etapa se realiza una tokenización más robusta que la anterior, ocupando la herramienta FreeLing para dejar las palabras correctamente separadas.

8. **Lematización:** se realiza la transformación de la palabra a su lema, este proceso conlleva un análisis morfológico de las palabras, es realizado para obtener palabras en su forma única, y evitar problemas con las diversas formas que pueden tener. Por ejemplo, “silla” es el lema de “sillas”, “niño” es el lema de “niñito”. El lematizador utilizado es el provisto por FreeLing para español.

9. **PosTagging:** En esta etapa del proceso se realiza la clasificación de la palabra según su categoría morfosintáctica. Para esto, nuevamente se han utilizado las herramientas provistas por FreeLing. Las etiquetas que provee FreeLing son las etiquetas Eagles, que pueden ser encontradas en el anexo 4.

10. **Eliminación de palabras que no expresan opinión:** como las palabras de opinión son consideradas como verbos, adjetivos, sustantivos y adverbios, se eliminan todas aquellas que no se encuentran dentro de esta clasificación. Por tanto, se mantienen en el lexicón las palabras que tengan el código entregado por “FreeLing” empezado en “A”, “R”, “N” y “V”.

11. **Corrección de la negación:** teniendo en cuenta que FreeLing sólo detecta un adverbio de negación, la palabra “no”, se ha considerado para la corrección de palabras que todas aquellas que se encuentran inmediatamente a continuación del adverbio “no” sean consideradas como palabras con polaridad contraria. Ejemplo: “No te quiero :(“ en este caso se la palabra “quiero” es etiquetada como positiva porque se encuentra luego de una negación al eliminar stopwords.

Luego de todo este proceso de clasificación y normalización de comentarios, para las palabras resultantes se almacena la frecuencia en comentarios negativos y en comentarios positivos, para luego aplicar la fórmula PMI vista en la sección 4.2.3.

A continuación en la Figura 9 se muestra un esquema que ejemplifica el procesamiento de comentarios realizado:

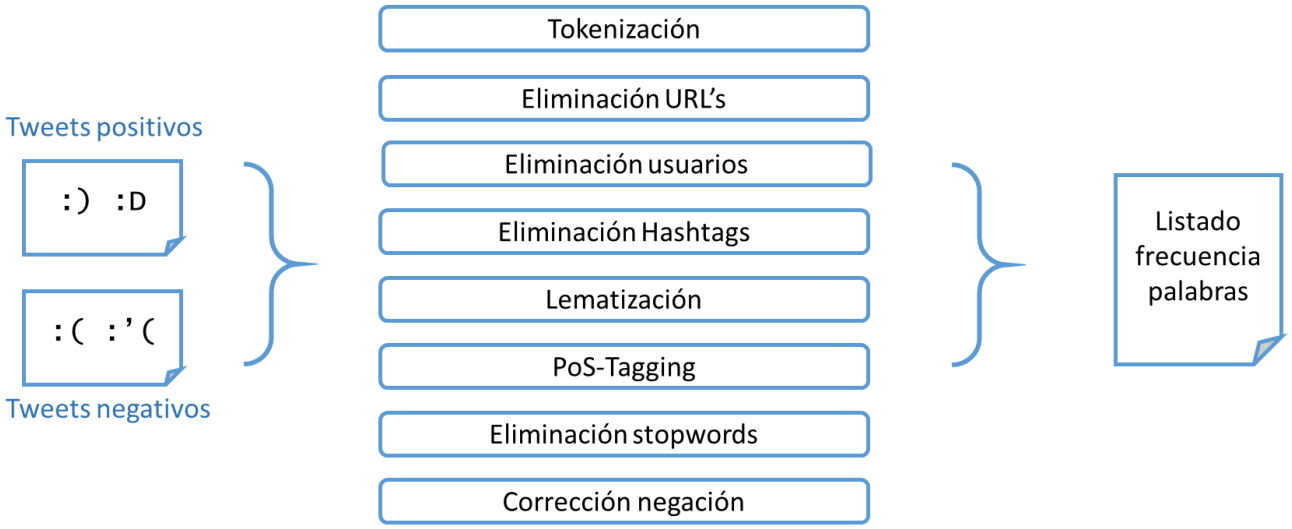


FIGURA 9: PROCESAMIENTO COMENTARIOS TWITTER
FUENTE: ELABORACIÓN PROPIA

5.2.2 Ajuste de polaridad

Luego de tener las frecuencias de cada palabra en el corpus de comentarios positivos y la frecuencia en el corpus de comentarios negativos, se calcula la polaridad de cada palabra.

Las reglas para el cálculo de la polaridad por PMI es que la frecuencia debe ser igual o mayor a 5 veces en cada uno de los corpus. Dentro de los datos utilizados, se observan palabras que no se encuentran en ambos corpus, por lo que se ha determinado al igual que en trabajos anteriores, como en [32] y [59], un ajuste al cálculo de polaridad para evitar la división por cero.

El cálculo de la polaridad de una palabra p se ha determinado con la siguiente fórmula:

$$Polaridad\ (p) = \ PMI\ (p, positivo) - \ PMI\ (p, negativo)$$

La fórmula PMI que se propuso en la sección 4.2.3 se ha ajustado como sigue:

$$PMI(p, positivo)_{ajustada} = log_2 \frac{Frec(w, positivo) * N}{Frec(w) * Frec(positivo)}$$

Donde:

- $Frec(p, positivo)$ = número de veces que aparece p en tweets positivos, considerando el ajuste. Calculado como la frecuencia de la palabra p en comentarios positivos + el ajuste r.

$$Frec(p, positvo) = frecuencia_p + r$$

- $Frec(p)$ = frecuencia total que aparece el término p en el corpus. Calculado como la suma de $Frec(p, positivo)$ y $Frec(p, negativo)$.

$$Frec(p) = Frec(p, positivo) + Frec(p, negativo)$$

- $Frec(positivo)$ = número total de tokens en tweets positivos. Calculado como el número total de palabras dentro del lexicón.

- N = total de tokens dentro del corpus. Se ha calculado como la suma de las frecuencias totales de todas las palabras dentro del corpus.

$$N = \sum_{p\ en\ lexicón} Frec(p)$$

Para determinar el ajuste r, se ha analizado el comportamiento propiamente tal que se tiene dentro del lexicón, considerando que las limitantes para este ajuste es que debe ser cercano a 0 para no influir significativamente en las frecuencias de las palabras y no mayor a 1, ya que un ajuste de 1 se considera la agregación de una palabra extra al corpus.

La distribución de polaridad en las palabras sin considerar ajuste corresponde a una distribución normal centrada en 0, como se muestra en la Figura 10. Al incluir un ajuste, la distribución se ve afectada directamente en las colas, por lo que es necesario tener en cuenta no perjudicar significativamente el tipo de distribución de los datos.

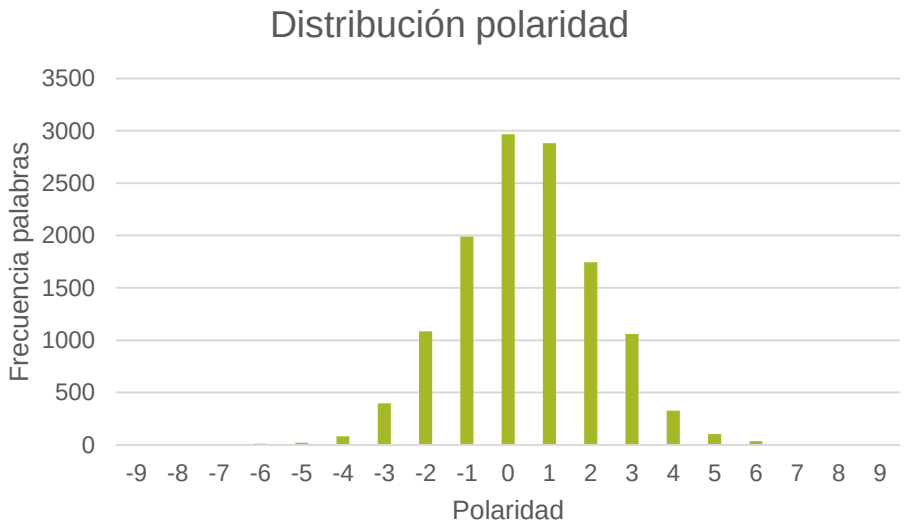


FIGURA 10: DISTRIBUCIÓN POLARIDAD
FUENTE: ELABORACIÓN PROPIA

Teniendo en cuenta el comportamiento de la polaridad según el valor de ajuste y no afectar significativamente su distribución, se ha determinado que el valor de ajuste debe considerar los siguientes criterios:

- Valor cercano a 0, menor a 1.
- El rango de polaridades no debiese variar negativamente (diminución rango).
- Para el rango de polaridades se tolera una variación del 10% (aumento de rango).
- Valor más pequeño que no varíe alguno de los límites de la polaridad.

A continuación en la Figura 11 podemos observar el comportamiento para 5 casos de valores de ajuste: 0.01, 0.1, 0.2, 0.4 y 0.6.

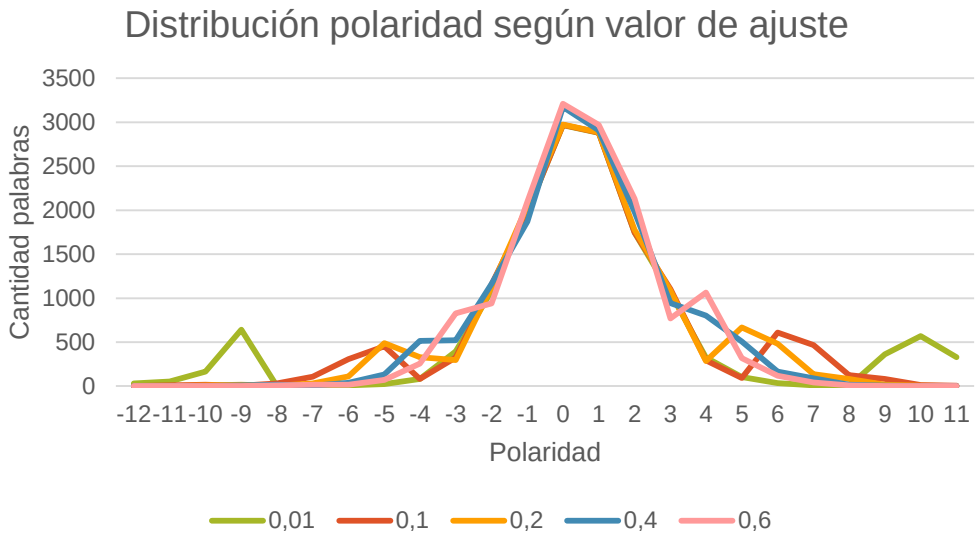


FIGURA 11: DISTRIBUCIÓN POLARIDAD SEGÚN VALOR DE AJUSTE
FUENTE: ELABORACIÓN PROPIA

Respetando los criterios mencionados anteriormente el ajuste r determinado para este lexicón es $0{,}4$, dejando una distribución de polaridad como se muestra en la Figura 12.

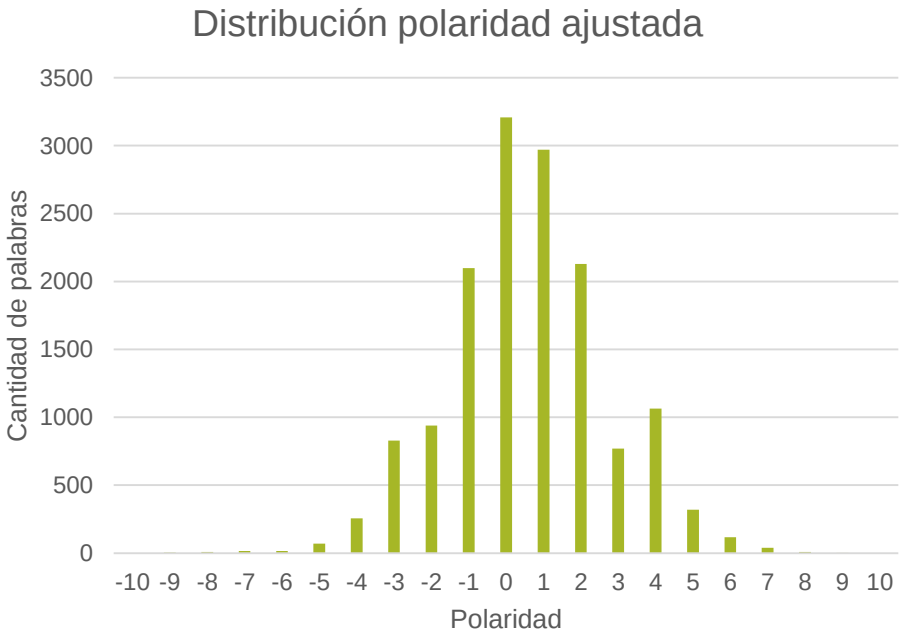


FIGURA 12: DISTRIBUCIÓN DE POLARIDADES AJUSTADAS
FUENTE: ELABORACIÓN PROPIA

5.2.3 Tamaño lexicón

La construcción del corpus se realizó tomando en consideración los tweets recolectados diariamente, por lo que el grupo de palabras que eran ingresadas al lexicón se fueron almacenando según su frecuencia diaria.

A continuación, en la Figura 13 se muestra el aumento en cantidad de palabras dentro del lexicón durante 2 semanas:

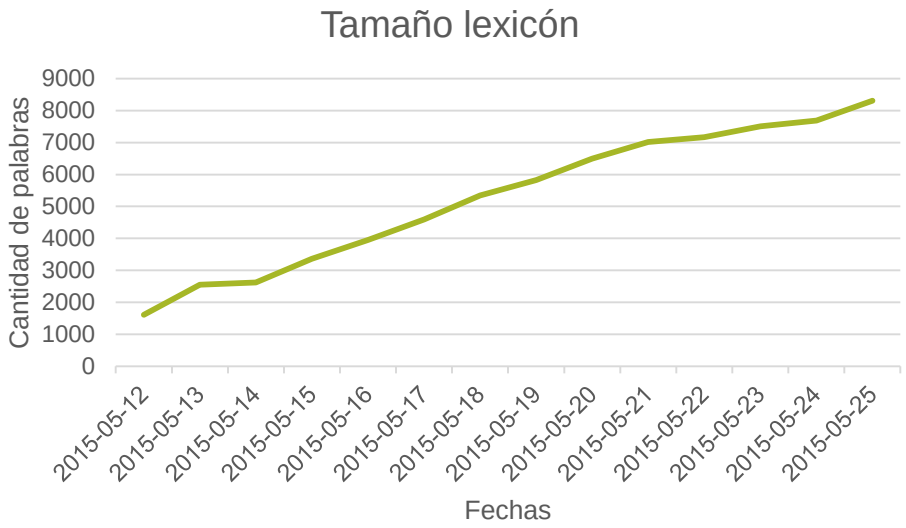


FIGURA 13: CRECIMIENTO LEXICÓN DE OPINIÓN
FUENTE: ELABORACIÓN PROPIA

Viendo el crecimiento que tiene el lexicón de opinión al ir agregando más comentarios de Twitter, surge el problema de cuántos comentarios finalmente utilizar para tener un lexicón de opinión lo suficientemente representativo de las palabras que ocupan los chilenos en Twitter.

Considerando que el tamaño de este lexicón puede aumentar cada día, por palabras nuevas, inventadas, mal escritas, etc., es necesario fijar un límite. Como para el cálculo de polaridad se necesita el dato de la frecuencia de estas palabras en comentarios positivos y en negativos, el límite a considerar debe tomar en cuenta una estabilización en cuanto a frecuencias de estas palabras en comentarios positivos y negativos.

En la Figura 14 se muestra el comportamiento para 2 semanas del promedio de frecuencias de las palabras dentro del lexicón en comentarios positivos y negativos.

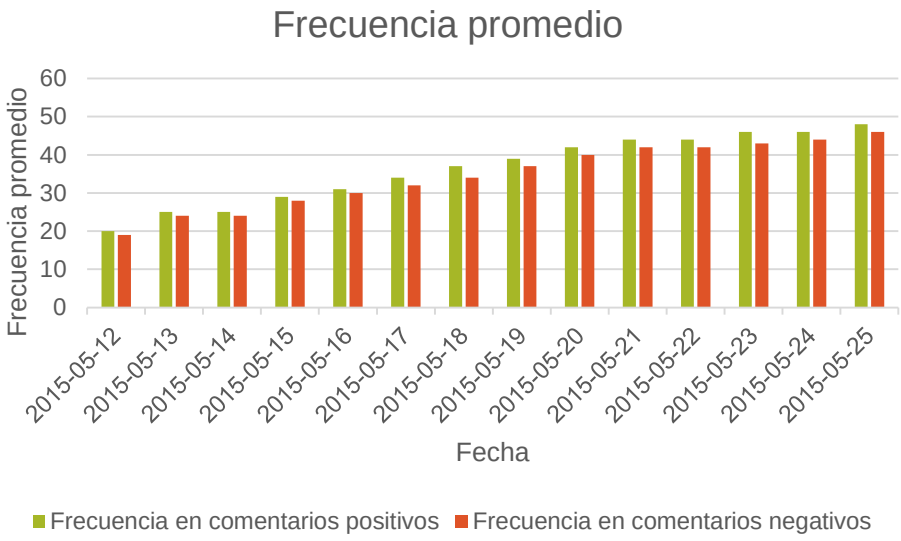


FIGURA 14: FRECUENCIA PROMEDIO
FUENTE: ELABORACIÓN PROPIA

Cómo métrica para evaluar el límite que debe tener el lexicón de opinión se han considerado los siguientes 3 indicadores:

- **Ganancia palabras**: considerada como la ganancia diaria de palabras por aumento en la cantidad de ellas, calculada como el porcentaje de palabras extra agregadas al lexicón sobre las palabras que se tienen.

- **Ganancia frecuencia positiva:** considerada como la ganancia en frecuencia promedio de palabras en comentarios positivos, calculada como el porcentaje de aumento en frecuencia promedio, sobre la frecuencia promedio de palabras en comentarios positivos que se tiene.

- **Ganancia frecuencia negativa:** considerada como la ganancia en frecuencia promedio de palabras en comentarios negativos, que es calculada de forma análoga al indicador anterior, considerando la frecuencia promedio de palabras en comentarios negativos.

Observando en la Tabla 9 los indicadores propuestos anteriormente para las primeras 2 semanas de análisis, se ve una tendencia a la estabilización, por lo que, con estos 3 indicadores, es suficiente para determinar el tamaño límite del lexicón de opinión.

Fecha	Ganancia palabras	Ganancia Positiva	Ganancia Negativa
12-05-2015	-	-	-
13-05-2015	59%	25%	26%
14-05-2015	3%	0%	0%
15-05-2015	28%	16%	17%
16-05-2015	18%	7%	7%
17-05-2015	16%	10%	7%
18-05-2015	16%	9%	6%
19-05-2015	9%	5%	9%
20-05-2015	12%	8%	8%
21-05-2015	8%	5%	5%
22-05-2015	2%	0%	0%
23-05-2015	5%	5%	2%
24-05-2015	2%	0%	2%

TABLA 9: INDICADORES TAMAÑO LEXICÓN DE OPINIÓN
FUENTE: ELABORACIÓN PROPIA

Como límite para determinar el tamaño del lexicón de opinión se ha decidido que los 3 indicadores propuestos tengan un valor menor a 1%, logrado luego de la recolección de comentarios de 33 días. En este caso, la ganancia por palabras es cerca del 0,01% y la ganancia en frecuencia en comentarios positivos y negativos es menor al 0,01%.

A continuación en la Figura 15 se puede apreciar el comportamiento que ha tenido el lexicón de opinión a medida que se agregan una mayor cantidad de comentarios diariamente, donde se aprecia el aumento en cantidad de palabras por agregación de comentarios y la frecuencia promedio de palabras en comentarios positivos y negativos.

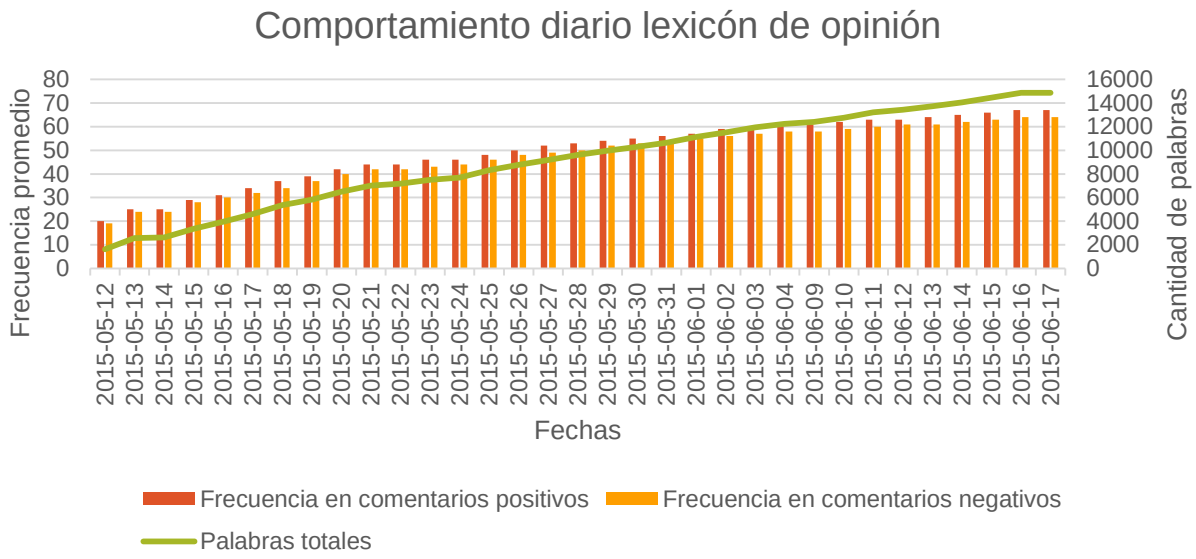


FIGURA 15: COMPORTAMIENTO DIARIO DE LEXICÓN DE OPINIÓN
FUENTE: ELABORACIÓN PROPIA

Observando el comportamiento que tiene el lexicón de opinión durante esos días, se aprecia una tendencia a la estabilización en cuanto a frecuencia de palabras y en tamaño del lexicón.

Finalmente, el lexicón de opinión utilizó comentarios de 33 días, comprendidos entre el 12 de mayo de 2015 y 17 de junio de 2015. Sin considerar los días 5, 6, 7 y 8 de junio donde no se almacenaron tweets.

El total de palabras únicas presentes en el lexicón de opinión es de 14.861, en que cada palabra dentro del lexicón presenta una frecuencia promedio de 67 en comentarios positivos y una frecuencia de 65 en comentarios negativos.

Considerando que en [60] se estudia que los adultos pueden usar hasta 25.000 palabras y que no todas corresponden a palabras de opinión, la cantidad de presentes en el lexicón construido es suficiente para representar las palabras de opinión utilizadas dentro de la plataforma de Twitter.

5.2.4 Resultados lexicón

El lexicón construido se compone de un listado de palabras, con su frecuencia asociada en el corpus con comentarios positivos y en el corpus con comentarios negativos, además de la polaridad calculada para cada una de ellas.

A continuación en la Figura 16 se presentan las palabras con mayor frecuencia en el corpus, se muestra una representación visual de las palabras contenidas de forma en que aquellas con mayor frecuencia en el corpus se encuentran diagramadas de mayor tamaño:

FIGURA 16: PALABRAS CON MAYOR FRECUENCIA
FUENTE: ELABORACIÓN PROPIA

También en la Tabla 10 se observan las palabras con mayor frecuencia en comentarios positivos y negativos, respectivamente. Se destacan los verbos “ser”, “tener”, “ver”, “haber” que son altamente utilizados dentro del corpus, en sus distintas formas. Se debe notar, que las palabras “gracia”, “seguir”, “follow” se encuentran asociadas a comentarios como “Gracias por seguirme” o “Gracias, te devuelvo el follow”, altamente utilizadas en la plataforma de twitter.

Palabras con mayor frecuencia en corpus

	Frecuencias en corpus
no	79.843
ser	76.216
tener	31.268
estar	28.935
gracia	26.828
ir	26.101
hacer	25.314
bueno	23.093
ver	23.079
seguir	23.045
querer	21.861
haber	20.101
ya	16.995
poder	15.989
día	14.002
dar	13.082
decir	12.226
más	11.572
muy	9.891
follow	9.783
devolver	9.638
breve	9.340
hoy	9.270

Palabras con mayor frecuencia en comentarios positivos

	Frecuencias	
	Comentarios positivos	Comentarios negativos
ser	37.655	38.561
gracia	25.233	1.595
no	23.707	56.136
bueno	19.161	3.932
seguir	18.946	4.099
tener	14.633	16.635
estar	12.673	16.262
ver	12.182	10.897
ir	11.135	14.966
poder	10.795	5.194
haber	10.299	9.802
querer	10.254	11.607
día	9.700	4.302
follow	9.654	129
devolver	9.411	227
breve	9.320	20
hacer	8.806	16.508
ya	7.444	9.551
dar	6.823	6.259
más	5.836	5.736
saludo	5.810	1.211
muy	5.766	4.125
mejor	5.727	2.088

Palabras con mayor frecuencia en comentarios negativos

	Frecuencias	
	Comentarios positivos	Comentarios negativos
no	23.707	56.136
ser	37.655	38.561
tener	14.633	16.635
hacer	8.806	16.508
estar	12.673	16.262
ir	11.135	14.966
querer	10.254	11.607
ver	12.182	10.897
haber	10.299	9.802
ya	7.444	9.551
decir	4.802	7.424
dar	6.823	6.259
más	5.836	5.736
pasar	3.736	5.403
poder	10.795	5.194
llegar	3.721	5.114
siempre	3.545	5.017
llorar	416	4.430
ahora	3.668	4.369
día	9.700	4.302
muy	5.766	4.125
seguir	18.946	4.099
hoy	5.219	4.051

TABLA 10: FRECUENCIAS EN COMENTARIOS

FUENTE: ELABORACIÓN PROPIA

FIGURA 17: PALABRAS CON MAYOR FRECUENCIA EN COMENTARIOS POSITIVOS
FUENTE: ELABORACIÓN PROPIA

FIGURA 18: PALABRAS CON MAYOR FRECUENCIA EN COMENTARIOS NEGATIVOS
FUENTE: ELABORACIÓN PROPIA

Para tener una visualización de los resultados obtenidos se elaboraron las siguientes figuras que representan palabras ordenadas según su magnitud de polaridad. En Figura 19 se aprecian las palabras con polaridad positiva y en la Figura 20 se aprecian las palabras con polaridad negativa.

FIGURA 19: PALABRAS CON POLARIDAD POSITIVAS
FUENTE: ELABORACIÓN PROPIA

Palabras con alta polaridad positiva	
	Polaridad
breve	9
eespero	9
karolconcompañia	8
epico	8
maar	8
ntilde	8
lúnes	8
feromonas	8
lisboa	8
mejicano	8
diccionario	8
tourterrallisboa	8
elsueñodemartinamg	7
ferraristas	7
sonries	7
regaloo	7
perfíl	7
composer	7
pisci	7
quearon	7
firescue	7
rcn	7
listado	7

Palabras con alta polaridad negativa	
	Polaridad
damasiado	-10
festi	-10
ñskdsldk	-10
gameloft	-10
fuerzajorge	-9
vuelaaltojorge	-9
jorgevilches	-9
qepdjorgevilches	-9
antártico	-9
answer	-9
acojonar	-9
gandalf	-9
interino	-9
nicohablamekawaii	-8
astucia	-8
yyyooooooooo	-8
qdep	-8
fifiaparece	-8
prepareis	-8
saludín	-8
horrible	-8
rcn	-7
listado	-7

TABLA 11: PALABRAS CON ALTA POLARIDAD
FUENTE: ELABORACIÓN PROPIA

Del primer análisis exploratorio de palabras dentro del lexicón junto al análisis realizado en la sección 5.1.4, se puede apreciar que aquellas palabras que presentan mayor nivel de polaridad, ya sea positivo o negativo, consisten en palabras que se encuentran en comentarios spam o altamente repetidos en el corpus lingüístico utilizado.

Dentro de las palabras positivas que pueden ser asociadas a comentarios repetidos o spam, podemos encontrar “ferraristas”, “feromonas”, “tourterrallisboa”, “breve”. Mientras que podemos encontrar palabras negativas con un alto nivel de polaridad asociadas a comentarios repetidos o spam como: “quedjorgevilches”, “vueltaaltojorge”, “nicohablamekawaii”.

FIGURA 20: PALABRAS CON POLARIDAD NEGATIVA
FUENTE: ELABORACIÓN PROPIA

Estas representaciones gráficas de la polaridad de las palabras presentes en el corpus, ayudan a visualizar la gran cantidad de palabras que a simple viste no parecen tener polaridad, pero que la presentan al pertenecer a comentarios con un alto nivel de retweets, o estar en comentarios altamente repetidos.

Capítulo 6

6 Evaluación y discusiones

En el presente capítulo se muestran los resultados de la evaluación realizada del lexicón construido que consta de 2 evaluaciones. La primera, de la calidad del recurso construido para elaborar el lexicón de opinión, considerando 2 casos de evaluación. Y una segunda evaluación del desempeño del lexicón en comentarios de Twitter utilizados en el WIC.

6.1 Evaluación del corpus construido

El propósito de esta evaluación es determinar si el uso de emoticones para construir un corpus lingüístico donde se le identifique su polaridad es suficiente para lograr el recurso base para la construcción del lexicón. Se tomó en consideración como base los resultados de [61], [62] y [40] que representan algunas investigaciones anteriores que ocuparon un modelo parecido y las investigaciones en las cuales se basó esta propuesta.

Para la evaluación del corpus construido se utilizó una muestra significativa de los comentarios utilizados. Dentro del corpus construido se tienen 432.191 comentarios de Twitter donde el 50% de los comentarios corresponde a comentarios clasificados como positivos, y el 50% restante como comentarios negativos como se muestra en la Tabla 12 a continuación:

Comentarios positivos	216.064
Comentarios negativos	216.127
Total comentarios	432.191

TABLA 12: COMENTARIOS CORPUS
FUENTE: ELABORACIÓN PROPIA

Se tomó una muestra de 1.064 comentarios dentro del corpus para evaluar, que corresponden a la muestra que maximiza su tamaño y que es estadísticamente significativa a una confianza del 95%, tolerando un error del 3%.

Para la evaluación se utilizó el criterio de expertos en el tema, considerando a expertos como aquellos que tienen una cuenta activa en Twitter y que presentan residencia en Chile por más de 5 años.

Se utilizaron 10 evaluadores expertos que no tenían conocimiento del trabajo realizado en este proyecto, donde cada uno evaluó un set de 141 comentarios según su polaridad en positivo, negativo o indeterminado (aquellos comentarios neutros, comentarios que no se podía identificar polaridad clara o que no presentaban idioma español).

Teniendo en cuenta este tipo de evaluación, según la clasificación realizada, los resultados que se obtuvieron se presentan en la Tabla 13.

	Comentarios	
	Positivos	**Negativos**
Precisión	58%	65%
Exhaustividad	93%	95%
Estadístico F	71%	77%

TABLA 13: RESULTADOS PRIMERA EVALUACIÓN
FUENTE: ELABORACIÓN PROPIA

La $Exactitud$ $(Accuracy)$ de la clasificación es de 68%.

		Clasificación por emoticones		
		Positivo	**Negativo**	TOTAL
Clasificación por expertos	**Positivo**	**407**	33	440
	Negativo	24	**456**	480
	Indeterminado	275	212	487
	TOTAL	706	701	**1407**

TABLA 14: CLASIFICACIÓN PRIMERA EVALUACIÓN
FUENTE: ELABORACIÓN PROPIA

Como resultados de investigaciones anteriores tenemos que la utilización de emoticones para [61] la $exactitud$ es de 61,4% y los resultados para la $exactitud$ en [40] es de 81,7%. En modelos similares donde se emplearon el uso de emoticones en corpus en español podemos ver que la exactitud por ejemplo en [62] es de 48% utilizando un modelo que recurre a un enfoque basado en aprendizaje automático, que no es exactamente lo realizado en este trabajo pero se tomó como referencia en cuanto a los resultados que se pueden obtener en clasificaciones de este tipo de datos, donde la subjetividad juega un rol fundamental.

De estos primeros resultados, se observa que en los comentarios con emoticones negativos es más probable que sean efectivamente comentarios negativos, a diferencia de los comentarios positivos donde la probabilidad de que sean efectivamente comentarios positivos disminuye, pensando en que se obtuvo una precisión de 65% en comentarios con emoticones negativos, a diferencia del 58% en comentarios positivos.

Mirando los altos niveles de exhaustividad podemos decir que es altamente probable que un comentario que tenga polaridad y tenga algún emoticón presente, la polaridad del emoticón sea suficiente para determinar la polaridad del tweet, por lo que el clasificador logra encontrar la polaridad dentro de estos comentarios.

Haciendo un análisis de los resultados obtenidos, se puede observar que los comentarios indeterminados dentro del corpus representan cerca de un tercio de éste y no aportan mayor valor al análisis porque su polaridad no puede ser identificada claramente por expertos y el clasificador construido se realizó para lograr capturar la polaridad que tienen los comentarios.

Tomando en consideración que los comentarios indeterminados no agregan mayor valor al análisis se realizó una segunda evaluación utilizando sólo aquellos comentarios que presentaban polaridad positiva y negativa, evitando por tanto tener distintas clases de clasificación.

Se analizaron 920 comentarios que presentaban polaridad clara y que fueron clasificados por expertos como positivos y negativos. En Tabla 15 se muestran los resultados de esta evaluación de acuerdo a la matriz de confusión:

		Clasificación por emoticones		
		Positivo	**Negativo**	TOTAL
Clasificación por expertos	**Positivo**	**407**	33	440
	Negativo	24	**456**	480
	TOTAL	431	489	**920**

TABLA 15: CLASIFICACIÓN CORPUS LINGÜÍSTICO
FUENTE: ELABORACIÓN PROPIA

Considerando esta segunda evaluación donde sólo fueron considerados los comentarios que presentaran polaridad, se tiene una $Exactitud\ (accuracy)$ del 94% y las métricas presentadas en la Tabla 16.

Métricas	**Precisión** (Precision)	**Exhaustividad** (Recall)	**Estadístico F** (F-measure)
Positivos	94%	93%	93%
Negativos	93%	95%	94%

TABLA 16: MÉTRICAS DE EVALUACIÓN
FUENTE: ELABORACIÓN PROPIA

Observando los resultados obtenidos y al evaluarlo con un corpus que presenta polaridad conocida, la $precisión$ que tiene el clasificador aumenta significativamente, mientras que la $exhaustividad$ en ambos corpus representaba una alta tasa. Resulta natural que el estadístico F también se vea aumentado ya que tiene directa relación con las medidas de $precisión\ y\ exhaustividad$.

Estos resultados muestran que si se evalúa este tipo de clasificación en un corpus donde claramente se aprecie la polaridad de los comentarios, la utilización de emoticones representa un indicador clave para determinar la polaridad de los comentarios. Es necesario considerar que los valores mostrados en la Tabla 16 corresponden a la evaluación en un corpus donde la polaridad era claramente expresada.

Como estamos frente a una evaluación que presenta una calificación subjetiva de los comentarios, es necesario también considerar el grado de acuerdo que tienen los evaluadores respecto al tema. Para determinar el grado de concordancia que tienen los observadores se utilizó la medida estadística $Kappa$.

Para ello se evaluaron por 2 expertos distintos una muestra de los comentarios evaluados anteriormente. La muestra utilizada corresponde a la muestra que maximiza su tamaño, que es estadísticamente significativa a una confianza del 95% y tolerando un error del 5%.

La muestra de tamaño 285 fue dividida en 5 grupos de comentarios para evitar que la clasificación demandara mucho tiempo para los expertos. Cada grupo fue evaluado por 2 expertos y el resultado de esa evaluación se muestra en la Tabla 17. La muestra contiene tweets aleatorios de los 33 días con que fue construido el corpus, los comentarios del corpus corresponden en igual cantidad a comentarios con emoticones positivos y negativos, pero no necesariamente las muestras que se hicieron de éstos contienen en igual cantidad comentarios positivos y negativos.

Juicio expertos	Positivo	Negativo	Indeterminado	TOTAL
Positivo	61	5	31	97
Negativo	1	60	36	97
Indeterminado	14	17	60	91
TOTAL	76	82	127	285

TABLA 17: CÁLCULO MEDIDA KAPPA
FUENTE: ELABORACIÓN PROPIA

El valor de kappa para esta evaluación es de 0.45, si tenemos que en cuenta que los valores aceptables de $kappa$ en una clasificación subjetiva se encuentran entre 0.4 y 0.75 como se plantea en [63], la clasificación de la muestra que se evalúo tiene valores aceptables de concordancia, pero también refleja que la evaluación de comentarios con información subjetiva es compleja, aún para expertos en el tema. Si analizamos kappa para cada grupo clasificado (0.53, 0.41, 0.41, 0.46 y 0.53), tenemos también un valor suficiente, mayor a 0.4.

Dado que los expertos utilizados conocían el lenguaje empleado en la plataforma de Twitter, pero no tenían un conocimiento avanzado respecto a análisis de opinión, específicamente en análisis de polaridad, los resultados de esta evaluación pueden tener un sesgo producto de la poca experiencia en el tema. Aun considerando esto, el índice Kappa, muestra que la clasificación realizada es suficiente para poder evaluar el corpus lingüístico construido. Para obtener mejores resultados en cuanto a concordancia de los evaluadores, hizo falta un periodo de entrenamiento donde los evaluadores fueran capacitados para la tarea que desempañaban.

6.2 Validación en proyecto OpinionZoom

Para validar el resultado de este trabajo, se ha implementado el lexicón de opinión construido en el sistema utilizado en el proyecto "OpinionZoom". Para evaluar el cambio producido en el sistema con el lexicón, se construyó un corpus de 475 comentarios que se encuentran en la base de datos con la cual trabaja el centro de investigación WIC.

El corpus utilizado corresponde a tweets entre el 20 y 27 de julio de 2015, extraídos considerando la presencia de *hashtags* para discriminar aquellos que expresaban alguna polaridad. Se extrajeron aquellos comentarios que presentaban en sus hashtags palabras con polaridad conocida, estas palabras corresponden a las palabras asociadas a positivo y negativo en el tesauro de Roget y que se encuentran en el anexo 1.

Luego de extraer los comentarios, se consideró la eliminación de aquellos comentarios que presentaran desacuerdo en cuanto a la polaridad presente en ellos según expertos de Twitter. Considerando que la clasificación por expertos no demandara una gran cantidad de tiempo, en el corpus se consideraron 1723 comentarios doblemente evaluados por expertos, eliminando los comentarios que no presentaban concordancia entre ambos evaluadores.

El corpus construido presenta 239 comentarios positivos y 234 comentarios negativos que fueron doblemente evaluados como tal por expertos de Twitter. A estos comentarios se les aplicó una clasificación de polaridad con la API diseñada en el WIC determinando si eran comentarios positivos o negativos.

Esta API actualmente considera como base un lexicón de opinión de uso académico y elaborado a partir del español utilizado en España, los resultados de la clasificación dentro del sistema actual se observan en la Tabla 18, se muestran además los resultados al utilizar el lexicón de opinión construido en este trabajo.

		Clasificación en API OpinionZoom				
		Positivos		**Negativos**		
		Lexicón Actual	Lexicón Construido	Lexicón Actual	Lexicón Construido	Total
Clasificación por expertos	**Positivos**	**148**	**171 (+9,6%)**	91	68	239
	Negativos	99	82	**135**	**152 (+7,3%)**	234

TABLA 18: RESULTADOS IMPLEMENTACIÓN EN SISTEMA DE CLASIFICACIÓN
FUENTE: ELABORACIÓN PROPIA

Se observa de los resultados obtenidos que tanto en comentarios positivos como negativos la aplicación del lexicón construido en este trabajo encuentra una mayor cantidad de comentarios para cada clase. En el caso de comentarios positivos encuentra un 10% más de comentarios y en el caso de los comentarios negativos se encuentra un 7% más. Considerando estos porcentajes ya se aprecia un beneficio por la utilización del recurso construido en este trabajo.

Analizando con mayor profundidad los resultados en cuanto a las métricas relevantes para estos sistemas de clasificación, podemos observar en la Tabla 19 los resultados obtenidos, donde se aprecia una mejora significativa para cada una de las métricas de desempeño.

Se ha calculado la $Exactitud\ (Accuracy)$ del sistema considerando la utilización de ambos lexicones y se han encontrado los siguientes resultados:

- $Exactitud\ (Accuracy)$ sistema con lexicón actual: 60%.
- $Exactitud\ (Accuracy)$ sistema con el lexicón construido: 68%.

	Precisión (Precision)		**Exhaustividad (Recall)**		**Estadístico F (F-measure)**	
	Actual	Construido	Actual	Construido	Actual	Construido
Positivos	60%	68%	62%	72%	61%	70%
Negativos	60%	69%	58%	65%	59%	67%

TABLA 19: MÉTRICAS DESEMPEÑO CON DIFERENTES LEXICONES
FUENTE: ELABORACIÓN PROPIA

Considerando que las métricas de desempeño presentadas en la Tabla 18 se ha calculado el aumento en el rendimiento del sistema al utilizar el lexicón de opinión construido que se muestran en la Tabla 20.

	Precisión (Precision)	**Exhaustividad (Recall)**	**Estadístico F (F-measure)**
Positivos	+8%	+10%	+9%
Negativos	+9%	+7%	+8%

TABLA 20: BENEFICIOS UTILIZACIÓN LEXICÓN
FUENTE: ELABORACIÓN PROPIA

El aumento en estas métricas de desempeño nos muestra:

- Se ha aumentado en un 8% el porcentaje de datos que son clasificados correctamente por el sistema.
- Se ha aumentado en un 8% y 9% el porcentaje de datos clasificados como positivos y negativos respectivamente que son efectivamente positivos y negativos.
- Se ha aumentado en un 10% y 7% el porcentaje de comentarios positivos y negativos respectivamente que fueron clasificados como tal.

Entendiendo que la medida de $Precisión$ considera que tan precisa o concisa es la evaluación y la medida de $Exhaustividad$ indica que tan completa es la evaluación, el $estadístico\ F$ muestra un aumento del 9% y 8% en comentarios positivos y negativos respectivamente caracterizando con un único valor la precisión y cobertura en cada tipo de comentario.

Una desventaja que tiene la utilización de este lexicón de opinión es el tiempo de procesamiento del sistema, ya que producto de la cantidad de palabras dentro de éste ha producido un aumento en el tiempo de procesamiento que se tiene. La medición se realizó en base al tiempo de procesamiento de los comentarios dentro del corpus utilizado para la validación según los distintos lexicones utilizados.

El tiempo de procesamiento promedio de comentarios para los distintos lexicones se presenta en la Tabla 21 mostrada a continuación:

Tiempo (ms)	Actual	Construido
Positivos	12.461	12.957
Negativos	17.747	18.348

TABLA 21: COMPARACIÓN TIEMPOS DE PROCESAMIENTO
FUENTE: ELABORACIÓN PROPIA

Si tomamos en cuenta que en una hora pueden procesarse con el lexicón construido menos cantidad de tweets que con el lexicón que actualmente se utiliza, el beneficio en la precisión de tweets considerando los resultados en la Tabla 20 es considerablemente mayor; para comentarios positivos en una hora existe una diferencia de 11 tweets que dejan de procesarse, pero se encuentran 22 comentarios más que con el lexicón anterior. En el caso de negativos, por tiempo se dejan de procesar 7 comentarios, pero se ganan por el aumento de precisión 18 comentarios. Por lo que la utilización del lexicón construido, tomando en cuenta la desventaja de tiempo de procesamiento es un beneficio considerable para el proyecto “OpinionZoom”.

Capítulo 7

7 Conclusiones

En el presente capítulo se analizan las conclusiones obtenidas del trabajo realizado, para finalmente proponer recomendaciones y trabajos futuros dentro del área.

7.1 Conclusiones generales

En este trabajo se propuso una metodología para el diseño y construcción de un lexicón de opinión aplicado a las características del español utilizado en Chile. El lexicón de opinión construido fue implementado para formar parte del sistema de análisis de opiniones que se lleva a cabo en el proyecto “OpinionZoom” dentro del Centro de Investigación de Inteligencia Web (WIC).

En cuanto a los objetivos específicos planteados para el desarrollo de este trabajo se explica a continuación el cumplimiento de cada uno de ellos:

- **Estudio del estado del arte de “Opinion Mining”:** dentro del marco teórico de este trabajo se estudian los diferentes sistemas de análisis de opinión y modelos ocupados dentro de la literatura para las diversas aplicaciones que se tienen, así como también la red social, Twitter, en la cual se aplica el análisis de opiniones dentro del proyecto. El capítulo 2 presenta en tanto, el contexto en el cual se enmarca el trabajo realizado.
- **Implementación de metodología:** en el capítulo 3, se estudian los 3 enfoques metodológicos planteados en la literatura para la generación de lexicones de opinión y algunas de las implementaciones que existen para diversos idiomas. Para este análisis se detectaron algunos pros y contras de cada uno de estos enfoques. Considerando este análisis y los requerimientos para el proyecto “OpinionZoom” se desarrolló en el capítulo 4 el diseño del lexicón de opinión, proponiendo la metodología y el proceso para llevarlo a cabo.
- **Construcción del lexicón de opinión**: en cuanto a la elaboración del lexicón en el capítulo 5 se muestra el proceso llevado a cabo para la construcción de éste, así como también se detalla el proceso llevado a cabo para la construcción del recurso clave para la realización de lexicones de opinión en base al enfoque basado en un corpus lingüístico, enfoque metodológico determinado para el desarrollo este trabajo.
- **Validación en proyecto “OpinionZoom”:** finalmente el último objetivo específico de este trabajo fue realizado dentro del proyecto “OpinionZoom”, implementando el lexicón de opinión construido dentro del sistema de análisis de opinión utilizado en el centro de investigación.

En cuanto al desarrollo del trabajo, los enfoques metodológicos planteados en la literatura representan un simple acercamiento a la generación de lexicones, que depende en gran medida de la aplicación para la cual será utilizada. Teniendo en cuenta los requerimientos que tenía el proyecto, la mejor alternativa para la elaboración de un lexicón de opinión era utilizar un enfoque basado en un corpus lingüístico, que fue lograda gracias a la construcción de un corpus lingüístico de comentarios de Twitter en base a los emoticones presentes en ellos.

Del corpus lingüístico construido se puede apreciar el alto contenido de comentarios repetidos, y una gran cantidad de spam que posteriormente fue reflejada en el lexicón de opinión dado que éste se basaba en la frecuencia de palabras y, por tanto, palabras dentro de estos comentarios tenían un alto nivel de polaridad que no necesariamente corresponde al nivel adecuado. Podemos ver el caso particular del término “breve” que está asociado a un comentario bastante repetido dentro del corpus de comentarios positivos, “en breve te devuelvo follow :)” y que no representa necesariamente uno de los más altos niveles de polaridad positiva.

El uso de emoticones para identificar la polaridad de los comentarios obtiene buenos resultados si se combina con alguna técnica que identifique la presencia de polaridad o comentarios neutros dentro del corpus. Esto se puede ver analizando los resultados en un corpus con presencia de comentarios sin polaridad conocida o difícil de determinar, puesto que sin el contexto adecuado no se obtienen resultados suficientes, a diferencia de lo que es identificar la polaridad en base a emoticones en comentarios donde se sabe que existe presencia de polaridad.

Se observó que la identificación de polaridad en un corpus donde existe presencia de emoticones y en el que se puede identificar la existencia de polaridad, la utilización de emoticones para identificarla va a acertar en un alto porcentaje de casos. Cerca de un 95% de los casos en que existe un comentario con polaridad, el tipo de polaridad (positiva o negativa) se puede identificar en base a los emoticones presentes. Este hallazgo puede ser utilizado para complementar los sistemas de análisis de opinión en que el análisis sintáctico se perfeccione considerando los emoticones presentes y que éstos tengan relevancia dentro del análisis.

De la construcción del lexicón este trabajo entrega herramientas que facilitan su elaboración, considerando los problemas que puedan existir al generar nuevos lexicones, como es el caso del tamaño del lexicón, donde se entregaron algunos indicadores asociados para determinar la cantidad de palabras dentro del lexicón, así como también se consideró agregar palabras que no estuvieran en ambos tipo de comentarios, sino que podrían presentarse sólo en un tipo de comentarios.

Teniendo en cuenta que el lexicón construido propiamente tal, se pueden apreciar varios fenómenos, uno de ellos se da al observar las palabras que se encuentran en los extremos del lexicón, palabras que presentan una alta frecuencia en comentarios positivos, o negativos, principalmente asociadas a comentarios que se repiten bastante dentro del corpus y que dentro del lexicón generan bastante ruido. Estas palabras no entregan valor adicional a lexicón de opinión, por lo que en este caso particular pueden

ser eliminadas de éste. Si se generan nuevos lexicones con la metodología propuesta y se toma en consideración la eliminación de comentarios que se repitan considerablemente dentro del corpus, el problema de las palabras en los extremos no debería generar dificultades.

Otro fenómeno asociado al lexicón construido es la presencia de palabras que con un simple análisis de expertos no son consideradas con polaridad, dado que estas palabras son presentadas sin un contexto específico. Es por ejemplo el caso de “metrodesantiago” que corresponde a un medio de transporte público, que a simple vista puede verse como una palabra sin polaridad, pero si se analizan los comentarios que presentan asociado este término, se aprecia que en su mayoría corresponden a comentarios que reclaman por el mal servicio o problemas asociados a interrupciones de la red de metro durante los días en que fueron recolectados los comentarios del corpus.

Estos fenómenos, por tanto, afectan la evaluación del lexicón de opinión porque presentan palabras con polaridad asociadas a los comentarios propios del corpus lingüístico del cual fue elaborado. Luego, la evaluación debe ser considerada en cuanto al contexto de comentarios de donde nació el lexicón de opinión y no como palabras libres de contexto, puesto que los resultados no expresan la polaridad propia de la palabra en ese entorno determinado.

De la evaluación realizada, se puede apreciar que expertos considerados como usuarios de Twitter chilenos es suficiente para analizar comentarios que presentan polaridad evidente, sin embargo, al considerar comentarios donde la polaridad requiere mayor atención y análisis, los expertos requieren un mayor conocimiento de análisis de sentimientos, así como también son necesarios una mayor cantidad de recursos en cuanto a tiempo para obtener una evaluación más exhaustiva.

Con respecto a la validación del lexicón de opinión construido, ésta fue elaborada en base al sistema de análisis de opiniones que posee el centro de investigación, ya que de esta forma la evaluación considera las palabras dentro de su contexto pudiéndose determinar de mejor forma la polaridad de los comentarios y así evaluar si realmente el lexicón de opinión construido cumple o no el objetivo para el cual fue elaborado.

Este trabajo, por tanto, constituye una de las primeras elaboraciones de lexicones de opinión realizadas en base al lenguaje utilizado por chilenos dentro de la plataforma de Twitter. Si bien existían comentarios dentro del corpus que eran realizados por extranjeros, en gran parte representaban la terminología utilizada por chilenos en Twitter. Este lexicón de opinión representa además un recurso útil dentro del área, para continuar las investigaciones sobre análisis de opiniones dentro de la región. Teniendo en consideración que el área de la minería de opiniones es un área relativamente nueva, en que aún queda mucho espacio para investigación y que en general, los recursos en español son escasos; este lexicón de opinión representa un gran avance para poder realizar análisis con respecto a las características de la región y extraer de mejor forma conocimiento desde comentarios de Twitter de usuarios chilenos.

Dentro del proyecto “OpinionZoom”, el lexicón de opinión construido genera valor considerando que gracias a éste se puede tener una mayor precisión en el análisis de los datos y también se capturan un porcentaje superior de comentarios. Teniendo en cuenta que el lexicón de opinión utilizado en el centro de investigación presentaba una licencia sólo de uso académico, los servicios que se brindan en “OpinionZoom” en cuanto a polaridad no podían ser comercializados. Por tanto, gracias a la construcción de este lexicón de opinión el sistema de análisis de opiniones presenta posibilidades de tener un producto comercializable en la región.

La mayor precisión de datos que se logró con el lexicón dentro del proyecto “OpinionZoom” es una ventaja considerable, aun cuando el lexicón construido genera un procesamiento más lento de los comentarios. Esta ventaja se puede apreciar en que el proyecto podrá capturar de mejor forma la opinión pública, generando mayor ventaja competitiva, debido a que el sistema presenta más palabras de uso en Twitter que puedan representar polaridad en los comentarios.

7.2 Recomendaciones y trabajo futuro

- Respecto de la construcción del corpus, uno de los ajustes realizados fue eliminar comentarios sin presencia de idioma español, en primera instancia con herramientas posteriores a la extracción de comentarios desde la base de datos, pero se obtuvieron mejores resultados determinando el idioma con la información entregada desde la API de Twitter. Eliminar comentarios que no presenten idioma español es un filtro que debe realizarse dentro de la base de datos que tiene el centro, ya que este tipo de comentario genera bastante ruido dentro del lexicón.

- La implementación para la construcción del lexicón de opinión no toma en consideración la categoría gramatical para diferenciar que una misma palabra pueda tomar diferente polaridad si se ocupa por ejemplo como verbo o como sustantivo. La arquitectura presentada para la construcción permite incluir la diferenciación entre palabras con categorías gramaticales diferentes, por lo que se propone medir qué tan significativa es realizar esta diferenciación.

- Para la construcción del corpus lingüístico también se recomienda hacer una limpieza exhaustiva en cuanto a comentarios repetidos que pueden ser dirigidos a diferentes usuarios, pero que representan la misma información que no agrega valor a la construcción del lexicón, sino que se ensucian sus resultados.

- Considerando los buenos resultados que se obtienen determinando la polaridad de comentarios utilizando como indicador los emoticones presentes en él, se puede avanzar en las investigaciones que consideran el uso de *emojis* para la determinación de polaridad y así, tener una mayor cantidad de indicadores de

polaridad en los comentarios, que mejoren la precisión de los resultados y logren capturar más opiniones presentes.

- El trabajo realizado en este proyecto fue desarrollado en comentarios de Twitter, comentarios que presentaban características especiales en cuanto a largo de las oraciones. Uno de los desafíos para próximas reproducciones de este proceso, es lograr extraer lexicones desde otras plataformas. La metodología utilizada fue propuesta tomando en consideración que sería utilizada en la plataforma de Twitter, pero puede ser fácilmente replicable en otras plataformas si se tienen corpus lingüísticos clasificados según polaridad. Para clasificar comentarios de otras plataformas usando emoticones es necesario realizar mayor análisis de la estructura de éstos.

- Otro desafío que queda del trabajo realizado es lograr tener lexicones de opinión para diversos dominios, si se logra ocupando la metodología propuesta en este trabajo es necesario construir distintos corpus clasificados para cada uno de los dominios requeridos. Si tomamos, por ejemplo, el análisis de opinión que se pueda realizar respecto a comentarios deportivos, se puede obtener un lexicón de opinión que represente las características de los chilenos al opinar sobre deporte, elaborado a partir un corpus lingüístico que tenga únicamente comentarios deportivos.

- Este trabajo fue desarrollado con un enfoque basado en un corpus lingüístico, pero puede ser complementado con otras técnicas como lo es por ejemplo un enfoque manual que ayude a corregir errores en algunas palabras presentes en el lexicón de opinión y con esto disminuir notablemente el desarrollo de un lexicón de opinión de forma netamente manual.

Bibliografía

[1] W3C, « About W3C: World Wide Web Consortium,» [En línea]. Available: http://www.w3.org/Help/#funds. [Último acceso: 28 Diciembre 2015].

[2] F. Ponce de León, Uso de la ingeniería de negocios en diseño e implementación, Santiago de Chile: Universidad de Chile - Facultad de Ciencias Físicas y Matemáticas, 2015.

[3] Internet world stats, «Internet world users by languaje,» Internet world stats, [En línea]. Available: http://www.internetworldstats.com/stats7.htm. [Último acceso: 28 Diciembre 2015].

[4] E. Constantinides and S. J. Fountain, «Web 2.0: Conceptual foundations and marketing issues,» *Journal of direct, data and digital marketing practice,* vol. 9, nº 3, pp. 231-244, 2008.

[5] S. Murugesan, "Understanding Web 2.0.," *IT professional,* vol. 9, no. 4, pp. 34-41, 2007.

[6] S. Aghaei, M. A. Nematbakhsh, and K.H. Farsani, "Evolution of the world wide web: from Web 1.0 to Web 4.0.," *International Journal of Web & Semantic Technology,* vol. 3, no. 1, pp. 1-10, 2012.

[7] Twitter, «Hitos de Twitter,» [En línea]. Available: https://about.twitter.com/es/company/press/milestones. [Último acceso: 2016 Enero 15].

[8] Twitter, «Empresa,» [En línea]. Available: https://about.twitter.com/company. [Último acceso: 1 Octubre 2015].

[9] B. O'Connor, et al., «From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series,» *ICWSM,* vol. 11, nº 1-2, pp. 122-129, 2010.

[10] O. Maimon and L. Rokach, Data mining and knowledge discovery handbook, Springer, 2005.

[11] E. Marrese-Taylor, J.D. Velásquez, F. Bravo-Marquez & Y. Matsuo, «Identifying customer preferences about tourism products using an aspect-based opinion mining approach,» *Procedia Computer Science,* vol. 22, pp. 182-191, 2013.

[12] E. Marrese-Taylor, J.D. Velásquez, & F. Bravo-Marquez, «A novel deterministic approach for aspect-based opinion mining in tourism products reviews,» *Expert Systems with Applications,* vol. 41, nº 17, pp. 7764-7775, 2014.

[13] B. Pang and L. Lee, «Opinion mining and sentiment analysis,» *Foundations and trends in information retrieval,* vol. 2, nº 1-2, pp. 1-135, 2008.

[14] B. Liu, Web data mining: exploring hyperlinks, contents, and usage data, Springer Science & Business Media, 2007.

[15] F. Bravo-Marquez, E. Frank and B. Pfahringer, «Positive, Negative, or Neutral: Learning an Expanded Opinion Lexicon from Emoticon-annotated Tweets,» de *IJCAI '15: Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015.

[16] B. Liu, «Sentiment analysis and opinion mining,» *Synthesis Lectures on Human Language Technologies,* vol. 5, nº 1, pp. 1-167, 2012.

[17] J. A. Balazs and J. D. Velásquez, «Opinion Mining and Information Fusion: A Survey,» *Information Fusion,* vol. 27, pp. 95-100, 2016.

[18] M. Hu and B. Liu, "Mining and summarizing customer reviews.," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2004, pp. 168-177.

[19] F. Bravo-Marquez, E. Frank and B. Pfahringer, "From Unlabelled Tweets to Twitter-specific Opinion Words," in *SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research & Development in Information Retrieval.*, Santiago, Chile, 2011.

[20] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu and B. Liu., "Combining lexicon based and learning-based methods for twitter sentiment analysis," *HP Laboratories, Technical Report HPL-2011,* vol. 89, 2011.

[21] R. Kohavi and F. Provost, «Glossary of terms,» *Machine Learning,* vol. 20, nº 2-3, pp. 271-274, 1998.

[22] C.D. Manning, P. Raghavan and H. Schütze, «Evaluation in information,» de *Introduction to Information Retrieval*, Cambridge: Cambridge university press, 2008, pp. 151-175.

[23] A. Viera and J. Garrett, «Understanding interobserver agreement: the kappa statistic,» *Fam Med,* vol. 37, nº 5, pp. 360-363, 2005.

[24] J.R. Landis and G.G. Koch, «The measurement of observer agreement for categorical data,» *Biometrics,* vol. 33, nº 1, pp. 159-174, 1977.

[25] M. Souza et al., «Construction of a portuguese opinion lexicon from multiple resources,» de *STIL*, 2011.

[26] J. Kamps, M. Marx, R.J. Mokken and M. De Rijke, "Using WordNet to Measure Semantic Orientations of Adjectives," in *LREC*, 2004.

[27] A. Esuli & F.Sebastiani, «Determining the semantic orientation of terms through gloss classification.,» *Proceedings of the 14th ACM international conference on Information and knowledge management,* pp. 617-624, 2005.

[28] A. Andreevskaia & S. Bergler, «Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses,» *EACL,* vol. 6, pp. 209-216, 2006.

[29] A. Esuli & F. Sebastiani, «Sentiwordnet: A publicly available lexical resource for opinion mining,» *Proceedings of LREC,* vol. 6, pp. 417-422, 2006.

[30] F. L. Cruz, J. A. Troyano, B. Pontes, and F.J. Ortega, «ML-SentiCon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas,» *Procesamiento del Lenguaje Natural,* vol. 53, pp. 113-120, 2014.

[31] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, 1997.

[32] P. D. Turney, «Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.,» de *Proceedings of the 40th annual meeting on association for computational linguistics.*, 2002.

[33] K.W. Church and P. Hanks, «Word association norms, mutual information, and lexicography,» *Computational linguistics,* vol. 16, nº 1, pp. 22-29, 1990.

[34] H. Yu and V. Hatzivassiloglou, «Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences,» de *Proceedings of the 2003 conference on Empirical methods in natural language processing.*, 2003.

[35] H. Kanayama and T. Nasukawa., «Fully automatic lexicon expansion for domain-oriented sentiment analysis,» de *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006.

[36] X. Ding, B. Liu, and P. S. Yu, «A holistic lexicon-based approach to opinion mining,» *Proceedings of the 2008 International Conference on Web Search and Data Mining,* pp. 231-240, 2008.

[37] S.M. Mohammad & S. Kiritchenko, «Using hashtags to capture fine emotion categories from tweets,» *Computational Intelligence,* vol. 31, nº 2, pp. 301-326, 2015.

[38] S.M. Mohammad, S. Kiritchenko & X. Zhu, «NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets,» *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013),* 2013.

[39] S. Mohammad, B. Dorr, and C. Dunne, «Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus,» *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* vol. 2, pp. 599-608, 2009.

[40] M. Araújo, P. Gonçalves and F. Benevenuto, «Métodos para Análise de Sentimentos no Twitter,» 2015.

[41] A. Go,R. Bhayani and L. Huang, «Twitter sentiment classification using distant supervision,» *CS224N Project Report, Stanford,* vol. 1, p. 12, 2009.

[42] M.D. Molina González et al. , «eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico,» 2015.

[43] P. Chikersal, S. Poria and E. Cambria , «SeNTU: Sentiment analysis of tweets by combining a rule-based classifier with supervised learning,» de *Proceedings of the International Workshop on Semantic Evaluation (SemEval 2015)*, 2015.

[44] D. Vilares, M.A. Alonso and C. Gómez-Rodríguez, «Una aproximación supervisada para la minería de opiniones sobre tuits en español en base a conocimiento lingüístico,» *Procesamiento del lenguaje natural,* vol. 51, pp. 127-134, 2013.

[45] J. Balazs, «Diseño, desarrollo e implementación de una aplicación de web opinion mining para identificar el sentimiento de usuarios de Twitter con respecto a una campañía de retail,» Universidad de Chile - Facultad de Ciencias Físicas y Matemáticas, Santiago de Chile, 2015.

[46] M. Taboada et al., «Lexicon-based methods for sentiment analysis. Computational linguistics,» *Computational linguistics,* vol. 37, nº 2, pp. 267-307, 2011.

[47] D. Vilares, M.A. Alonso and C. Gómez-Rodríguez, «Clasificación de polaridad en textos con opiniones en espanol mediante análisis sintáctico de dependencias.,» *Procesamiento del lenguaje natural,* vol. 50, pp. 13-20, 2013.

[48] R. Saurí, «A factuality profiler for eventualities in text,» *ProQuest,* 2008.

[49] A. O'Keeffe & M. McCarthy, The Routledge handbook of corpus linguistics, Routledge, 2010.

[50] T. McEnery & A. Hardie, Corpus linguistics: Method, theory and practice, Cambridge University Press, 2011.

[51] F. Vera, «Caracterización de perfiles influyentes en Twitter de acuerdo a tópicos de opinión y la generación de contenido interesante,» de *Universidad de Chile - Facultad de Ciencias Físicas y Matemáticas*, Santiago, Chile, 2015.

[52] PostgreSQL , «About,» [En línea]. Available: http://www.postgresql.org/about/. [Último acceso: 28 Diciembre 2015].

[53] R.L. Chapman and P.M Roget, Roget's international thesaurus, New York: Harper & Row, 1984.

[54] L. Padró and E. Stanilovsky, «FreeLing 3.0: Towards Wider Multilinguality,» de *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.*, Istanbul, Turkey, Mayo, 2012.

[55] Oracle, «Java Basics: The Java Programming Language and the Java Platform,» [En línea]. Available: http://www.oracle.com/technetwork/topics/newtojava/downloads/index.html. [Último acceso: 29 Diciembre 2015].

[56] D. MacKay, «Chapter 20. An Example Inference Task: Clustering,» de *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003, p. 284–292.

[57] Adsocia, «Ranking de cuentas de Chile en Twitter,» [En línea]. Available: http://adsocia.com/index.php?m=website&a=ranking&place=Chile. [Último acceso: 13 Enero 2016].

[58] I. Alegria et al. , «Introducción a la Tarea Compartida Tweet-Norm 2013: Normalización Léxica de Tuits en Español,» *In Tweet-Norm@ SEPLN,* pp. 1-9, 2013.

[59] P. Turney and M.L. Littman, «Unsupervised learning of semantic orientation from a hundred-billion-word corpus.,» 2002.

[60] G. Boeree, «Language development,» *General psychology,* 2003.

[61] F. Jiang et al., «Every Term Has Sentiment: Learning from Emoticon Evidences for Chinese Microblog Sentiment Analysis,» *Natural Language Processing and Chinese Computing,* pp. 224-235, 2013.

[62] J.M Perea-Ortega and A. Balahur , «Experiments on feature replacements for polarity classification of spanish tweets.,» de *Proceedings of the TASS workshop at SEPLN.*, 2014.

[63] J.L. Fleiss, B. Levin, and M. C. Paik, Statistical methods for rates and proportions, John Wiley & Sons, 2013.

[64] S. Estévez-Velarde and Y. A. Cruz, «Evaluación de algoritmos de clasificación supervisada para el minado de opinion en twitter,» 2013.

Anexos

1 Palabras con polaridad conocida

Palabras positivas
positivo
ideal
belleza
bien
maravilloso
bonito
estupendo
perfecto
deslumbrante
excelente
perfección
asombroso
bonito
espléndido
magnífico
deseable
maravilloso
sensacional
exquisito
elegancia
súper
fabuloso
elegante

Palabras negativas
negativo
despreciable
insignificante
insatisfactorio
malo
detestar
grave
indigno
vergonzoso
detestable
enfermo
horrible
pecaminoso
terrible
patético
sucio
lamentable
infernal
pobre
falta
miserable
vil
triste
abominable
siniestro
inferior
deplorable
indeseable
inaceptable

2 Abreviaciones más frecuentes

Abreviación	Corrección		
aca	acá	pense	pensé
adios	adiós	plis	por favor
ahaha	jaja	pls	por favor
ahi	ahí	po	pues
Ahi	ahí	porfa	por favor
aki	aquí	porq	porque
aprox	aproximadamente	porqe	porque
aqui	aquí	pq	porque
aunq	aunque	q	que
aver	a ver	Q	que
bueh	bueno	qe	que
cel	celular	Qe	que
celu	celular	quee	que
cm	como	razon	razón
cn	con	rio	río
d	de	salio	salió
escu	escuela	sere	seré
estan	están	sii	sí
fb	Facebook	Sii	sí
FB	Facebook	siii	sí
finde	fin de semana	Sip	sí
frio	frío	soi	soy
gim	gimnasio	sorry	perdón
grax	gracias	tbn	también
grx	gracias	tia	tía
gym	gimnasio	tio	tío
haha	jaja	tkm	te quiero mucho
info	información	tmb	también
ire	iré	tmbn	también
xq	porque	toy	estoy
uds	ustedes	tqm	te quiero mucho
Nose	no sé	ud	usted
peli	película	unica	única
k	que	unico	único
ojala	ojalá	voi	voy
min	minutos	x	por
nose	no sé	xfa	por favor

3 Risas para normalización

Risas más frecuentes	
haha	jaja
Jaja	jaja
JAJA	jaja
Jajaj	jaja
jeje	jaja
lol	jaja

4 Etiquetas Eagles para POS-TAGGING

ADJETIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
3	Grado	Aumentativo	A
		Diminutivo	D
		Comparativo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Función	-	0
		Participi	P

DETERMINANTES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Determinante	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
		Artículo	A
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Poseedor	Singular	S
		Plural	P

ADVERBIOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adverbio	R
2	Tipo	General	G
		Negativo	N

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5-6	Clasificación semántica	Persona	SP
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

VERBOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

CONJUNCIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Conjunción	C
2	Tipo	Coordinada	C
		Subordinada	S

INTERJECCIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interjección	I

PRONOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I
		Interrogativo	T
		Relativo	R
		Exclamativo	E
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		ImpersonalMInvariable	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblicuo	O
7	Poseedor	Singular	S
		Plural	P
8	Politeness	Polite	P

PREPOSICIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída	C
3	Género	Masculino	M
4	Número	Singular	S

SIGNOS DE PUNTUACIÓN			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Puntuación	F

NUMERALES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Cifra	Z
2	Tipo	partitivo	d
		Moneda	m
		porcentaje	p
		unidad	u

FECHAS Y HORAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Fecha/Hora	W

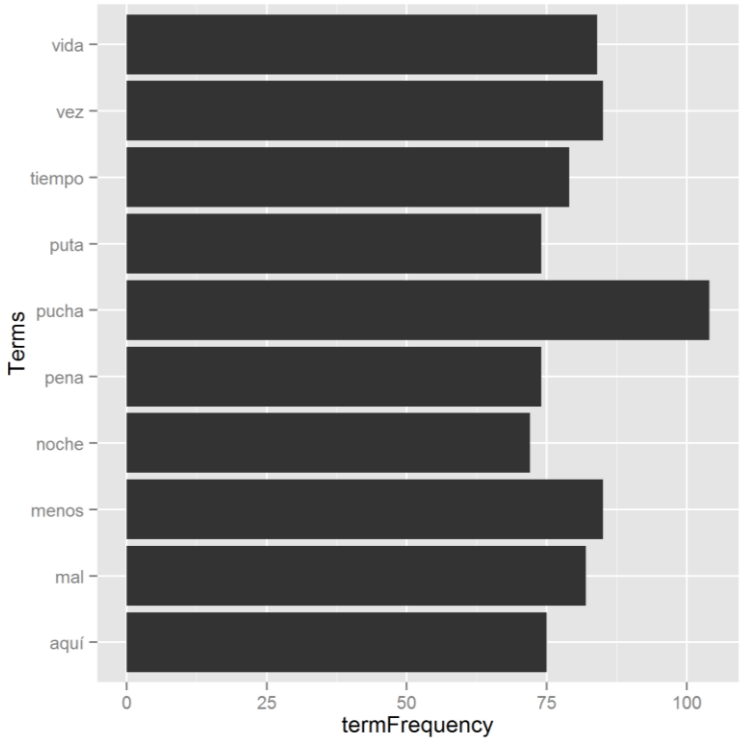
5 Usuarios iniciales corpus

@GermanGarmendia	@KarendTV	@pedroruminot	@christianpino	@anatijoux
@CHISTE	@FelipeAvello	@vmoulian	@contreras_kathy	@MesiasVega
@tv_mauricio	@elmostrador	@alejoferrada	@andresbaile	@DiarioLaHora
@SoledadOnetto	@DomiGallego	@reddeemergencia	@solebacarreza	@andydellacasa
@24HorasTVN	@tomasgonzalez1	@danielexhuevo	@SV_CHV	@consejocultura
@sebastianpinera	@_FALOON_	@AhoraNoticiascl	@bonvallet	@AndresVelasco
@T13	@tolerancia0	@lagosweber	@lucialopezchile	@MyriamHN
@TVN	@chilevision	@panchosagredo	@ivanguerrerom	@KATHYBODIS
@biobio	@lgolborne	@franciscamusic	@JanisPope	@romizalazar
@CNNChile	@fabriziocopano	@PanchaMerino1	@cata_edwards	@pablosimonetti
@SoloFilosofia	@tv_Amaro	@tvn_gonzalo	@Giia_marengo	@jasalfate
@rubionatural	@sergiolagos	@gutierreznacho	@ciper	@kingarturo23
@Cooperativa	@PublimetroChile	@evagomezs	@CDF_cl	@nicolemusica
@alejodorowsky	@fernandahansen	@udechile	@intrusoslared	@pelotazo
@StefanKramerS	@deportes13cl	@franciniamaral	@Carolina_Toha	@angelicacastro_
@benjavicunaMORI	@13AR	@RadioCarolina	@muchogustoMEGA	@antonellarios
@camila_vallejo	@lfranzani	@vivirodriguesof	@nelsonavila	@barriolapeli
@TonkaTP	@marcoporchile	@jpcrettino	@allamand	@dj_emilio
@Karol_LuceroV	@Guatonsalinas	@JoseManuelR	@patricionavia	@Nabih_Chadud
@iambetocuevas	@Carabdechile	@ramirez_polo	@andressilvaa	@Porlaputa
@buenosdiatodos	@iamdelafuente	@mariseka	@mgsubercaseaux	@ingridcruztoro
@mxperez	@fuentesilva	@KathySalosny	@entel	@albertoplaza
@Lafrangh	@tv_monica	@Alisonmandel	@LaRedTV	@EAlbasetti
@carolaurrejola	@javilarusia	@Los_Bunkers	@rodrigosepu	@RicardoLagos
@RinconSalfate	@pillanes	@Enportada	@Fr_parisi	@ferurre
@josemvinuela	@metrodesantiago	@vale_ortega	@INFORMADORCHILE	@jmanalich
@consuelosaav	@bianchileiton	@Lucilavit	@vilchesip	@MarketingHoy
@fernandopaulsen	@mvacarezza	@enavonbaer	@felipekast	@futurofm
@matiasdelrio	@zabaletachile	@PDI_CHILE	@Vardoc1	@fcokaminski
@halconmatinal	@lacuarta	@caro_mestrovic	@elquenoaporta	@webonomia
@copano	@adribarrientos	@MovistarChile	@Ariel_Levy	@BelenHidalgo
@latercera	@alvarez_monse	@sergiofraude	@giancarlopeta	@tvn_prog
@CruzRojaInforma	@el_mago_oficial	@lamaldito	@Orrego	@arriagadabizaca
@canal13	@onemichile	@K3LCALDERON	@felipeharboe	@40ChileOficial
@Teleton	@jumastorga	@Cumparini	@sismoguc	@HocicoTarro
@RafaAraneda	@JPQueralto	@ColoColo	@humbertosuazop	@ScarlethCardena
@Emol	@chvdeportes	@MaiteOrsini	@Jason_heredia	@lilyperez
@TerraChile	@Leonorvarela	@chico_perez	@Rayenaraya	@ChileActualidad
@Lavivirocks	@carolinademoras	@NissRosenthal	@Mega	@revistaQP
@GobiernodeChile	@rhinzpeter	@felipevidalc	@UltimoMinutoCL	@blanquifran
@elfergonzalez	@PatricioDelSol	@DMatamala	@Educacion2020	@GondwanaChile
@hoytschile	@fayerwayer	@ceciliamorel	@Rumpy1000	@cqcchile
@Calle7_TVN	@La_Segunda	@SamsungChile	@JuanPedroTV	@polaco_goldberg
@theclinicc l	@sismos_chile	@Catavallejos	@gustavohuerta	@_PrimerPlano
@leonardofarkas	@SERNAC	@blancalewin	@nacioncl	@Jhendelyn
@adnradiochile	@GiorgioJackson	@asisomosoficial	@jordicastell	@Gerente2012
@YINGO_oficial	@DjMendezmusic	@RodrigoGallina	@renenaranjo	@ElGraficoChile

6 Grupos de clustering

6.1 Grupo 1

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  En este grupo se encuentran comentarios que demuestran tristeza, pena, decepción frente a algún evento en particular.

- Ejemplos:

$$@TenchaSalvaje\ Pucha\ qué\ pena, minas\ así\ les\ ponen\ la\ pata\ encima\ siempre.$$

$$Ojalá\ tire\ pa'rriba!\ :($$

$$@psicotonio\ pucha, tu\ bio\ es\ la\ tristeza\ máxima, la\ vida\ \text{sin}\ palta\ y\ ajo\ no\ vale$$

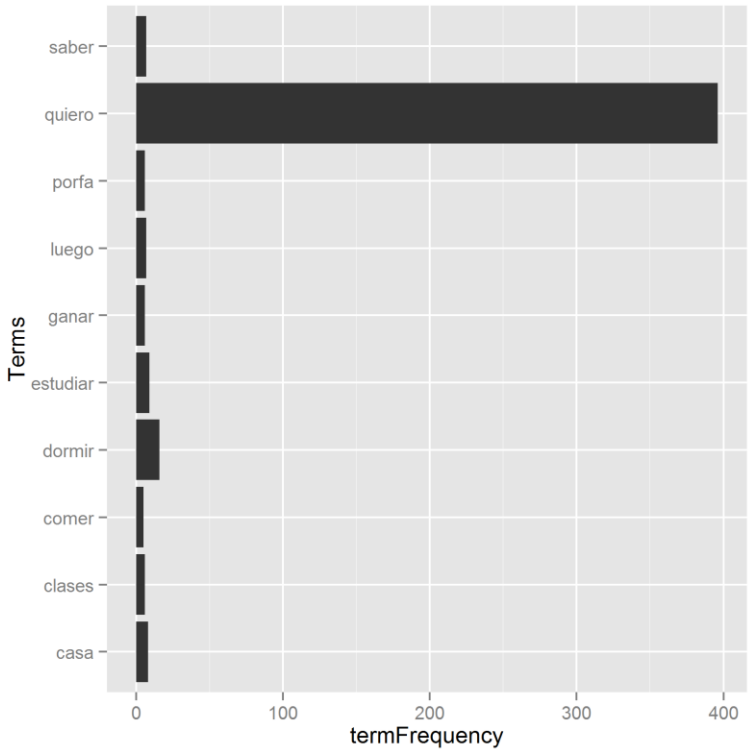
$$la\ pena\ vivirla\ po!:($$

$$@ZukyLyon\ @dalcahue1\ jajaja\ \ldots\ Menos\ mal\ que\ mis\ apoderados\ y\ mis\ niñas$$

$$(alumnas)\ me\ adoran\ jajaja\ldots y\ apoyan\ el\ paro\ indefinido:D$$

6.2 Grupo 2

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Presenta comentarios donde usuarios expresan su voluntad, deseo de realizar o lograr algo.

- Ejemplos:

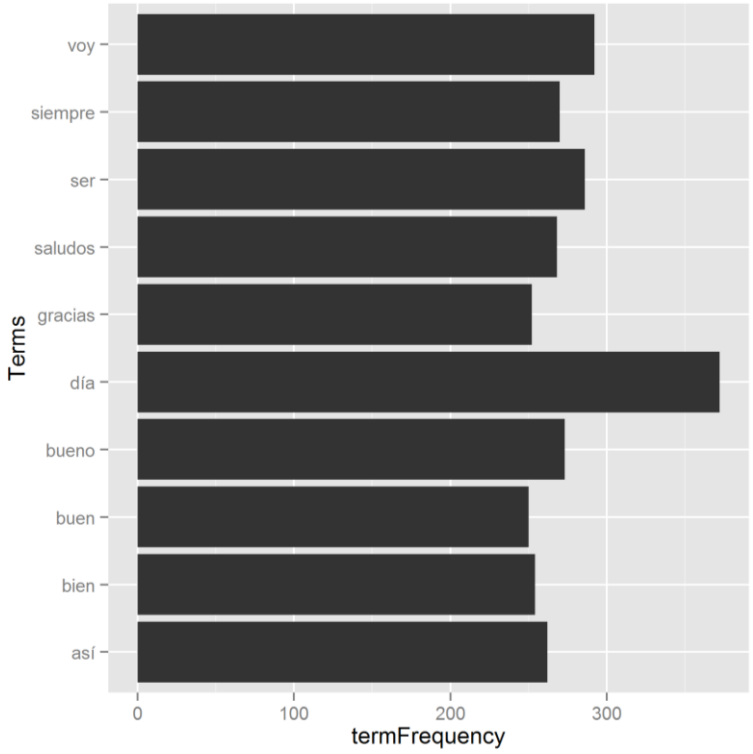
$$No\ quiero\ estudiar\dots ya\ me\ habia\ acostumbrado\ a\ no\ estar\ en\ la\ U\ :'($$

$$Quiero\ dormir\ 24\ horas\ nada\ mas\ :($$

$$Me\ quiero\ ir\ pa\ mi\ casa\ :($$

6.3 Grupo 3

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Se encuentran comentarios de saludos y de buenos deseos para el día, este tipo de comentario representa gran cantidad de tweets dentro del corpus.

- Ejemplos

$$Buenos\ días\ gente\ guapa.\ Por\ fin\ es\ viernes.\ A\ pasar\ buen\ día!!!!\ \ Se\ agradecen\ los$$

$$RT\ :)\ http://t.co/TZJEwp6K4L$$

$$@catherine\_fulop\ Cathy\ ¡!!!!\ \ BUENOS\ DÍAS\ :\ D$$

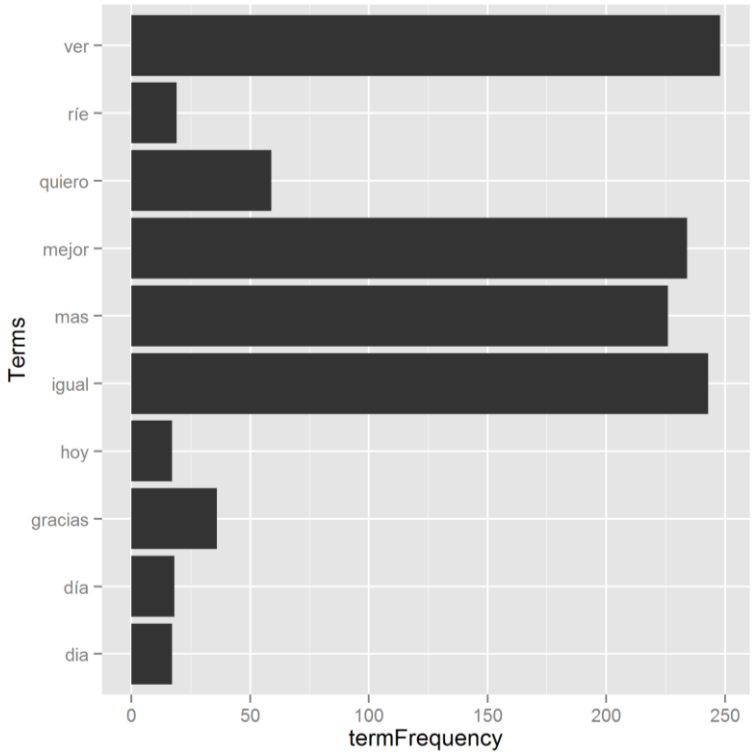
$$hace\ mucho\ no\ te\ jodo\ jaja\ que\ tengas\ un\ hermoso\ Viernes\ y\ un\ buen\ finde\ :)))$$

$$@labragandiash:\ Buenos\ días\ mundo\ mundial,\ el\ capitán\ del\ amor\ os\ desea\ un$$

$$buen\ día\ :\ D\ Jaja\ @myh\_tv\ http://t.co/vxvflqBjyx$$

6.4 Grupo 4

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Dentro de este grupo se aprecian comentarios que realizan comparaciones frente a distintos eventos, situaciones, principalmente mejoras a estos eventos o situaciones.

- Ejemplos:

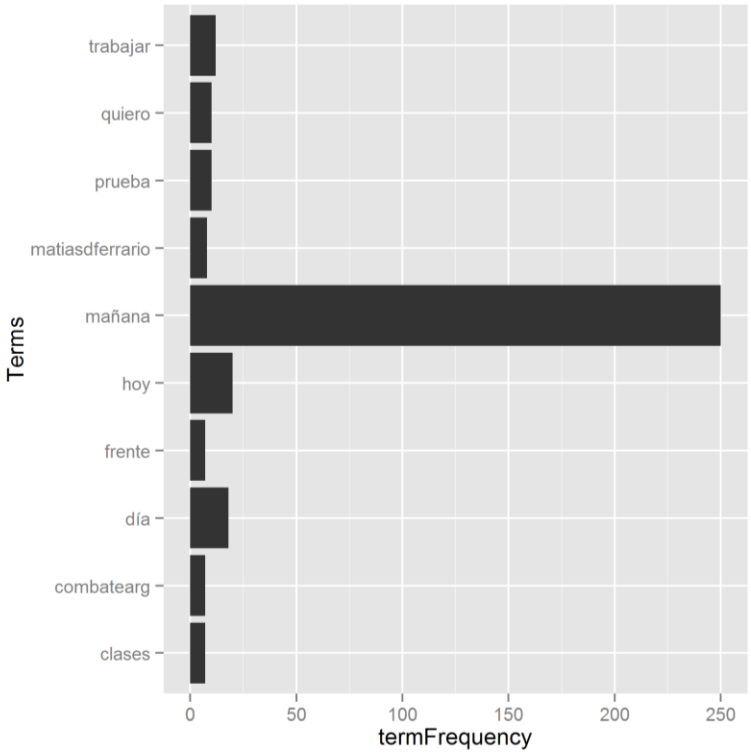
$$@Katapiumpium\ al\ menos\ no\ se\ es\ mejor\ ver\ series\ k\ pasan\ piola\ pero\ son\ igual\ de\ buenas: D$$

$$Te\ me\ caiste\ Arica: (\ Mas\ arreglado\ no\ pudo\ estar.\ Pero\ igual\ felicidades\ Sexta\ Región\ Campeones\ de\ Chile\ Cueca\ Adulto\ Arica.$$

$$@pedrokayak\ @NoAltoMaipo\ \ deberían\ hacer\ algo\ mas\ masivo\ en\ stgo\ igual :)$$

6.5 Grupo 5

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Corresponde a comentarios donde se expresa alguna intención principalmente para realizar durante la mañana o el día siguiente. Dentro de este grupo también se aprecia gran cantidad de retweets de comentarios de "Matt Ferrario"[9], modelo y actor argentino con gran cantidad de seguidoras.

- Ejemplos:

$$@CariFutbolera\ yo\ vengo\ carretiando\ del\ lunes, hoy\ libre\ mañana\ de\ papa,$$

$$responsablemente\ :)$$

$$Y\ hoy\ fue\ mi\ último\ día\ en\ la\ pega\ desde\ mañana\ soy\ cesante\ :)$$

$$@matiasdferrario: Me\ Gusta\ mi\ bronceado\ Natural\ by\ @BSASBRONZE$$
$$:)\ Animate\ a\ visitarlos!\ http://t.co/8VEybFjTte$$

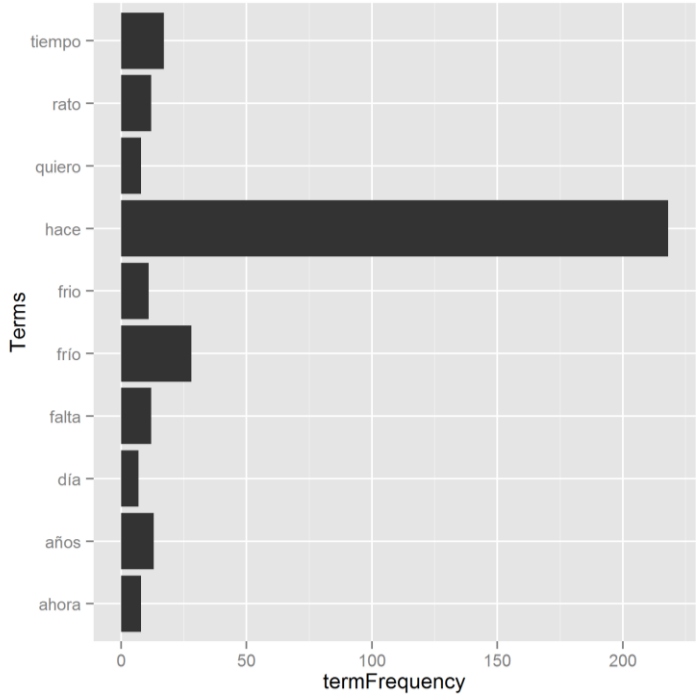
$$@matiasdferrario: Mañana\ Juntada\ Solidaria\ junto\ a\ @GonzaGravano$$

[9] https://twitter.com/matiasdferrario

$@combatearg\ 15hs\ Frente\ al\ Planetario!\ :)\ +\ http://t.co/nmUFTsR2sI$

6.6 Grupo 6

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Dentro de este grupo se observan comentarios que hacen alusión a la condición que se presenta en cuanto al clima, y comentarios que hablan respecto a la duración de eventos, situaciones.

- Ejemplos:

$$Que\ fríoo\ hace, a\ ponerle\ onda\ que\ es\ viernes\ :)$$

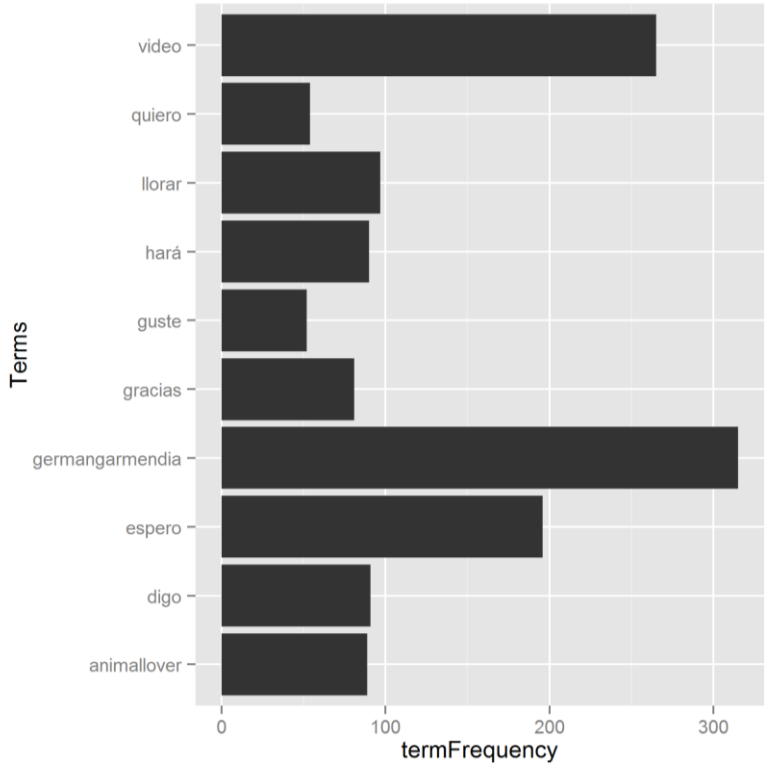
$$@dhelstein\ falta\ TAAAAANTO\ tiempo\ D: necesitamos\ más\ fotos\ como\ esa,$$

$$millones\ más\ xD$$

$$Estoy\ pensando\ en\ la\ perdida\ de\ tiempo\ que\ tendre\ hoy\ :($$

6.7 Grupo 7

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Este grupo está altamente influenciado por Germán Garmendia, un youtuber chileno, reconocido en el territorio y a nivel internacional. Dentro de este grupo se presentan interacciones con Germán Garmendia y retweets a los comentarios que él realiza.

- Ejemplos:

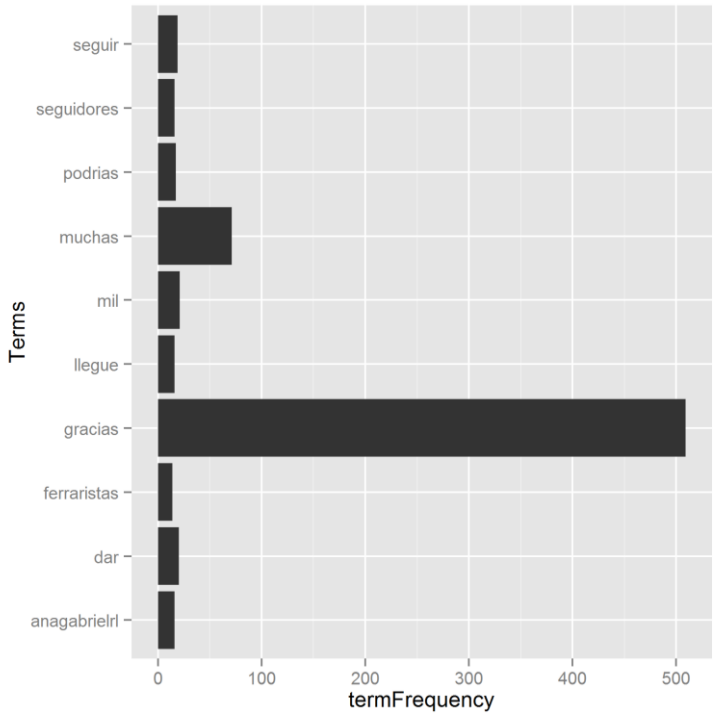
$$@GermanGarmendia\ ok\ creamos\ un\ charco\ de\ lodo\ entonces: -)$$

$$@GermanGarmendia\ a\ que\ hora\ sale\ el\ video\ ...?\ Si\ podes\ saludame\ en\ unos\ de\ tus\ videos\ soy\ fanatico\ tuyo: D$$

$$@GermanGarmendia: \text{https://t.co/SaaZ6eqXcP}\ Este\ video\ te\ hará\ llorar\ ...\ con\ eso\ lo\ digo\ todo!: '(\ \#AnimalLover$$

6.8 Grupo 8

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Corresponde comentarios que hacen referencia a agradecimientos, incluyendo un comentario que se repite bastante dentro del corpus para seguir a una cuenta de la cantante mexicana Ana Gabriel.

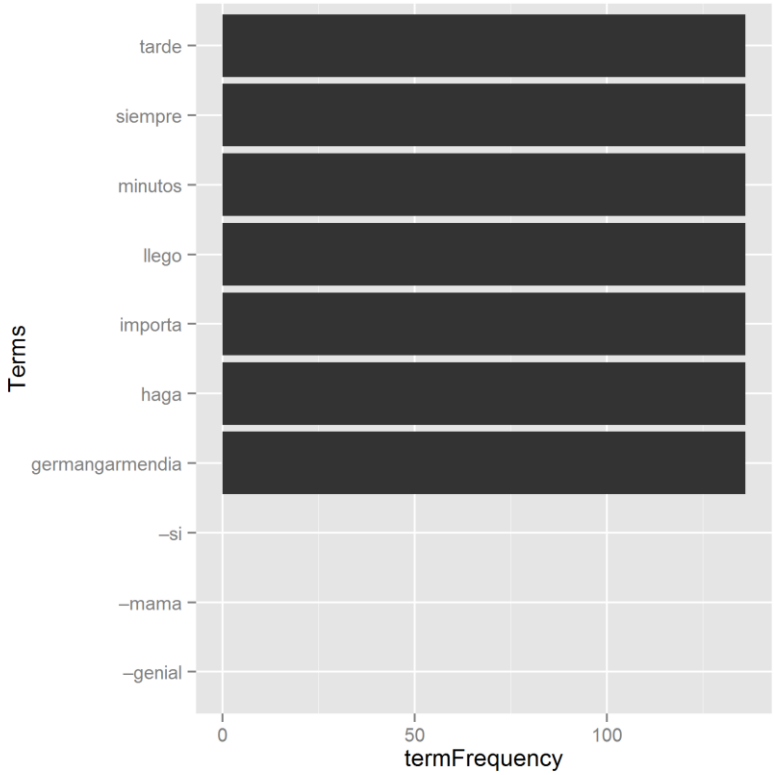
- Ejemplos:

$$@PilarOsso\ muchas\ gracias!!!\ \ Nos\ vamos\ contentos\ a\ la\ cama\ ;)$$

$$@perezvelard\ Hola\ podrias\ seguir\ a\ @ANAGABRIELRL\ para\ que\ llegue\ a\ 500\ mil\ seguidores?\ \ Gracias\ :)$$

6.9 Grupo 9

- Gráfico frecuencia palabras dentro del grupo:



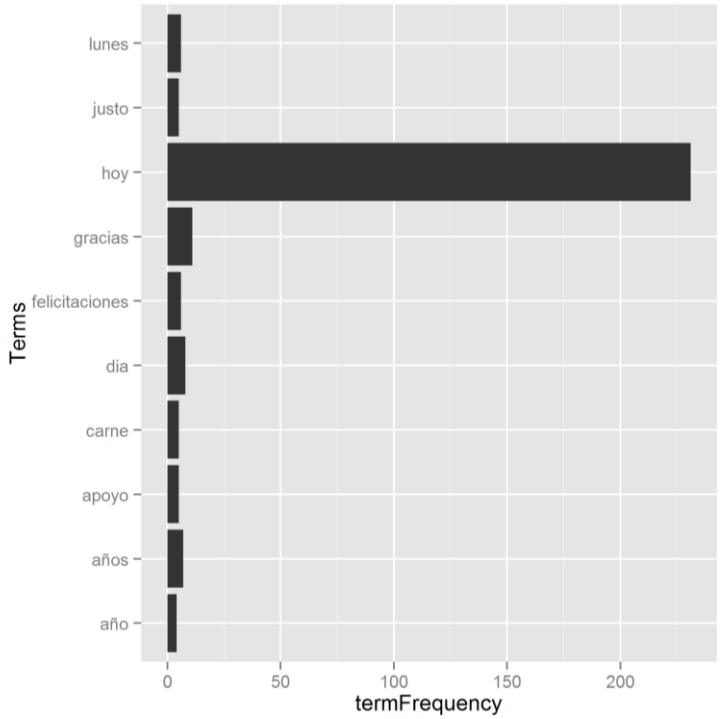
- Descripción del grupo:
  Este grupo está conformado por un comentario que se repite bastante en el corpus, por el nivel de retweets presentes.

- Ejemplo:

$$RT\ @GermanGarmendia{:}\ No\ importa\ lo\ que\ haga, siempre\ llego$$
$$10\ minutos\ tarde\ a\ todo!:($$

6.10 Grupo 10

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Este grupo presenta comentarios donde se habla principalmente de situaciones, eventos producidos en el mismo día, se aprecia que una gran cantidad de personas hablan sobre el día lunes.

- Ejemplos:

$$@edu\_castillo\ @Katherista\ @TomasPardo\ Buen\ dia\ mucha\ energia\ y\ animo$$

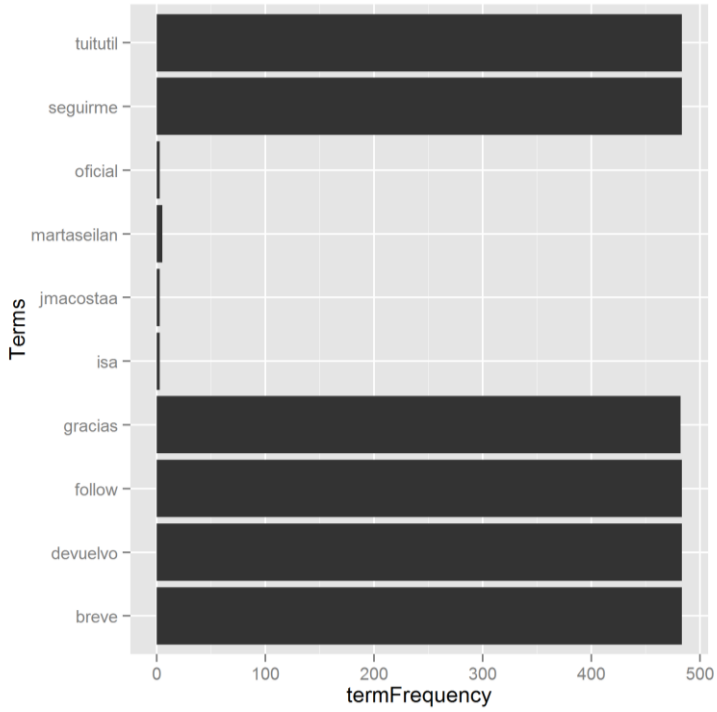
$$para\ Hoy\ Lunes\ que\ tengan\ una\ Buena\ Semana\ :)$$

$$RT\ @Karol\_LuceroV: Almorzando\ con\ @Yoxilla\_Ka\ la\ conocí\ un\ día\ como\ hoy\ hace\ 7$$

$$años \ldots Gracias\ por\ tu\ cariño\ y\ apoyo\ siempre\ TQM\ :)\ http:// \ldots$$

6.11 Grupo 11

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Corresponde a un comentario que realiza la cuenta @TuitUtil[10] agradeciendo a sus seguidores, hablándole a diferentes usuarios de Twitter.

- Ejemplos:

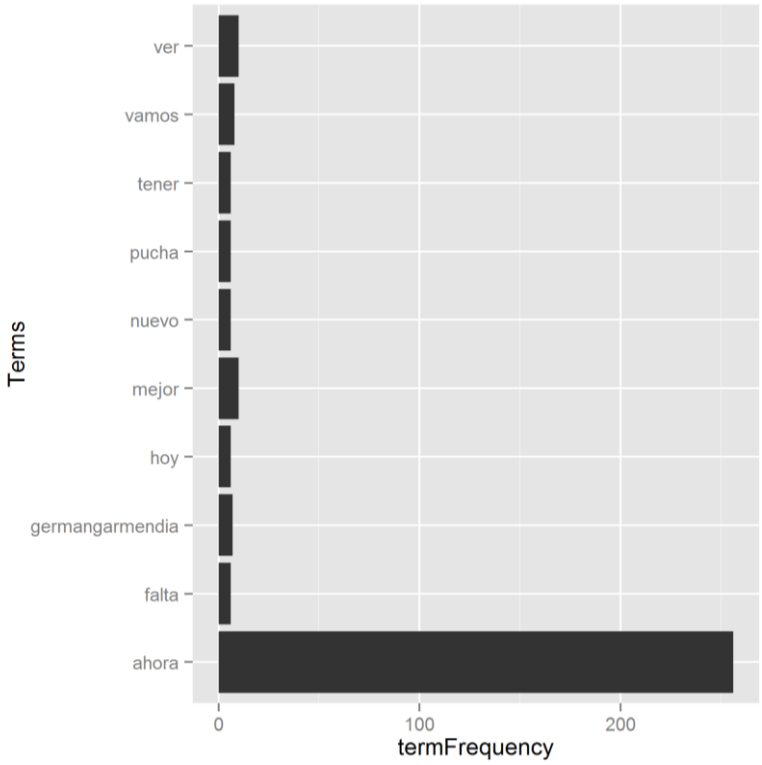
$$@elmejorpaisaje\ Gracias\ por\ seguirme, en\ breve\ te\ devuelvo\ follow$$
$$:)\ \#TuitUtil\ http://t.co/yjiKpxEL43$$

$$@GorkaAhal\ Gracias\ por\ seguirme, en\ breve\ te\ devuelvo\ follow$$
$$:)\ \#TuitUtil\ http://t.co/PAHlypOQ9j$$

[10] https://twitter.com/tuitutil

6.12 Grupo 12

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Principalmente dentro de este grupo se encuentran comentarios sobre eventos que ocurren en el instante o a reacciones frente a diversas situaciones, dentro de este grupo también se incluyeron comentarios asociados a German Garmendia.

- Ejemplos:

$$FULL\ ESTUDIO\ AHORA\ MAÑANA\ PRUEBA\ Y\ NO\ E\ TENIDO\ TIEMPO\ PARA\ ESTUDIAR$$

$$\ldots.VAMOS\ A\ VER\ QUE\ SE\ PUEDE\ LOGRARA\ A\ MENOS\ DE\ 20\ HORA$$

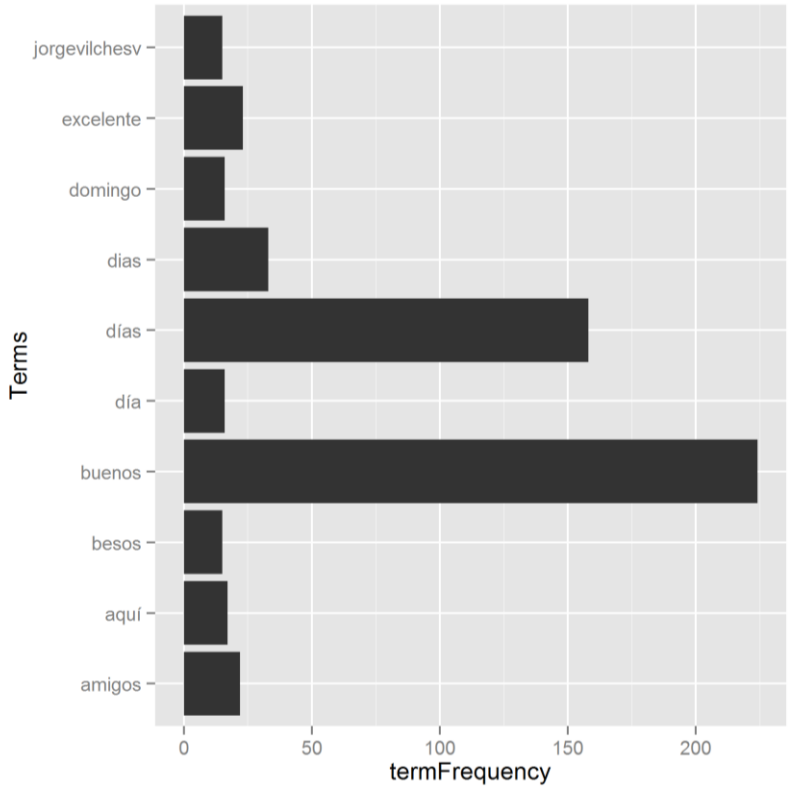
$$DE\ LA\ PRUEBA: p$$

$$@CaatalinaPaz\_\ ahora\ cache: (\ pucha\ era\ la\ mejor\ fonda\ pobre, igual\ fueron$$

$$los\ pioneros\ ahora\ todos\ copian\ la\ idea\ como\ los\ de\ sf$$

6.13 Grupo 13

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Se incluyen dentro de este grupo comentarios que desean buenos días. En este grupo también se incluye un comentario que tuvo gran cantidad de retweets dentro del corpus utilizado que se encuentra asociado a JorgeVilchesV[11], un publicista que se ha especializado en redes sociales y que es parte de los usuarios iniciales con los que fue creado el corpus.

- Ejemplos

$$Buenos\ dias\ querido\ TL!\ :)$$

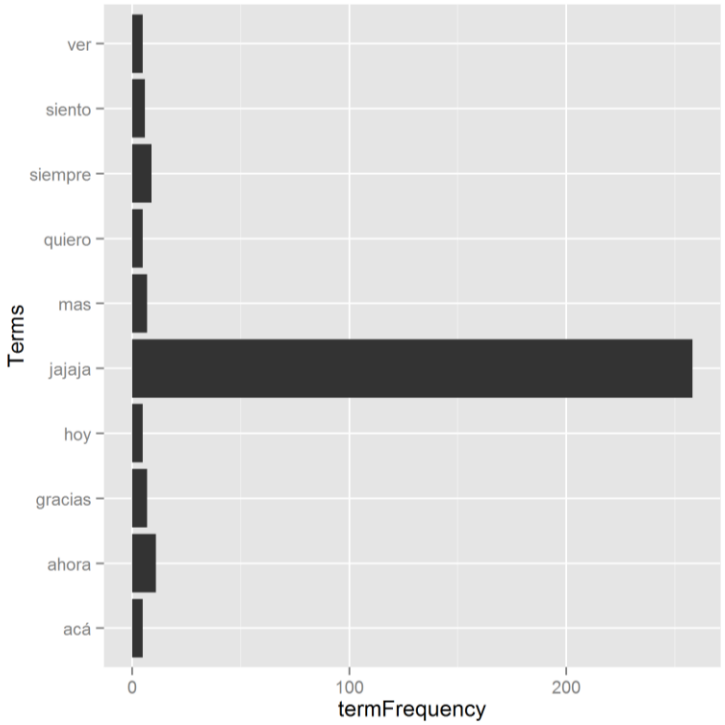
$$@JorgeVilchesV\ Estamos\ esperando\ el\ informe\ OFICIAL\ por\ TV\ :($$

$$\#QEPDJorgeVilches\ \#FuerzaJorge\ \#JorgeVilches\ \#VuelaAltoJorge\ ?$$

[11] https://twitter.com/jorgevilchesv

6.14 Grupo 14

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:
  Dentro de este grupo se encuentran comentarios donde se expresa algún tipo de risa o burla, ocupando principalmente la expresión “jajaja”.

- Ejemplos:

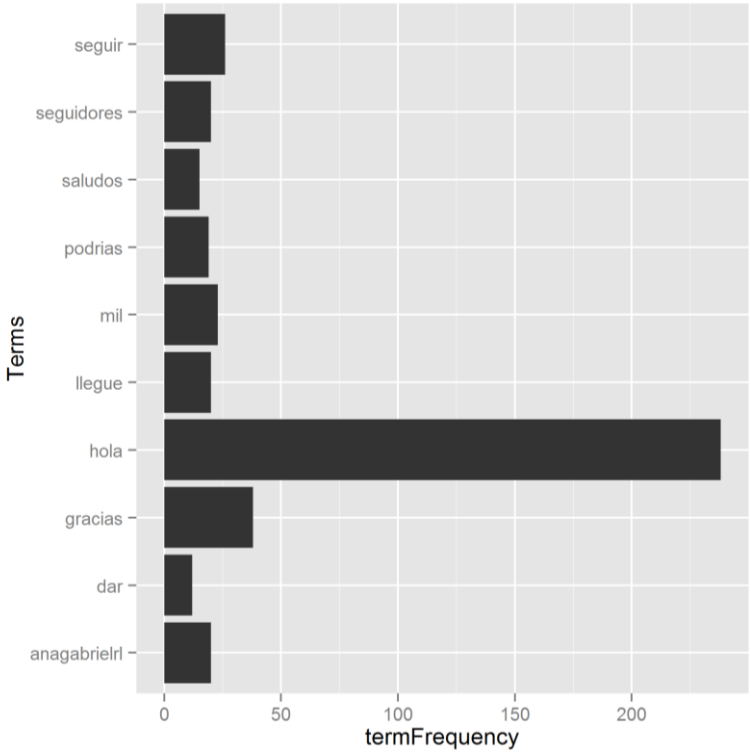
@adnradiochile jajaja conozco una que ahora se hace hincha del colo.

Jajaja cambia la unión por colo colo. Menos mal no cb mi equipo : −)

@AlonzoNicolas ¿?? siento que me van a rajar : (pero es la última que me queda, así que trataré de morir digna jajaja

6.15 Grupo 15

- Gráfico frecuencia palabras dentro del grupo:



- Descripción del grupo:

  Corresponde a una gran cantidad de comentarios de saludos, donde se incluye mayoritariamente el término “hola”. Dentro de este grupo se encuentra también un comentario que se repite bastante dentro del corpus, que busca seguidores.

- Ejemplos:

$$Hola :)$$

$$Hola : P$$

$$@yanka18m\ Podrias\ seguir\ a\ @ANAGABRIELRL\ y\ dar\ RT\ para\ que\ llegue\ a\ 500\ mil\ seguidores?\ Gracias :)\ http://t.co/CIjsXEFB7p$$

$$@hosky74\ Podrias\ seguir\ a\ @ANAGABRIELRL\ y\ dar\ RT\ para\ que\ llegue\ a\ 500\ mil\ seguidores?\ Gracias :)\ http://t.co/CIjsXEFB7p$$

7 Dendograma