PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

ESCUELA DE SISTEMAS

DISERTACIÓN DE GRADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS Y COMPUTACIÓN

ANÁLISIS DE REDES SOCIALES ENFOCADO A LA RED SOCIAL TWITTER MEDIANTE EL USO DE ANÁLISIS DE SENTIMIENTO U OPINION MINING

AUTOR:

ISRAEL MORENO GUZMÁN

Dedicatoria.

Este trabajo de titulación junto con mi trayectoria como estudiante de la carrera de Ingeniería en Sistemas y Computación se la quisiera dedicar primero a mi tía Saba Guzmán y a mi abuelito Moisés Guzmán que hoy me cuidan a mí y a mi familia desde el cielo.

A mi familia más cercana mi madre Eva Guzmán, mi padre Luis Moreno, mi hermana Isaí Moreno y mi abuelita Alicia Palacios les quiero dar eternamente las gracias por acompañarme en mi día a día hasta este momento por siempre haberme apoyado cuando más cansado, triste o golpeado me he sentido durante este camino. Sinceramente sin ustedes cuatro mi vida no sería la misma que es desde que me despierto hasta que me acuesto por las noches.

A otros maravillosos seres humanos que siempre llevare en el corazón por haber compartido conmigo aula, clases y gratos momentos juntos es a mis amigos Sofía, Francisco, Adrián, José y Paúl. Debido a que seguramente sin ellos jamás hubiera culminado deberes, trabajos o proyectos que en su momento parecían imposibles y hoy en día son solo memorias.

A todos mis compañeros y amigos dentro del LTIC PUCE Luis, Sebastián, Christian, Leonardo, Isa, José y Sergio les agradezco por haber sido una gran compañía y soporte durante todos los momentos que compartí junto a ustedes trabajando. Igual de importante y sin olvidarme de Daniel por haber sido el mejor mentor y jefe con el que jamás me haya encontrado.

A mi pequeña familia de amigos conformada por Kevin, Alejandro, Valerie y Kelly quiero darles las gracias por ser los mejores amigos que jamás haya tenido. El soporte y amistad brindados por ustedes trasciende a niveles que jamás pensé encontrar en mi vida.

A mis docentes, tribunal y revisores Ing. Alfredo Calderón, Ing. Fabian de la Cruz e Ing. Oswaldo Espinosa les doy las gracias por su tiempo, paciencia, ayuda y enseñanzas tanto dentro como fuera del aula. Cada uno de ustedes para mí ha sido importante dentro de mi formación profesional.

A mi tío Rafael y mis primos María José, Rafael y Juan Esteban les quiero dar las gracias eternas por ser siempre mi segundo hogar y refugio del mundo. Sinceramente sin ustedes mi vida no tendría el mismo significado que hoy en día tiene.

Finalmente quisiera dedicar este trabajo a mi persona Israel Moreno Guzmán para demostrarme que nada es imposible, inalcanzable o difícil cuando la fe, la familia, el amor y los verdaderos amigos te acompañan durante el camino.

ÍNDICE

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

Capítulo 1 Análisis de Redes Sociales

Dentro del primer capítulo de está disertación se explorará el crecimiento de las redes sociales junto con el del internet en los últimos 10 años y cómo esto ha generado la necesidad de analizar todo el contenido dentro de las mismas. Además, se explorará el modelo de madurez que IBM propone para el análisis de redes sociales dentro de cualquier organización que desee aprovechar los beneficios que esta técnica puede brindar. Finalmente se analizará paso a paso todo el proceso del análisis de redes sociales.

1.1 La Necesidad del Análisis de Redes Sociales

El surgimiento de avances en campos como las ciencias de la computación, estadística, análisis de redes y computación lingüística han proporcionado nuevas técnicas para el seguimiento, modelado, análisis y minería datos que se pueden obtener de las redes sociales.

Estas nuevas metodologías corresponden al:

1. Análisis/Minería de Texto
2. Análisis de Red Social
3. Análisis de Tendencias.

Las tres técnicas mencionadas serán abordadas con más profundidad en los capítulos venideros, debido a que cada una de ellas merece su propio **análisis** para su total comprensión.

1.1.1 Crecimiento en las Redes Sociales

El número de usuarios en redes sociales ha experimentado un crecimiento constante a partir del año 2009, debido a dos factores: el abaratamiento del costo del internet en el mundo y el boom de Facebook. Estos acontecimientos han sido los que prácticamente multiplicaron a los usuarios de internet en todo el mundo.

	Asia	Europe	North America	Latin America / Caribbean	Africa	Middle East	Oceania / Australia
2009	764.4	425.8	259.6	186.9	86.2	58.3	21.1
2010	825.1	475.1	266.2	204.7	110.9	63.24	21.3
2011	1,016.8	500.72	273.07	235.82	139.88	77.02	23.93
2012	1,076.68	518.51	273.79	254.92	167.34	90	24.29
2013	1,265.14	566.26	300.29	302.01	240.15	103.83	24.8
2015	1,563.21	604.12	313.86	333.12	313.26	115.82	27.1
2016	1,792.16	614.98	320.07	384.75	339.28	132.59	27.54
2017	1,938.08	659.63	320.06	404.27	388.38	146.97	28.18
2018	2,062.14	704.83	345.66	438.25	455.84	164.04	28.44
2019	2,300.47	727.56	327.57	453.7	522.81	175.5	28.64

Tabla 1.1 (Clement, Number of internet users worldwide from 2009 to 2019, by region(in millions), 2019)

En la Tabla 1.1 se muestra la cantidad de usuarios en internet a nivel mundial por zona geográfica entre (2009-2019) en millones.

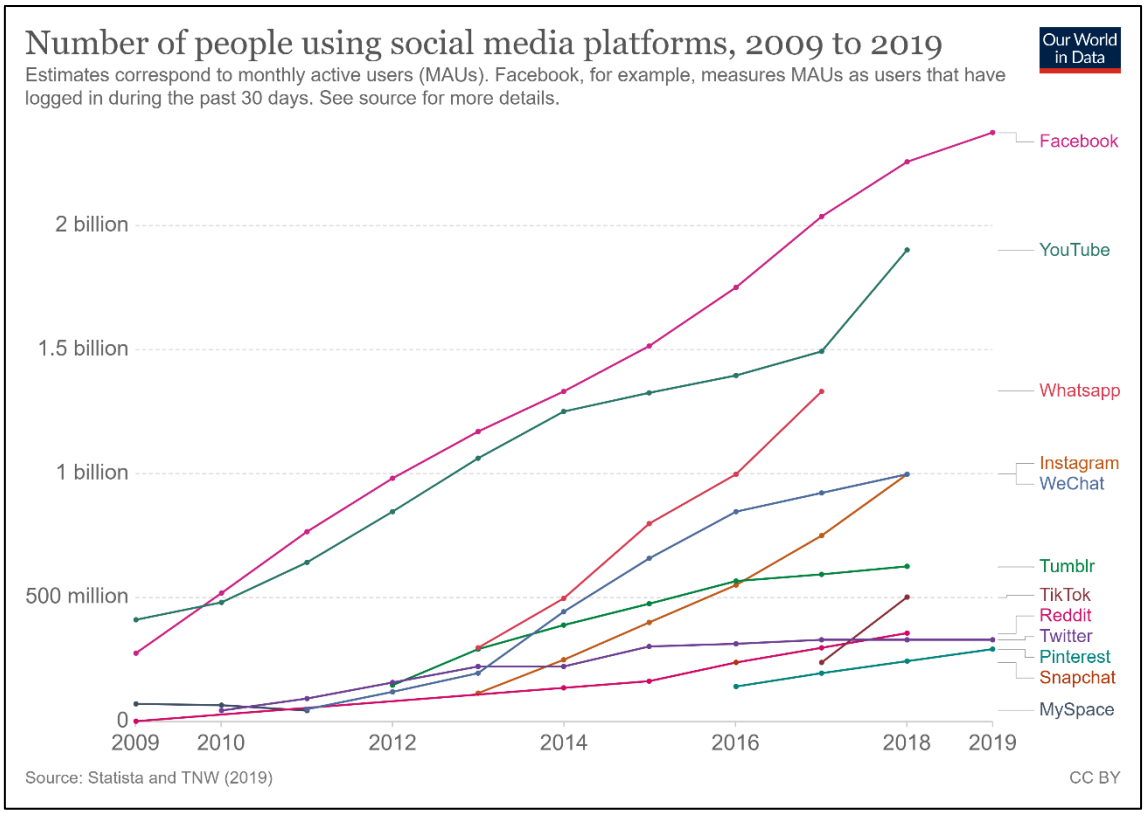


Ilustración 1.1 (Statista & TNW, 2019)

La Ilustración 1.1 muestra el número de usuarios en redes sociales entre (2009-2019) en millones.

Para inicios de la década pasada el número de usuarios registrados en redes sociales era cercano a los mil millones, para enero de 2020 el número asciende a los 3.8 billones de usuarios con un crecimiento anual del 9% (Chaffey, 2020).

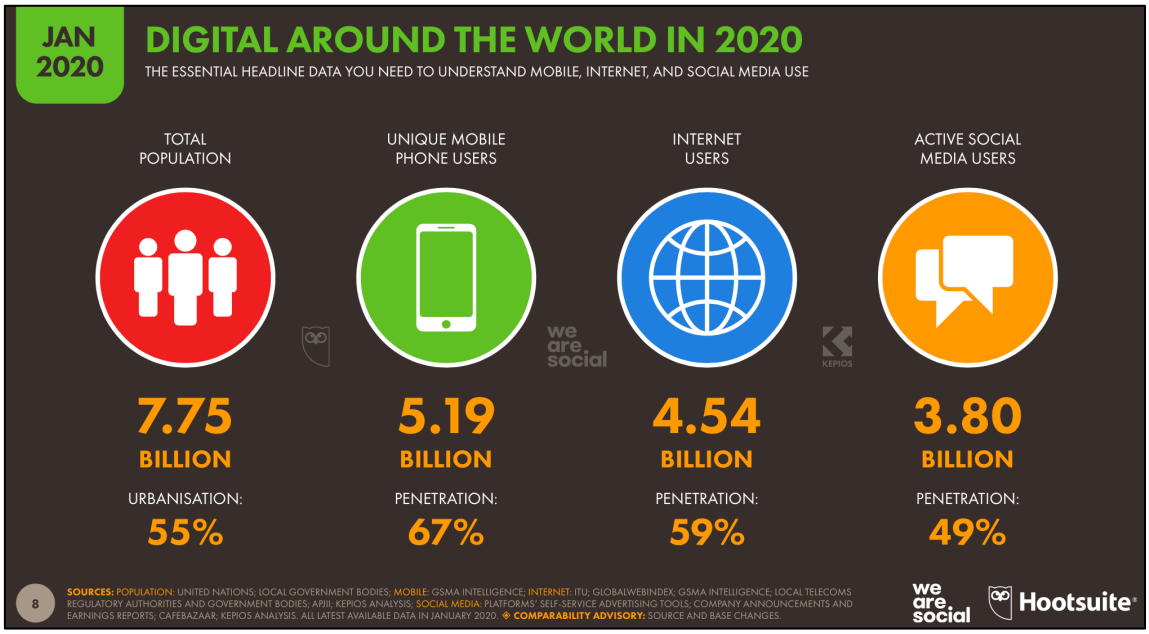


Ilustración 1.2 (Chaffey, 2020)

La Ilustración 1.2 se encarga de mostrar las cifras y porcentajes de urbanización, usuarios únicos de telefonía móvil, usuarios de internet y usuarios activos en las redes sociales a inicios del año 2020 en billones.

Como se puede observar en las ilustraciones y tablas antes presentadas, las redes sociales han pasado a formar parte de nuestra vida y durante la última década han transformado nuestra forma de ver y experimentar el mundo.

1.1.2 Un nuevo campo de estudio para la Ciencia, los Negocios y la Política

El crecimiento casi desenfrenado de las redes sociales ha creado un mundo completamente nuevo e inexplorado para estudios científicos que pueden ser llevados a cabo como un trabajo en conjunto realizado por las ciencias de la computación, estadística, sociología, psicología, etc. Con el objetivo de estudiar a las personas en función de sus opiniones, puntos de vista, actitudes, emociones hacia entidades, individuos, problemas o demás atributos que podamos llegar a pensar.

Estos nuevos estudios representan oportunidades de acercarse a la gente nunca vistas o exploradas por los grandes negocios, políticos o empresas del mundo. Debido a que jamás en la historia estos se han encontrado en una posición más favorable como para financiar estudios científicos de los que puedan llegar a obtener la aceptación de la gente al momento de vender un producto, servicio o campaña política.

Este nuevo panorama en el cual se encuentra el mundo moderno ha sido analizado por parte de IBM sobre todo en el ámbito de los negocios, dentro del cual la empresa propone varios tipos de soluciones para quienes estén

interesados en hacer crecer su valor de negocio mediante el análisis de redes sociales.

Áreas de negocio que IBM promete solucionar con el uso de IBM Social Media Analytics.

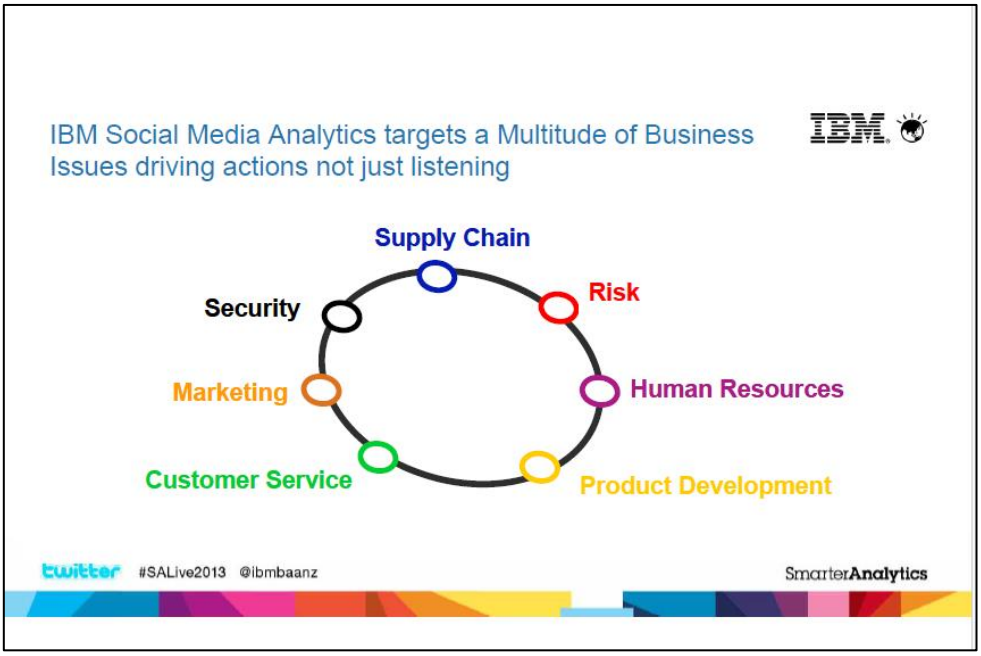


Ilustración 1.3 (IBM, IBM Smarter Analytics Libe 2013, 2013)

En la figura 1.3 se presentan las áreas de negocio en las cuales IBM promete soluciones en base al uso de su producto IBM Social Media Analytics.

El producto ofrecido por IBM es bastante avanzado debido a que no es solamente una herramienta por si sola, sino más bien un conjunto de herramientas que se valen de un marco de trabajo y modelo de madurez con el cual se busca hacer crecer a un negocio valiéndose de los datos que se pueden encontrar en las redes sociales.

El **Social Analytics Maturity Model-SAMM** (modelo de madurez de analítica de redes sociales) permite a cualquier negocio encontrar las oportunidades que necesite dentro del caos inicial que pueden parecer las redes sociales en un inicio. Mediante el uso de tres capas fundamentales dentro las cuales se resumen en **"escuchar", "pensar"** y **"actuar/hacer"**. Dichas capas complementadas con el **"cociente de analítica"** AQ (por sus siglas en inglés) permiten iniciar el camino hacia la madurez.

Para saber en qué nivel de madurez nos encontramos dentro del **SAMM** de IBM se debe hacer uso del ya antes mencionado cociente de analítica, el cual es una evaluación que se encarga de medir el provecho que obtenemos al utilizar las herramientas de analítica que actualmente posee el negocio al momento de aplicar el conocimiento obtenido de las mismas en decisiones, estrategias y procesos.

Una vez obtenido el puntaje del cociente de analítica podremos ubicarnos dentro de uno de los cuatro niveles propuestos por el SAMM de IBM. Los niveles clasificados en orden corresponden a:

- **Novato**
  Individuos o equipos analizan sus propios datos en redes sociales utilizando herramientas de consulta básicas u hojas de cálculo.

  En este estado lo importante es crear presencia dentro de las redes sociales donde el objetivo principal será el monitorear las mismas. Toda organización empieza en este estado.

- **Constructor**
  Las organizaciones que se encuentran dentro de estado están realizando sus primeras incursiones en la analítica de redes sociales de forma colaborativa. Todo esto sucede por lo general en un solo departamento, y por lo general está tarea recae sobre la gente responsable de mercadeo. Los equipos que están trabajando dentro de este departamento se encuentran bastante avanzados en la construcción de presencia y a la par analizan métricas básicas, como pueden ser el número de seguidores activos y reactivos.

- **Líder**
  Aquello que caracteriza a una organización líder es el hecho de que la analítica de redes sociales se ha integrado profundamente con los datos internos de la organización. Este tipo de organizaciones tienen ya definidas métricas operacionales y financieras en más de un departamento.

  Estas organizaciones líderes combinan los datos de las redes sociales junto con datos de varios sistemas para alcanzar una visión donde se encuentren integrados conocimiento y oportunidades.

  La organización ha formalizado un centro de excelencia con roles, responsabilidades, un ambiente compartido, procesos y tecnología estandarizados en la cual las partes interesadas continuamente evalúan, revisan estrategias y prioridades.

- **Maestro**

  Este tipo de organizaciones han integrado a la analítica de redes sociales dentro de su estrategia empresarial. Esto les permite construir y generar objetivos de arriba hacia abajo y asignar recursos basados en prioridades estratégicas en tiempo real mientras estás se complementan con la analítica de redes sociales.

  En este punto cualquiera sin importar su rol dentro de la organización conoce sus objetivos y como estos deben colaborar para lograrlos utilizando las redes sociales.

  Aquellos que están encargados de tomar las decisiones tienen disponible toda la información que necesitan al alcance de sus manos, sin importar que estos se encuentren en la mesa directiva evaluando movimientos estratégicos basados en redes sociales o tomando decisiones de venta

que involucran a los clientes y la retroalimentación que estos generan al usar los sistemas automatizados de la organización.

En este estado las decisiones que se tomen haciendo uso de la analítica de redes sociales debe ser obligatoria.

Son todos estos factores expuestos hasta el momento los cuales nos hacen darnos cuenta de lo necesario que es realizar un análisis de redes sociales en la actualidad, para prácticamente todas y cada una de las actividades sociales o comerciales que se busquen realizar. Con el fin de obtener el máximo beneficio en cada una de ellas al haber entendido primero al consumidor, cliente o usuario final de aquello que vendemos o promocionamos.

1.2 El Proceso del Análisis de Redes Sociales

Para poder capturar todos aquellos datos generados dentro de los océanos de información sin procesar que son las redes sociales, debemos seguir un proceso para su recolección, análisis y entendimiento. Con el objetivo de finalmente presentarlos indicando patrones o tendencias que buscamos exponer a un público general o específico.

Este proceso de análisis de redes sociales debe definir en un inicio que es aquello que buscamos probar con los datos obtenidos, una vez que estos hayan sido recolectados, analizados y procesados.

A continuación, se describen en orden los pasos necesarios para realizar un análisis de redes sociales:

- **Captura**
  Dentro de este paso definimos cuales son las “palabras clave” que vamos a buscar o capturar dentro de la red o redes sociales que vayamos a monitorear durante este proceso. Debemos tener en cuenta que “pescaremos” todo tipo datos por lo que solo algunos serán de utilidad. Los datos que buscamos encontrar durante este paso son aquellos que denotan actividades o intereses de los usuarios dentro de redes sociales como Facebook o Twitter.

  Parte de este paso también es preparar los datos ya recolectados para la siguiente etapa que es la de entendimiento.

  La manera correcta de prepararlos consiste en realizar algunos pasos de preprocesamiento como; modelar los datos, relacionarlos entre sí, extraer características y realizar operaciones semánticas y sintácticas que apoyen el análisis (Fan & Gordon, 2014).

- **Entendimiento**
  Una vez que hemos recolectado la cantidad de datos que consideremos indispensables para continuar con este paso, debemos buscarles sentido para la generación de métricas útiles que identifiquen comportamientos o patrones dentro de las redes sociales previamente analizadas.

Como los datos obtenidos en el paso anterior provienen de una o varias fuentes de información, es seguro que una porción de ellos contenga ruido y tenga que ser removido antes de realizar algún análisis a éstos. Para realizar la depuración podemos utilizar clasificadores de texto basados en reglas o datos ya previamente clasificados haciendo uso de técnicas derivadas a partir de minería de textos o de datos.

Este paso nos provee de información acerca de los "sentimientos" y comportamientos de los usuarios frente a un determinado tema, acontecimiento o producto que es provisto por un usuario, perfil o empresa.

La importancia de este paso es ser el núcleo de todo el proceso de análisis de redes sociales. Debido a que de aquí parten las métricas e información final que se llegará a desplegar en el último paso que es el de presentación.

- **Presentación**
  Los resultados obtenidos después de haber realizado diferentes análisis serán resumidos, evaluados y mostrados en un formato de fácil comprensión. Por lo que el uso de gráficos será útil para que la información pueda ser visualizada por todos aquellos a quien va dirigida o enfocada.

A continuación, se presenta de forma gráfica cómo funciona el proceso del Análisis de Redes Sociales:

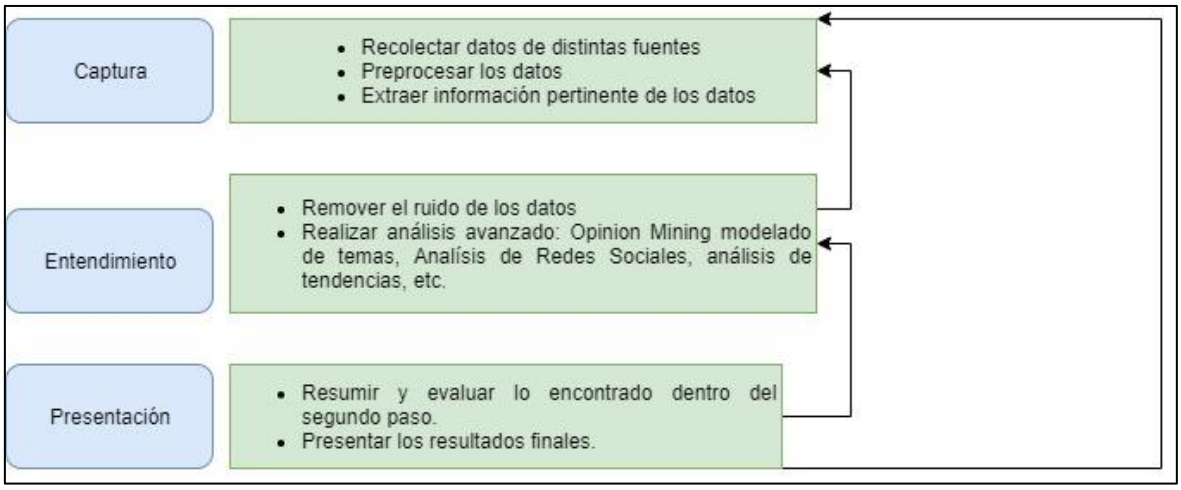


Ilustración 1.4 (Fan & Gordon, 2014)

Como se puede apreciar en la Ilustración 1.4, el proceso del análisis de redes sociales no es rígido. Debido a que pueden llegarse a dar iteraciones o retrocesos dentro del proceso, con el objetivo de ir refinando cada vez más los resultados finales y conclusiones que vamos encontrando a lo largo de nuestra investigación. Por lo que, si un paso nos dejó inconformes o con resultados inconsistentes siempre podemos regresar al anterior para cambiar el método de análisis, número de datos recolectados o forma en la que filtramos los mismos.

1.3 Los Estados del Análisis de Redes Sociales

Es importante revisar el significado del Análisis de Redes Sociales para poder diferenciarlo de otros conceptos (monitoreo de redes sociales e inteligencia de redes sociales) que muchas veces son utilizados como sinónimos.

1.3.1 ¿Qué es el Análisis de Redes Sociales?

(Khan, 2015, pág. 21) define al: "Análisis de Redes Sociales (ARS) cómo el arte y ciencia de extraer valiosas perspectivas ocultas dentro de grandes cantidades de datos semiestructurados y no estructurados provenientes de las redes sociales para permitir la toma de decisiones perspicaces".

Otra definición de este concepto es:

"Su objetivo principal es el de desarrollar y evaluar métodos científicos al igual que marcos de trabajo técnicos y herramientas de software para rastrear, modelar, analizar y minar grandes escalas de datos procedentes de las redes sociales para varios propósitos. Dentro del mundo de los negocios se puede llegar a considerar al **ARS** como un subconjunto de la inteligencia de negocios a la cual le conciernen temas como metodologías, procesos, arquitecturas, y tecnología que transforman datos en bruto de las redes sociales en información útil y significativa para los propósitos de un negocio" (Stieglitz, Dang-Xuan, Brans, & Neuberger, 2014, pág. 90).

1.3.2 ¿Qué es el Monitoreo de Redes Sociales?

Es un proceso que consiste en recolectar y monitorear que comenta el mundo dentro de las redes sociales. Estos comentarios pueden corresponder a hashtags, palabras claves acerca de una marca o sus competidores. A los datos recolectados no se los busca categorizar o entender.

Está técnica es bastante usada dentro del comercio electrónico para conocer qué opinan los consumidores acerca de una determinada marca o empresa.

1.3.3 ¿Qué es la Inteligencia de Redes Sociales?

Un proceso que abarca el monitoreo de redes sociales, la recolección y análisis del contenido de estas para brindar resultados útiles al momento de realizar decisiones estratégicas dentro de una empresa o marca.

1.3.4 Los Tres Estados del Análisis de Redes Sociales

A continuación, se explicarán **los tres estados** o formas que puede tomar el análisis de redes sociales:

- **Análisis Descriptivo**
  Consiste en recolectar y describir los datos de las redes sociales en forma de reportes, visualizaciones y grupos para entender un problema en concreto. El análisis de acciones (vistas, tweets, comentarios, etc.) y texto son ejemplos de un análisis descriptivo. Esto puede utilizarse para

entender los sentimientos del usuario o identificar tendencias emergentes al momento de agrupar los temas y opiniones encontradas. Actualmente este tipo de análisis es el más utilizado dentro del ARS.

- **Análisis Predictivo**
  Involucra el analizar grandes cantidades de datos acumulados de redes sociales para "predecir" un evento futuro. Como, por ejemplo, cuanta gente estará dispuesta a comprar, vender, usar o dejar de usar un determinado producto a partir de alguna intención expresada a través de las redes sociales.

- **Análisis Prescriptivo**
  Este tipo de análisis nos sugiere la mejor alternativa al momento de manejar algún escenario en concreto. Por ejemplo, si tenemos varios grupos de usuarios de los cuales hemos identificado patrones de compra. Con este análisis podríamos llegar a optimizar o personalizar la forma en la que se realizan ofertas a cada uno de estos grupos.

Capítulo 2 Análisis de Sentimientos

El capítulo actual explica el Procesamiento Natural del lenguaje y en que aplicaciones diarias lo podemos encontrar, además se muestra su utilidad en motores de razonamiento automático como WolframAlpha y Watson de IBM. También se aborda al análisis de sentimiento y los distintos retos con los que esta disciplina emergente tiene que lidiar. Finalmente se profundiza en los distintos niveles del análisis de sentimientos y cuál es el objetivo de cada uno de estos.

2.1 Procesamiento Natural del Lenguaje-PNL

El procesamiento natural del lenguaje es una parte de las ciencias de la computación e inteligencia artificial que trata con los lenguajes humanos.

Es probable que todos los seres humanos ya nos hayamos topado con aproximaciones del procesamiento natural del lenguaje y minería de texto sin siquiera haberlo notado. Algunas de dichas aproximaciones son la función de autocompletar y los correctores ortográficos, debido a que se encargan constantemente de analizar los textos que escribimos antes de enviarlos como correos electrónicos o mensajes. Otro ejemplo puede ser la función de autocompletar de Facebook la cual autocompleta el nombre que escribimos en la barra de búsqueda con lugares o amigos conocidos, esto se logra con una herramienta llamada “reconocimiento de una entidad nombrada” (named entity recognition). El objetivo de esta funcionalidad no es reconocer si lo que se escribe es un sustantivo o artículo, sino más bien lo que busca es encontrar a la persona o lugar al que uno se puede estar refiriendo y reconocerlo.

Otra empresa que también aplica PNL y minería de texto dentro sus herramientas es Google. Esto es bastante evidente debido a que todos alguna vez en nuestra vida hemos utilizado el buscador de internet Google y al buscar una palabra, tema o personalidad en particular encontramos varios resultados que no solo nos redireccionan a páginas de internet, sino que incluso pueden hacerlo a todo tipo de documentos ya sean presentaciones, archivos de texto, etc.

La forma en la que Google presenta todo tipo de resultados en internet al momento de ingresar cierto tipo de palabras se realiza de la siguiente manera según (Cielen, Meysman, & Ali, 2016):

- “Preprocesamiento de todos los documentos que recolecta para las entidades nombradas.
- Realizar identificación del lenguaje.
- Detectar a qué tipo de identidad se hace referencia.
- Empatar la consulta con un resultado.
- Detectar el tipo de contenido que se debe devolver (PDF, contenido sensible)”

Podemos apreciar que PNL y la minería de texto no necesariamente buscan encontrar el significado a los textos que analizan, sino que más bien su objetivo es relacionar a los textos/palabras con meta-atributos cómo es el lenguaje en el que están escritos o el tipo de documento al que pertenecen. Este tipo de aproximaciones han permitido la construcción y creación de “**motores de**

razonamiento automático" los cuales se encargan de procesar consultas en lenguaje natural para su funcionamiento.

Algunos "motores de razonamiento automático" conocidos dentro del medio de la tecnología son **WolframAlpha** y **Watson de IBM**.

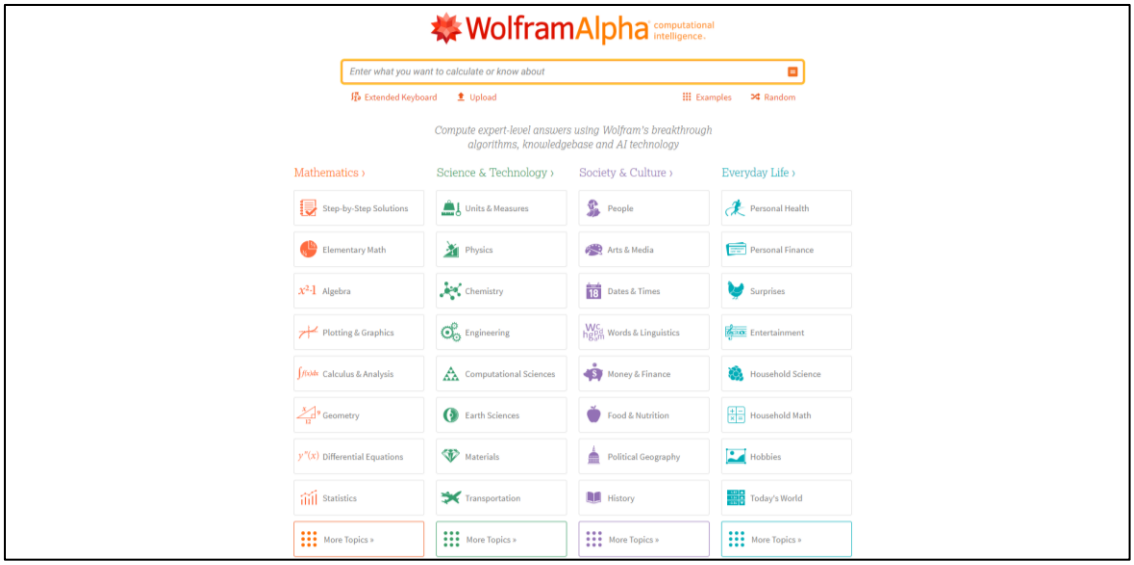


Ilustración 2.1 (Wolfram, 2020)

La Ilustración 2.1 se encarga de mostrar los distintos campos de acción en los cuales se puede utilizar a WolframAlpha.

Ilustración 2.2 (Wolfram, 2020)

La Ilustración 2.2 muestra cada uno de los componentes que construyen a WolframAlpha.

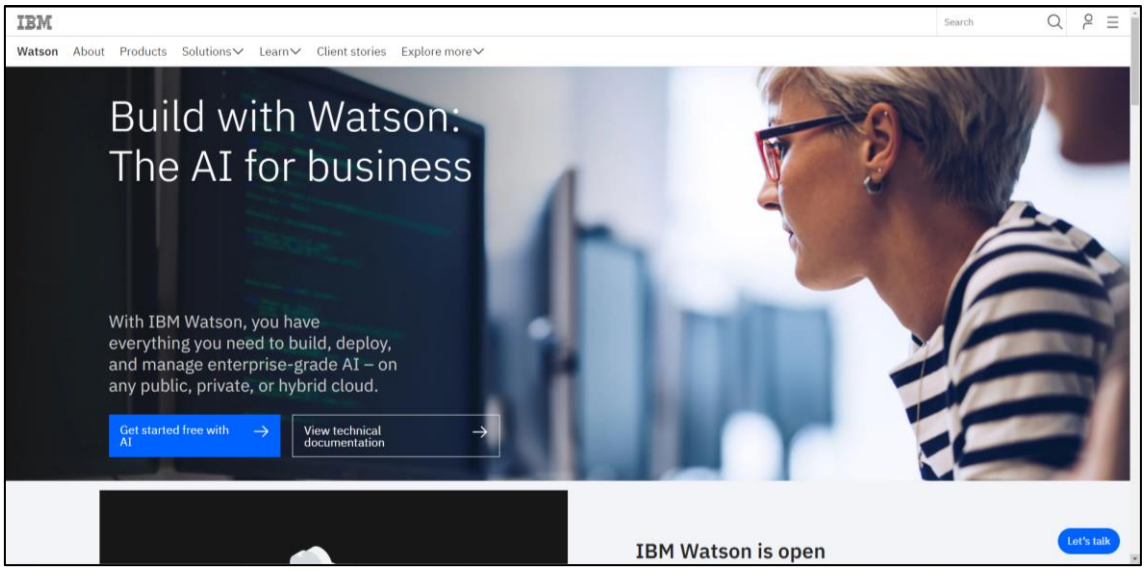


Ilustración 2.3 (IBM, IBM Watson Developer, 2020)

La Ilustración 2.3 muestra una propaganda de IBM Watson donde se indica que es “la inteligencia artificial para negocios”.

El caso de IBM Watson es particularmente interesante debido a que surgió como un proyecto de investigación y desarrollo en el año 2006 y fue mostrado por primera vez al público en 2011. El programa de televisión estadounidense Jeopardy mostro a Watson obteniendo la victoria sobre dos jugadores humanos.

Ilustración 2.4 (Cielen, Meysman, & Ali, 2016)

Ilustración 2.4 Watson de IBM obtiene una victoria contundente sobre dos jugadores expertos en el programa de televisión americano Jeopardy.

Desde aquel año 2011 donde se mostró la efectividad de Watson, su campo de aplicación y desarrollo se expandió primero al ámbito de la salud cómo Watson Health y después al de las finanzas cómo Watson for Financial Services en 2012.

Hoy en día Watson ofrece varias soluciones en las que incluye atención al cliente, cadena de suministro, riesgo y recopilación, publicidad, video, seguridad, etc.

Uno de los problemas que tanto WolframAlpha como IBM Watson pueden llegar a encontrar es la ambigüedad. Este problema también es propio del ser humano y es probable que en una situación donde se presente esta, un ser humano no se llegue a desempeñar mejor que estos motores de razonamiento automático a pesar de contar con cierto conocimiento en el campo. Otro problema también son las faltas ortográficas y las diferentes formas de pronunciar una misma palabra. Dentro de este problema es bastante probable que un ser humano sea capaz de identificar a las palabras a pesar de que se encuentren mal escritas o entender el significado de cierta palabra dado todo un contexto. Este tipo de soluciones son fáciles de resolver para nosotros los seres humanos, pero para las computadoras son retos bastante complejos mientras sus algoritmos no sean capaces de asociar ciertas cadenas de texto que dentro de su programación aparentemente no tienen ninguna relación.

La mejor forma de tratar de solucionar este tipo de problemas es crear algoritmos específicos para situaciones o contextos bien definidos. El crear un algoritmo multipropósito no es la solución ideal para tratar con un problema en particular, por ejemplo; un algoritmo de análisis de sentimientos diseñado para el idioma inglés o derivado de él no va a funcionar con efectividad al momento de trabajar con otros idiomas cómo pueden ser el español, francés, noruego, chino, etc.

2.2 La demanda de información sobre opiniones y sentimientos

La gran cantidad de usuarios que existen dentro de las redes sociales junto con sus reacciones, posturas y opiniones trajeron consigo nuevas técnicas y herramientas para su respectivo análisis. El surgimiento de estas nuevas metodologías para analizar los datos e interacciones de los usuarios dio como resultado el surgimiento del **análisis de redes sociales**.

El análisis de redes sociales trajo consigo distintas herramientas dentro las cuales se encuentra el **análisis o minería de texto**, el cual tiene como subcampo al análisis de sentimiento/minería de opinión.

Es importante conocer en que consiste el análisis de sentimientos para saber de qué manera ayuda a satisfacer la demanda de opiniones y sentimientos.

2.2.1 ¿Qué es el Análisis de Sentimientos?

“Es un nuevo tipo de **análisis de texto** que apunta a determinar la opinión y subjetividad de comentarios y reseñas.” (Collomb, Costea, Joyeux, Hasan, & Brunie, 2013) Aunque también se lo considera cómo una rama de las **Ciencias de la Computación** la cual abarca en gran parte el campo de **Machine Learning**, y la **Lingüística Computacional**.

Este campo ha ganado bastante popularidad en un mundo donde grandes portales de ventas en línea cómo **Amazon** acumulan dentro de sí una gran cantidad de opiniones, calificaciones y críticas sobre diferentes productos.

El reto del **Análisis de Sentimientos** es ser algo más que un simple análisis de texto basado en clasificación léxica. Debido a que muchas veces la cantidad de palabras positivas o negativas encontradas dentro de una opinión no revelan o indican la perspectiva global de dicha opinión. Es por ello por lo que el trabajo del **Análisis de Sentimientos** debe ser encontrar nuevos métodos de clasificación que no simplemente indiquen la polaridad de una opinión, sino que más bien se enfoquen en mostrar como resultado final la opinión general de un tema.

En algún instante de nuestra vida hemos deseado saber que piensa alguien o el resto acerca de nosotros antes de decir algo o tomar una decisión. Debido al miedo interno que provoca la incertidumbre del momento. Hoy en día prácticamente no tenemos que preocuparnos por la incertidumbre. Ya que en la actualidad tenemos al instante todo tipo de opiniones y recomendaciones sobre un tema, producto o decisión dentro de redes sociales cómo Twitter, Facebook o foros como Reddit. Las cuales pueden contener textos de 140 caracteres, posts complejos de varios párrafos o un foro de cientos de páginas dedicadas a un tema en específico.

A pesar de contar con aquella gran biblioteca de conocimientos y recomendaciones dentro de las redes sociales, muchos de los problemas que presentan la información y los datos que encontramos dentro de ellas son los siguientes:

- Información incompleta.
- Información confusa.
- Información o datos falsos.
- Información abrumadora.

Provocando de esta manera que la disciplina emergente del Análisis de Sentimientos lidie con el tratamiento computacional de las opiniones, sentimientos y subjetividad que se encuentra dentro de un texto. Con el objetivo en mente de crear un mejor sistema de acceso a la información para que los usuarios obtengan mejores experiencias al momento de buscar información o datos que abarquen el tema de su preferencia.

2.3 Clasificación de Sentimientos

La clasificación de sentimientos es un campo que se encuentra dentro del Análisis de Sentimientos, el cual se encarga de clasificar documentos enteros en base a la opinión de un tema. Otro tipo de clasificación de sentimientos es aquel que se enfoca solo en las características de un objeto más que en sus opiniones.

Los dos lenguajes donde más estudios sobre clasificación de sentimientos se han llevado a cabo son el inglés y el chino. Actualmente existen muy pocos estudios en lenguajes cómo el árabe, italiano y tailandés.

Los estudios de análisis de sentimientos que se apoyan de la clasificación de sentimientos toman datos sobre opiniones de usuarios para básicamente “juzgar la polaridad de sus comentarios”. Haciendo uso de los niveles del Análisis de Sentimientos.

2.3.1 Niveles del Análisis de Sentimientos

- **Nivel de Oración**
  Este nivel está enfocado en determinar si una oración es positiva, negativa o neutral. Por lo general se considera neutral a una oración cuando no expresa una opinión (polaridad).

  (Liu, 2012) afirma que este nivel "es parecido a la **clasificación de subjetividad**, la cual se encarga de diferenciar **oraciones objetivas** de aquellas **oraciones subjetivas** que expresen hechos sobre la opinión que proporcionan las oraciones" p,11. Otros autores cómo (Kharde & Sonawane, 2016) afirman que la **clasificación de subjetividad** es el primer paso que se debe realizar dentro del nivel de oración.

  La determinación del grado de una oración se da cuando se encuentra la orientación individual de las palabras de una oración y se las combina para determinar el sentimiento colectivo de la misma.

- **Nivel de Documento**
  El objetivo es clasificar toda una opinión negativa o positiva expresada sobre un documento. Este nivel solo funciona cuando todo el documento habla de **un solo** producto, tema, entidad, etc.

- **Nivel Basado en Características o Aspecto**
  Este nivel busca identificar y extraer características de un producto o entidad a partir del cual se tengan datos. Por lo general los datos se obtienen de una reseña, opinión o comentario de la cual se puedan extraer características que serán clasificadas cómo positivas, negativas y neutrales.

- **Nivel de Palabra**
  Este nivel se encarga de clasificar a las palabras según su polaridad tomando en cuenta principalmente a los adjetivos dentro de oraciones o documentos.
  El clasificar la polaridad de las palabras se usa dentro de otros niveles cómo el paso previo a clasificar un **documento** u **oración**.

Las distintas aproximaciones que el Análisis de Sentimientos puede tomar son:

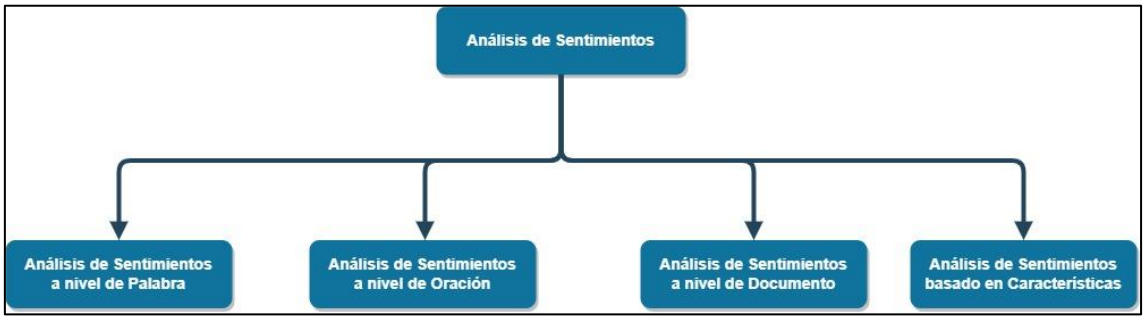


Ilustración 2.5 (Kharde & Sonawane, 2016)

La ilustración 2.5 muestra los cuatro niveles del Análisis de Sentimientos.

Independientemente del nivel escogido se pueden aplicar dos técnicas (machine learning o enfoque basado en la orientación semántica) que se llegarán a complementar con el procesamiento natural del lenguaje “el cuál se utiliza especialmente en la detección de sentimientos en documentos. El PNL comparte varias características con otras disciplinas como la extracción de información y minería de texto, lingüística computacional, psicología y análisis predictivo” (Vinodhini & Chandrasekaran, 2012).

Capítulo 3 Recolección de Datos

El siguiente capítulo profundiza en los datos y sus formas en el mundo moderno, además se expone como empresas del tamaño de Facebook, Microsoft o Tesla destilan los datos de sus consumidores para obtener beneficios. También se busca exponer la importancia de la minería de datos y quienes participan dentro de ella (científicos e ingenieros). Otros temas cómo la minería de texto, el algoritmo de Porter y la lematización son revisados previo a la preparación de preguntas internacionales y nacionales que serán resueltas en futuros capítulos de la disertación. Finalmente se prepara el entorno de trabajo en R Studio y se capturan los primeros datos en inglés y español.

3.1 ¿Qué son los Datos?

Generalmente se define a los datos cómo cualquier conjunto de caracteres que han sido agrupados y procesados para un propósito determinado, el cual en la mayoría de los casos es el análisis. Si los datos recogidos no poseen un contexto estos no tendrán ningún sentido para los seres humanos o las computadoras al momento de realizar alguna interpretación.

La definición anterior es bastante simple y queda bastante corta debido a que en la actualidad no solo contamos con **datos estructurados** a los cuales comúnmente interpretamos como datos organizados en tablas las cuales almacenan texto y valores numéricos que pueden relacionarse entre sí. Sino que también contamos con **datos semi estructurados** los cuales por lo general son almacenados en **estructuras de datos tipo árbol** que pueden ser manipuladas mediante el uso de **bases de datos NoSQL** y almacenadas en archivos tipo **JSON, XML, YAML**. Finalmente, también tenemos aquellos datos que no se encuentran estructurados en su totalidad y que prácticamente son la mayor cantidad de datos que podemos encontrar fuera de las **Bases de Datos SQL** y **NoSQL.** A este tipo de **datos no estructurados** es común encontrarlos en **Data Lakes** o **Data Warehouses**, debido a que son archivos de todo tipo, cómo texto, música, video, fotos, etc.

Esto demuestra que aquella realidad ideal en la que todos los datos están 100% estructurados y son fáciles de recuperar con un par de consultas SQL para satisfacer las necesidades de una compañía, organización o científico de datos está completamente errada. Debido a que los datos en la mayor parte de los casos por no decir en todos los casos se encuentran de forma desorganizada. Estos no están donde tendrían que estar, pueden encontrarse en formatos desconocidos, poco familiares, también podrían estar sesgados o simplemente no existir. Estas situaciones son de los más común para un científico datos y se presentan cómo un reto u oportunidad para que este se familiarice con los datos en sus distintas formas y cualidades.

El mundo de los datos hoy en día es un entorno salvaje y de gran valor para su exploración e investigación. Llegados a este punto dentro de la historia de la humanidad los datos y su respectivo valor no pueden llegar a ser ignorados. Por lo que para obtener aquel valor estos deben ser limpiados, organizados y preparados.

Aquellas personas, compañías o conglomerados que mejor sepan a aplicar las distintas técnicas de depuración que existen dentro del mundo de los datos

podrán **"explotar y refinar"** toda la información que los millones de usuarios del mundo digital generan a diario. Con el objetivo de beneficiarse al aprender sobre las distintas conductas de búsqueda, compra y movilización que los usuarios presentan frente a un producto, evento, campaña publicitaria o celebridad.

Está explotación masiva de datos para obtener beneficios ha hecho que medios de gran relevancia en el mundo como la revista **The Economist** mencionen que "el recurso más valioso actual ya no es el petróleo, sino los datos" (Economist, 2017). Debido a los grandes ingresos y poder que los grandes de la tecnología actual representan no solo en Estados Unidos (qué es donde se encuentran la mayoría) sino el mundo entero. Llegando a comparar a gigantes cómo **Google**, **Amazon**, **Microsoft** y **Facebook** con la **Estándar Oil Company** debido a que muchos consideran que estas compañías tienen **monopolizada** a la **economía de los datos** de la misma forma en la que en el siglo pasado la Estándar Oil Company tenía monopolizada a la economía del petróleo**.**

La preocupación de un posible **monopolio digital** es latente en el ambiente, pero ¿realmente sería una buena idea desintegrar a estas grandes compañías en otras pequeñas? Muchos tal vez coincidan en una respuesta negativa a pesar de que Facebook haya presentado prácticas monopólicas en un pasado reciente al haber comprado primero a **Instagram** en 2012 y luego a **WhatsApp** en 2014, los cuales se presentaron como una posible competencia a su modelo de negocios (obtención de datos a cambio de un servicio) y mercado el cual hoy en día debe compartir con **Google**, **Amazon**, **Tesla**, **Microsoft** y los recientes **General Electric** y **Siemens** que se han etiquetado como compañías digitales y centradas en los datos. Una posible respuesta positiva a la pregunta anterior generaría probablemente pequeñas empresas dedicadas a los datos que no puedan satisfacer la demanda de servicios que los gigantes actuales son capaces de suministrar.

El modelo de negocios de la **economía digital** no se basa en que **calidad sea igual a cantidad**, sino más bien en **cuanta calidad se puede extraer de una cantidad depurada de datos.** Esta realidad la ha entendido muy bien la empresa **Tesla Inc.** los cuales en el año 2016 tuvieron mayor cotización en bolsa que **General Motors** a pesar de vender menos automóviles que estos. Este éxito se debe al entendimiento e importancia que les dan a los datos que recogen de sus vehículos autónomos los cuales no solo les proporcionan formas de mejorar su producto o sistema de conducción sino también una ventaja a la hora de acercarse a sus clientes.

A continuación, podemos observar un par de ilustraciones que demuestran cómo se encuentra actualmente la cotización en bolsa de Tesla Inc. y General Motors después del artículo publicado por The Economist en 2016.

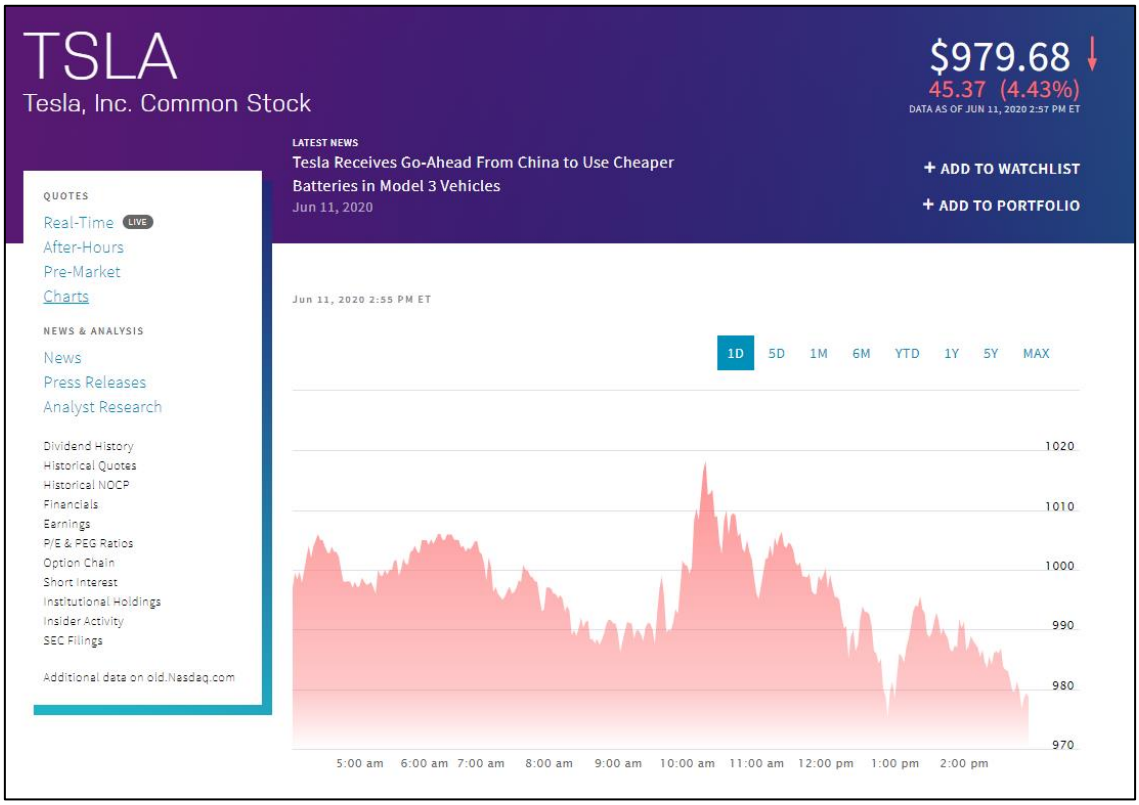


Ilustración 3.1 (NASDAQ, Tesla Inc. Common Stock, 2020)

La Ilustración 3.1 muestra el valor de las acciones de Tesla Inc. el día 11 de junio de 2020.

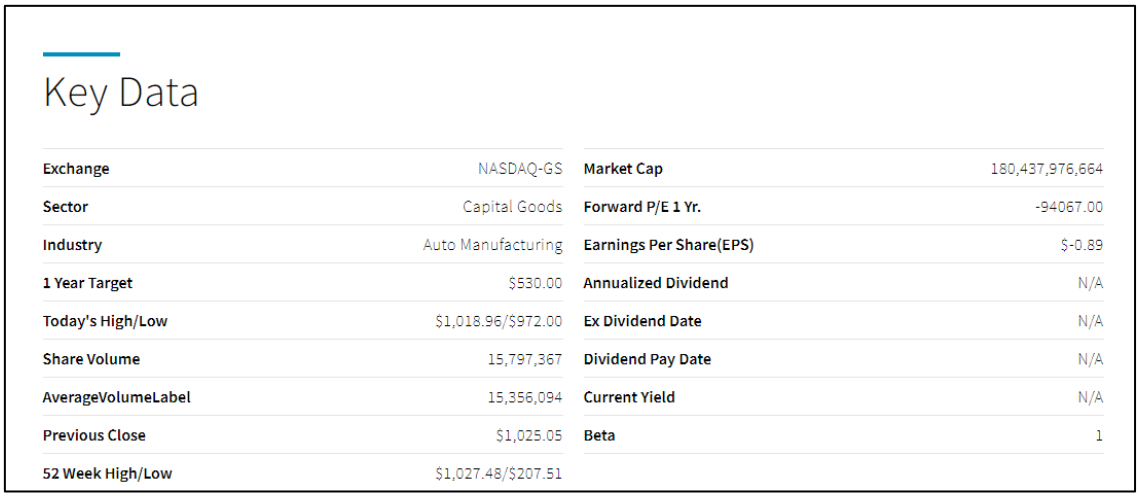


Ilustración 3.2 (NASDAQ, Tesla Inc. Common Stock, 2020)

Ilustración 3.2 datos claves de Tesla Inc. durante el 11 de junio del 2020, donde su capitalización en el mercado corresponde $ 180,437,976,664 dólares americanos.

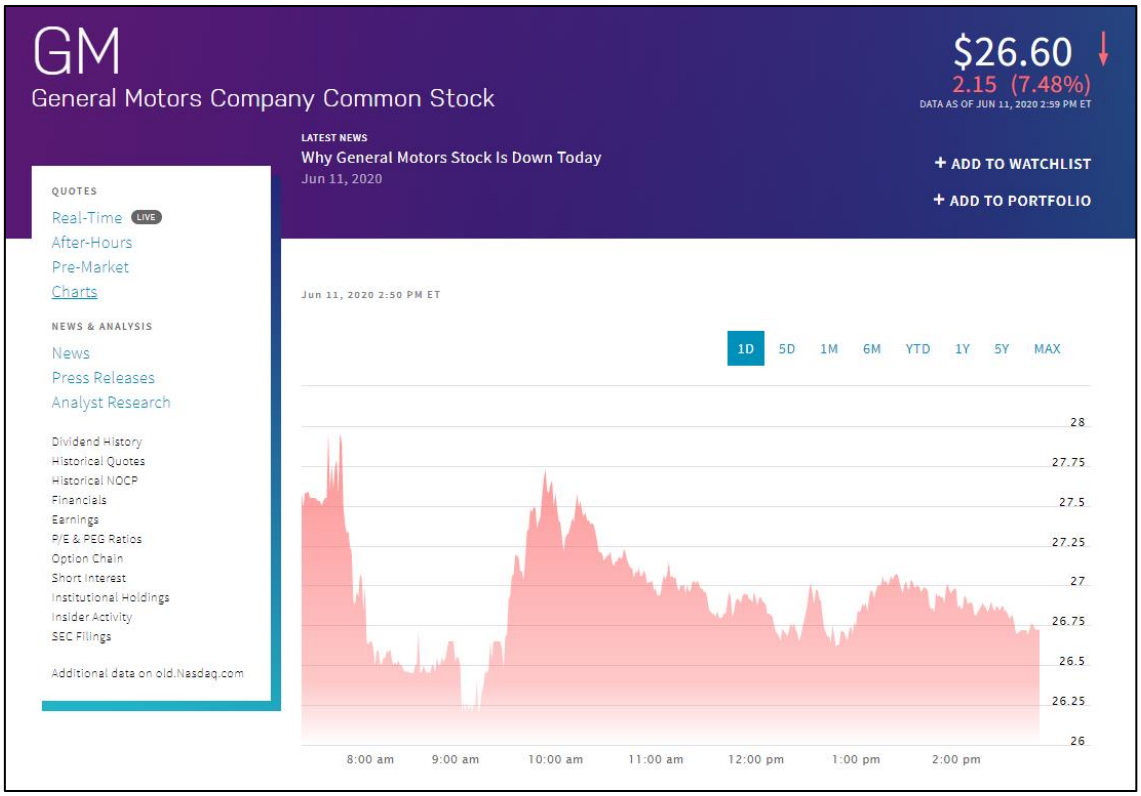


Ilustración 3.3 (NASDAQ, General Motors Company Common Stock, 2020)

La Ilustración 3.3 muestra el valor de las acciones de General Motors Company el día 11 de junio de 2020.

Ilustración 3.4 (NASDAQ, General Motors Company Common Stock, 2020)

La Ilustración 3.4 muestra datos claves de la General Motors Company durante el 11 de junio del 2020, donde su capitalización en el mercado corresponde a $38,469,451,842 dólares americanos.

A partir de lo observado en los gráficos anteriores podemos concluir que hoy en día no basta el simple hecho de vender un producto en grandes cantidades para generar mayor valor a nuestra empresa, sino que más bien lo que importa es cuanto valor podemos obtener de los datos generados por nuestros productos y

clientes para de esta manera mejorar y refinar nuestro acercamiento a nuevos consumidores a la vez que perfeccionamos el producto o servicio que ofrecemos.

Este uso de datos por parte de los gigantes de la industria tecnológica ha sido cuestionado por gobiernos, medios de comunicación y usuarios en todo el mundo. A pesar de que son ellos mismos los cuales "pagan" con sus datos a cambio de un perfil en Instagram, Twitter o un envío de Amazon Prime cuando dan al botón de aceptar en el ya famoso cuadro de dialogo de **Términos y Condiciones** que todos consienten ciegamente. Dentro de esta situación los usuarios tienen bastante culpa, pero las compañías también podrían ser más "transparentes" y notificar a los usuarios sobre cuantos datos e información poseen de cada uno de sus usuarios y como la están utilizando. El tema de conocer que se hace con nuestros datos y su privacidad probablemente sea uno de los más populares en los años venideros debido a la gran cantidad de escándalos sobre filtración y venta de datos que se han suscitado en años recientes, donde el caso más famoso seguramente es aquel en el que estuvieron involucrados **Facebook** y **Cambridge Analytica**.

Una posible solución al problema de la privacidad y protección de nuestros datos sea la creación de leyes universales o regionales que le exijan a las compañías publicar cuanta información y beneficio poseen de los datos de sus usuarios con el objetivo de evitar que se atente contra ellos mismos en un futuro. Esto con el propósito de evitar ser dominados por el "monopolio digital" de los datos a la vez que ofrecen nuevos servicios confiables y seguros a sus ciudadanos.

La definición ideal que tenemos sobre los datos se aleja bastante de la realidad actual. En la cual estos se han vuelto más valiosos que el petróleo y representan la nueva moneda de pago en la sociedad moderna. Donde quienes saben almacenarlos, interpretarlos y procesarlos pueden obtener conclusiones bastante interesantes a partir de compras, movilización y publicaciones dentro de todos los medios digitales que encontramos en nuestro día a día.

3.2 Minería de Datos

La problemática actual en el mundo de las ciencias de la computación y de la ciencia en general es el hecho de que nos encontramos rodeados de datos y no de información. Esto se debe a que con el paso del tiempo la tecnología ha abaratado sus costos y cualquier ser humano es capaz de crear un perfil en una red social o administrar su propio sitio web por un módico precio a la par que un científico molecular realiza estudios sobre el genoma humano y almacena los datos de su investigación en un servidor externo.

La gran cantidad de datos almacenados en servidores y bases de datos crece a una velocidad impresionante, provocando que año tras año estos ya no solo se almacenen en terabytes, sino que incluso los lleguen a almacenar en petabytes o zettabytes. Este crecimiento desenfrenado no solo produce nuevos datos como se mencionó con anterioridad, sino que también provoca que datos realmente valiosos queden almacenados en un disco duro en algún lugar del mundo. Muchos de estos datos que han quedado atrás podrían contener respuestas, información y soluciones a muchos de los problemas con los que tenemos que lidiar en la actualidad cómo podrían ser el cambio climático, enfermedades, migración, distribución de espacios urbanos, etc.

Aquellos profesionales que dedican sus vidas e investigaciones dentro de la medicina, ciencia, ingeniería y negocios necesitan sacar provecho de todo dato que puedan, para generar soluciones y encontrar respuestas con las que logren beneficiar a la sociedad. Es por aquello que muchos investigadores dedicados a la búsqueda de datos han trabajado en conjunto durante los últimos tiempos para desarrollar una forma óptima en la cual se pueda capturar y procesar datos que no se encuentran en formatos convencionales, con el objetivo de generar información con la cual alimentar la toma de decisiones futuras y responder incluso a situaciones de vida o muerte. Es esta colaboración en conjunto la que ha dado origen a la minería de datos.

Los métodos tradicionales de recuperación e incluso de captación de información ya no responden a todas la necesidades y soluciones que deben proveer. Debido a que el volumen de los datos se ha inflado a niveles insospechados. La era del Big Data es el reto que los científicos de datos e ingenieros de datos deben afrontar. De ellos depende explotar este recurso que hoy en día vale más que el petróleo, debido a que son ellos los encargados de armar los oleoductos que van a captar, transformar y presentar la información que obtengamos de nuestra minería de datos.

3.2.1 Bases de Datos, Data Lakes, Data Warehouses y Data Swamps.

Las formas convencionales de almacenado y procesamiento de datos surgieron durante la segunda mitad del siglo **XX**, para ser más preciso durante los años sesenta y setenta cuando existieron los primeros postulados, estudios e investigaciones sobre bases de datos y sus respectivos modelos. Estos primeros acercamientos fueron realizados por Charles Bachman y Edgar Frank Codd respectivamente.

Bachman es reconocido por haber desarrollado en el año de 1962 el primer prototipo funcional de **IDS** (Integrated Data Store) el cual supuso la creación del primer sistema gestor de base de datos más conocido como **DBMS** por sus siglas en inglés. De igual manera Codd una década más tarde creo el modelo relacional de base de datos el cual es el corazón de los **DBMS** modernos. Algo que es de destacar y recalcar sobre estos dos grandes avances es que estos surgieron mucho antes de que un lenguaje de consulta de alto nivel existiera. A pesar de aquello las primeras versiones de **IDS** funcionaban utilizando **GECOM** el cual fue un lenguaje desarrollado por **General Electric**. La popularidad de **IDS** hizo que años más tarde el **CODASYL** lo considerará como el estándar para los sistemas gestores de bases de datos futuros a pesar de una gran negativa por parte de **IBM** el cual también contaba con un sistema gestor de bases de datos propio, el cual cabe recalcar que era bastante inferior al creado por Bachman. La estandarización del **IDS** haciendo uso del lenguaje de programación **COBOL** le permitió llegar a **IDS** a los mainframes insignia de **IBM** (IBM System/360) como un port de nombre **IDMS** (Integrated Database Management System). En los años posteriores a estos Bachman y Codd tuvieron bastantes discusiones sobre qué modelo de base de datos era mejor o cual era el correcto para utilizar, por su parte Bachman defendía arduamente a su modelo de relaciones basadas en redes y Codd defendía su modelo de relaciones basadas en algebra y calculo lineal. En la actualidad ambos modelos han persistido hasta nuestros días, el

modelo más utilizado y popular es el de Codd, mientras que el de Bachman es usado por mainframes de alto procesamiento.

Las dos aproximaciones anteriores construyeron el mundo digital en el que nos encontramos actualmente, pero hoy en día ya no satisfacen todas las necesidades del mundo moderno, en el cual un electrodoméstico, un satélite o un radar producen una gran cantidad de datos durante el transcurso del día. Muchos de estos datos no se almacenan de manera tradicional en forma de tablas, registros, nodos, redes o árboles. Sino que se encuentran en formato **JSON**, **XML**, **YAML**, mp3, mp4, archivos de texto plano etc. Son todas estas problemáticas las que han obligado a crear nuevas formas de almacenamiento. Estas nuevas formas de almacenamiento moderno fueron creadas para lidiar con grandes volúmenes de datos y todos sus formatos, estas soluciones reciben el nombre de:

- **Data Warehouse**
  Descrito como una base de datos empresarial o un almacén electrónico de donde una organización mantiene grandes cantidades de información. Esta información debe estar almacenada de forma segura y fiable, además debe ser fácil de recuperar y administrar. Se considera que su creación es el primer paso hacia la inteligencia de negocios (BI).

  El concepto de **Data Warehouse** surgió en el año de 1988 en **IBM** como un trabajo realizado por parte de los investigadores Barry Delvin y Paul Murphy. A pesar de aquello el que es considerado como el padre del **Data Warehousing** es el señor William H. Inmon, debido a que describió al Data Warehouse como una colección de datos que se encuentran orientados a un tema específico, integrado, que es capaz de variar en el tiempo y no ser volátil que sea capaz de soportar el proceso de la toma de decisiones.
  El Data Warehouse dentro de una empresa se puede alojar dentro de un servidor corporativo o en la nube. Este se encarga de funcionar como un repositorio unificado para todos los datos que se captan de varios sistemas dentro de una organización.

  El éxito de su aplicación depende del enfoque que se le dé al momento de escoger la arquitectura de Data Warehouse que se busca utilizar. Las arquitecturas dentro del Data Warehouse son tres:

  - **Básica**
    Sistemas operativos y archivos planos proporcionan datos en bruto que se almacenan junto con los metadatos. Lo usuarios finales de esta arquitectura pueden acceder a ellos para el análisis, generación de informes y minería.
  - **Básica con Área de Ensayo**
    Se añade un área de ensayo ente las fuentes de datos y el almacén de datos. Esta área permite limpiar los datos antes de que estos lleguen a entrar al almacén.

  - **Data Marts**

Estos son sistemas diseñados para un área del negocio en particular.

Su concepto y construcción parten de aquello que llamamos **ETL** (Extraer, Transformar y Cargar) donde cada letra representa las operaciones que se realizan dentro de la organización:

Extracción: obtener información de fuentes tanto internas como externas.

Transformación: corresponde a las tareas de filtrado, limpieza, depuración, homogenización y agrupación de la información.

Carga: actualizar y organizar a los datos y metadatos dentro de la base de datos.

- **Data Lake**
  Es un entorno de datos en bruto los cuales se encuentran en cualquier tipo de extensión o estructura. La cantidad de datos almacenados en un Data Lake común puede llegar a rozar el petabyte.

  Está herramienta puede parecer la antítesis del Data Warehouse debido a que los datos aquí no se encuentran relacionados y estructurados en tablas u hojas de cálculo. La idea de esta nueva forma de almacenamiento no tradicional es que está pueda llegar a complementarse con Data Warehouse y no ser su competencia. Un negocio u organización que busque mejorar implementando **BI** debería tener trabajando a amabas herramientas de manera simultánea.

  El uso principal del Data Lake es el de servir como fuente de explotación y descubrimiento de datos por parte de los científicos de datos dentro de una organización. Cada uno de los elementos dentro del Data Lake se encuentran allí hasta que sea necesario, durante aquel tiempo estos son asignados un identificador único y agrupados con un conjunto de metadatos.

  Algunos de los principales beneficios de un Data Lake son:

  - La **flexibilidad** de un Data Lake permite **normalizar**, **enriquecer** y **agrupar** los datos según sea necesario para la toma de decisiones dentro de la organización.

  - Los usuarios de todos los departamentos de la organización pueden tener acceso al Data Lake y su contenido para la **construcción de soluciones específicas de negocio**.

  - Al poner los datos en las manos de más personas dentro de la organización esta se vuelve más **"inteligente", ágil e innovadora**.

- **Data Swamp**
  El principio y estructura del Data Swamp es el mismo que el de un Data Lake, debido a que aquí también existen todo tipo de archivos con sus respectivas extensiones. Aquello que lo convierte en un **Swamp (pantano)** es la falta de organización, categorización e identificación con la ayuda de metadatos. Esto provoca que simplemente sea un repositorio de datos que puede acumular varios petabytes que no sean capaces de aportar algún tipo de valor al negocio u organización que lo posea.

  Este tipo de repositorio de datos debe ser evitado a toda costa por cualquier organización debido a que representa un dolor de cabeza principalmente a los científicos e ingenieros de datos que buscan realizar tareas de minería o análisis.

  Es muy probable que varias organizaciones en el mundo sean poseedoras de un Data Swamp sin siquiera saberlo debido a que han mal entendido o confundido el famoso termino anglosajón "**Schema on Read**" el cual dice que los datos solo se usan cuando se necesitan. Por lo que lo más probable es que estás organizaciones almacenen todo tipo de datos sin ningún criterio contaminando sus lagos de datos. A todos aquellos datos que son almacenados dentro de un Data Lake sin ningún criterio se les llama "piscinas estáticas" debido a se encuentran sobresaturadas de **datos no curados**. Afortunadamente los Data Lakes contaminados pueden ser limpiados haciendo uso de la curación de datos y de un buen Gobierno de Datos dentro de la organización.

  Las claves para limpieza de un Data Swamp son:

  - **Priorizar el Gobierno de Datos**
    Esto significa conocer las prioridades del negocio para que la construcción del Data Lake vaya de acuerdo con las necesidades del negocio.
    Una vez creado el Data Lake, este debe ser cuidado y mantenido con regularidad.

  - **Curar el Contexto de los Metadatos**
    Cuando se habla de curar a los datos se está hablando en realidad de curar el contexto de los metadatos. Una recomendación sobre cómo solucionar este problema es la implementación de un sistema federado de archivos en el cual existen unidades más pequeñas encargadas de mantener la autoridad sobre repositorios más simples. Una plantilla para implementar dicho modelo es el **Sistema OASIS** (**Arquitectura de almacenamiento basada en objetos para la Escalabilidad, la Inteligencia y la Seguridad**) el cual es un sistema de gestión de archivos que permite crear porciones de archivos a los cuales posteriormente empaquetará de forma lógica y física como una entidad.

  Sobrecurar o sobre organizar un Data Swamp no hace que regrese a ser un Data Lake, más bien lo que hace es convertirlo en un Data Warehouse. El mantener un Data Lake de manera saludable sin que caiga en Data

Swamp o Data Warehouse accidental depende del mantener un equilibrio entre cuanto almacenamos y organizamos.

3.2.2 Ingenieros de Datos, Científicos de Datos y la construcción del Conocimiento.

Estas nuevas formas almacenar de datos han comenzado a ganar popularidad entre las organizaciones recientemente, debido al boom de las redes sociales y otros medios digitales. Aquello que muchos ignoraron durante años como es el poder de los datos hoy en día es una receta asegurada para el éxito. El complementar Data Lakes con Data Warehouse permite mantener al día los informes de negocio a la vez que un científico de datos encuentra una nueva tendencia en datos históricos que fueron almacenados en formatos poco convencionales.

El mantenimiento, construcción y explotación de todas estas nuevas fuentes de datos, información e incluso conocimiento depende de individuos bastante hábiles y perspicaces. Este pequeño grupo de expertos por lo general representa una parte bastante pequeña dentro de cualquier organización y aun así son seguramente una de las partes más importantes. Este grupo de elegidos corresponde a los **ingenieros y científicos de datos.**

Muchas veces el terminó ingeniero de datos y científico de datos son utilizados como sinónimos de forma errónea debido a que se cree que ambos realizan las mismas tareas dentro del mundo de los datos y **Big Data** en tiempos más modernos. Lo cual es erróneo debido a que cada uno de los roles tiene sus propias tareas y responsabilidades dentro de cualquier organización.

Las responsabilidades propias de cada uno son:

- **Ingeniero de Datos**
  Su trabajo consiste en recolectar datos de diferentes fuentes, optimizar las bases de datos del negocio para análisis, remover archivos corruptos. Además, también se encarga de desarrollar, construir, probar y mantener la arquitectura de datos.

- **Científico de Datos**
  Su trabajo consiste en explotar los datos almacenados dentro de las bases de datos valiéndose de la arquitectura de datos diseñada por el ingeniero de datos para producir resultados. La característica principal del científico de datos es su capacidad analítica.

Su trabajo en conjunto les permite crear “oleoductos de datos”, los cuales servirán para extraer, procesar y producir resultados que terminarán en información. Está información al haber sido asimilada por la organización debe llegar a convertirse en **conocimiento**.

La **construcción del conocimiento** a partir de un "oleoducto de datos" donde los datos proceden de una sola fuente se presenta en la Ilustración 3.5:

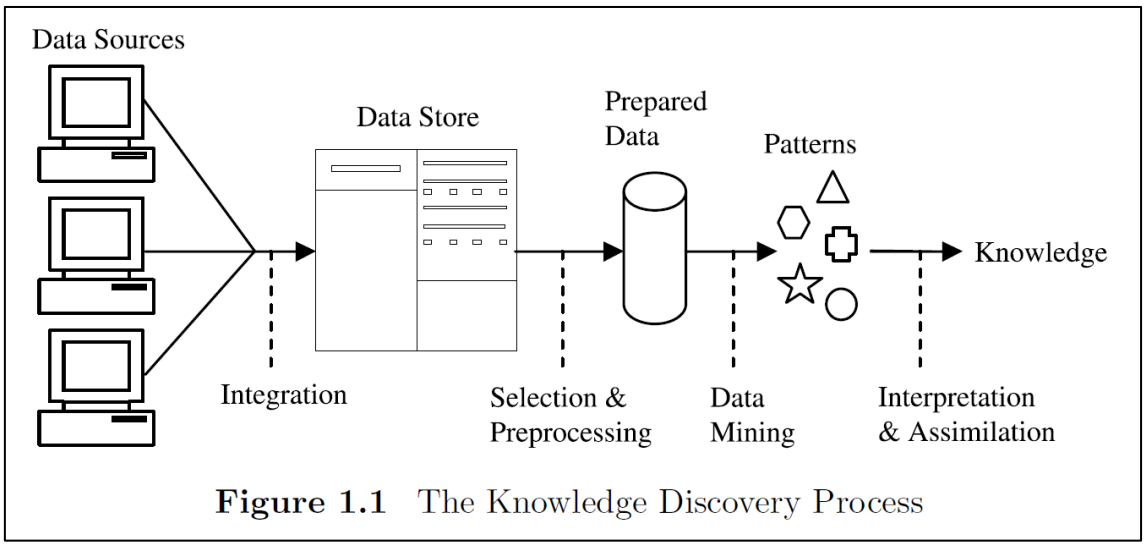


Figure 1.1 The Knowledge Discovery Process

Ilustración 3.5 (Bramer, 2016)

Una aproximación más moderna de la construcción del conocimiento es la que plantea Hadrien Lacroix en su curso "**Data Engineering for Everyone**" del portal **DataCamp** en la cual construye para su ejemplo una aplicación ficticia llamada Spotflix.

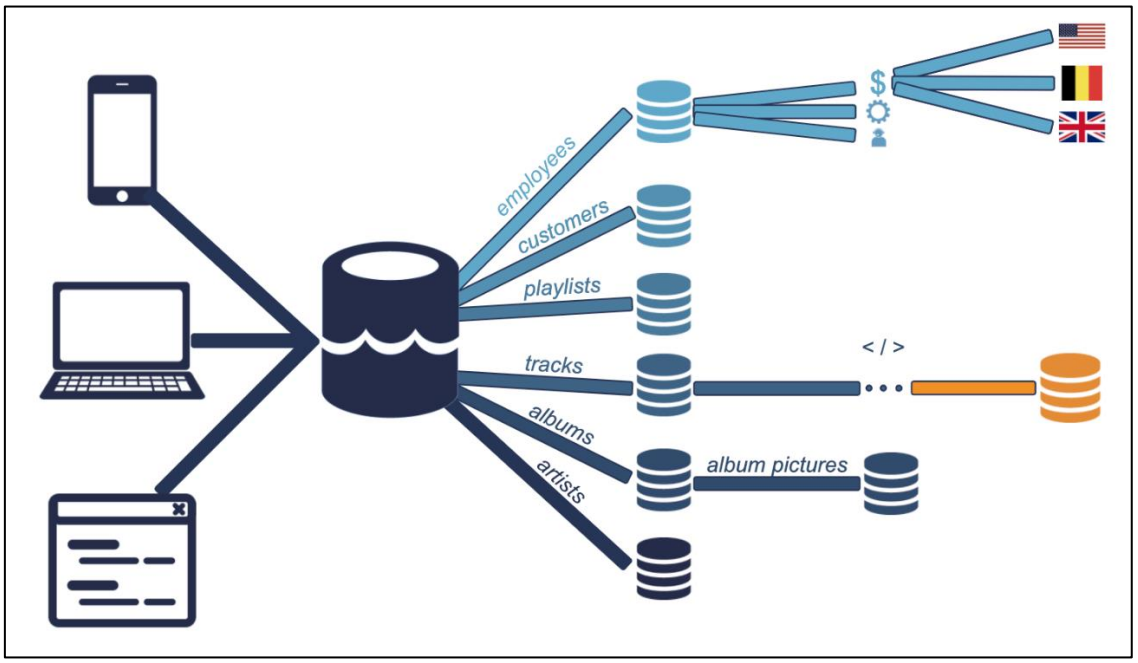


Ilustración 3.6 (Lacroix, 2020)

La ilustración 3.6 muestra la construcción de un "oleoducto de datos" moderno.

Como se puede apreciar, ambas aproximaciones son correctas pero la segunda es mucho más acertada y acercada a la realidad debido a que los **Data Lakes** modernos se componen por lo general de datos y metadatos que proceden de dispositivos móviles, computadores e incluso sitios web.

Esta construcción de oleoductos permite la correcta minería y explotación de los datos que posee la organización. Estos datos no son solo utilizados para mejorar y agregar valor al negocio sino también son usados para implementarlos en procesos de contratación, promoción y pago de los empleados.

3.2.3 El trabajo de la Minería de Datos

Gracias al crecimiento desmesurado e inimaginable de los datos hoy nos encontramos en la era del **Big Data.** La cual se presenta como una oportunidad inmejorable para construir, replantear y descubrir conocimiento.

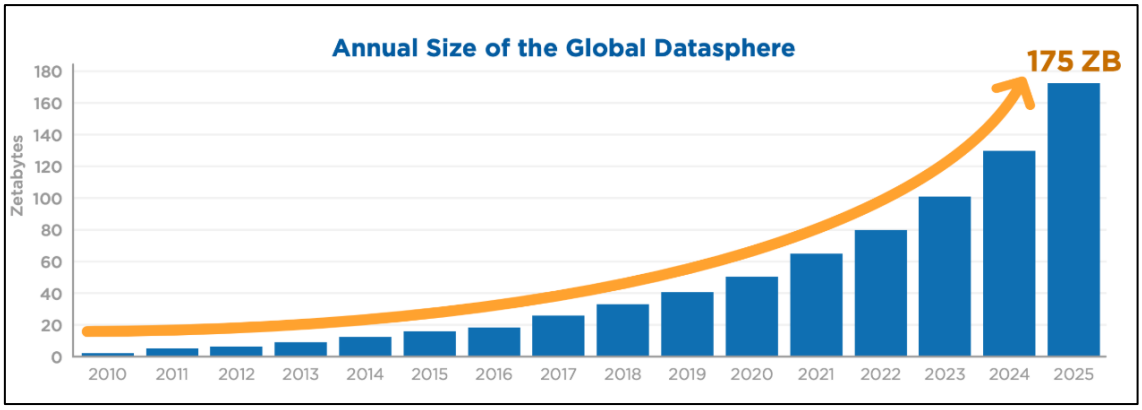


Ilustración 3.7 (Reinsel, Gantz, & Rydning, 2018)

Aquella cantidad de datos titánica que se puede apreciar en la Ilustración 3.7 que nos muestra el reto al cuál los ingenieros y científicos de datos deberán enfrentarse en los próximos 5 años.

El trabajo de la minería de datos es hacerle frente a este monstruo debido a que, si bien es cierto que mucho conocimiento se puede encontrar dentro de **175 ZB**, también hay muchos datos incompletos, inconsistencias e incluso falsos. Estas se presentan como grandes piedras en el zapato de los mineros de datos, los cuales por lo general son los científicos de datos de cualquier organización.

Para que el trabajo de un científico de datos sea considerado cómo minería de datos autentica, los resultados que obtenga deben representar un verdadero descubrimiento (encontrar un patrón o varios patrones de relaciones ocultos) al momento de introducirse en un repositorio de datos de tamaño considerable. Esto descarta de manera inmediata tareas que impliquen recuperar, asociar o buscar registros individuales dentro de una base de datos estructurada, debido a que las relaciones o patrones ya están construidos.

Con el objetivo de conocer que es aquello que permite a la minería de datos el seguir creciendo y evolucionando para ya no solo ofrecer soluciones negocios sino incluso soluciones de vida, debemos conocer sus desafíos motivadores:

- **Escalabilidad**
  Los años pasan y los datos crecen en escalas que superan al terabyte, por lo que hablar de petabytes o zetabytes se volverá común en un futuro. Los algoritmos de minería de datos deben ser escalables para lidiar con este tipo de problemas en el momento en el que se presenten.

Una solución aplicada a la escalabilidad debe ser el experimentar con nuevas estructuras de datos que permitan obtener registros únicos de una manera eficiente.

- **Alta Dimensionalidad**
  En la actualidad es común toparse con grandes conjuntos de datos que almacenan dentro de sí miles de atributos. Esto se debe a que en la actualidad almacenamos datos de genes en bioinformática, datos temporales o espaciales que tienden a presentar alta dimensionalidad.
  Lo que tendemos a entender o llamar dimensiones son la cantidad de características o atributos que son almacenados al momento de realizar una medición, experimento, etc.
  Este punto es importante debido a que saber relacionar las distintas dimensiones de un conjunto de datos puede aportar distintas perspectivas de un mismo problema y como este puede resolverse.

- **Datos Heterogéneos y Complejos**
  Los días en los que los datos de un mismo tipo se almacenaban juntos o en un mismo tipo de archivo se han terminado, por lo que el rol de la minería de datos dentro del mundo de las ciencias y negocios es encontrar técnicas que nos permitan relacionar a la gran cantidad de datos heterogéneos que existen dentro de nuestro mundo.
  Grandes cantidades de datos heterogéneos pueden ser encontrados en páginas web con texto semiestructurado y un par de hipervínculos al igual que en cadenas de ADN las cuales poseen estructuras secuenciales o en tres dimensiones.
  Las relaciones de los datos heterogéneos pueden incluir relaciones entre tiempo, espacio, complejidad de grafos, herencia, etc.

- **Propiedad y distribución de datos**

  Los datos no se encuentran dónde deben estar y tampoco pertenecen a quien le corresponde. Este problema nos muestra que los datos no siempre estarán centralizados o pertenecerán a una misma organización, estos se encontrarán distribuidos por todo el globo y le pertenecerán a mil y una organizaciones. El reto de la minería de datos es proponer una solución desarrollando algoritmos de minería distribuida que puedan lidiar con esto. Las tres soluciones propuestas son:

  1. **"¿Cómo reducir la cantidad de comunicación necesitada para realizar computación distribuida?**
  2. **¿Cómo consolidar de forma efectiva los resultados obtenidos a partir de diversas fuentes utilizando la minería de datos?**
  3. **¿Cómo tratar los problemas con la seguridad de los datos?"**

  (Tan, Steinbach, & Kumar, 2014)

- **Análisis no tradicional**
  Tradicionalmente un problema se analiza utilizando el método científico el cual aborda a un problema desde la hipótesis-tesis-resultado. Este proceso funciona siempre y cuando la cantidad de datos es capaz de manejarse con facilidad, pero en tiempos actuales el hecho de recoger

grandes cantidades de datos para su posterior evaluación es un proceso exhaustivo y bastante caro en algunos casos. Es en aquellos que la minería de datos busca automatizar el proceso de generación y evaluación de hipótesis para acelerar el proceso del análisis.
Otro problema que requiere el análisis no tradicional es lidiar con datos sesgados, erróneos, no convencionales y que presentan distribuciones nada convencionales.

3.2.4 Minería de Texto

La minería de texto es conocida como el proceso de obtener información valiosa a partir de los textos en **Lenguaje Natural.** El procesamiento natural del lenguaje junto con la minería de texto busca convertir a los textos en datos que puedan ser utilizados para el análisis de sentimientos, emociones, polaridad, etc.

Una gran parte de la información que posee el ser humano se encuentra escrita en forma de textos. A través de los textos podemos aprender y entender que piensan, sienten o conocen otras personas. Los negocios, empresas y grandes conglomerados piensan que se puede extraer información valiosa a partir de los textos que la gente es capaz de producir debido a que estos comunican aquello que nos gusta, nos desagrada o como se encuentra hasta nuestra salud y humor. El problema es que extraer conclusiones e información a partir de toda esta inmensa cantidad de texto es casi imposible para un solo ser humano y es por ello por lo que utilizamos a las computadoras para asistirnos en esta tarea tan laboriosa.

Extraer un significado potencial a los datos que recibimos e incluso filtrar aquello que es relevante de lo que no lo es sigue siendo hasta ahora algo netamente humano y en lo que las máquinas aún no son capaces de alcanzarnos hasta este momento. Cuando realizamos minería de texto debemos tener claro el horizonte o panorama donde queremos actuar y cuáles son los resultados que esperamos obtener debido a que estos son aquellos componentes humanos que una máquina no podrá tomar en cuenta por si sola.

Cuando la minería de texto es utilizada para analizar textos y convertirlos en alguna forma más estructurada, lo que esperamos es derivar o crear conclusiones a partir de ellos. Uno de los problemas de la minería de texto dentro del contexto de encontrar conclusiones aparece cuando nos encontramos con lenguajes no naturales cómo pueden ser los logs de registro de una computadora, las matemáticas, el código morse, el esperanto, etc. Estos lenguajes no caen dentro del campo del lenguaje natural debido a que su evolución no se debió a factores naturales, sino que más bien estos fueron creados deliberadamente por los seres humanos. Con excepción del esperanto el cual contiene características de la comunicación natural (habla y gramática).

3.3 Stemming y el Algoritmo de Porter

Stemming (derivado, proveniente) es el proceso en el cual se busca reducir la inflexión de las palabras a su “raíz”. Esto evita tener demasiada variación en los datos recolectados. Los algoritmos de **stemming** se han venido desarrollando desde los años 60 en el mundo de la computación.

Está técnica de normalización del texto en el campo del procesamiento del lenguaje natural sirve para preparar texto, palabras y documentos para un procesamiento más exhaustivo. Su uso se ha extendido a sistemas de etiquetado, indexación, motores SEO, etc.

3.3.1 El Algoritmo de Porter

El algoritmo de Porter fue desarrollado en el año de 1979 y busca quitar los sufijos presentes en las palabras. El objetivo es que todas las palabras que pertenezcan o se deriven de una misma raíz puedan ser identificadas.

El algoritmo de Porter cuenta con 5 reglas las cuales se deben ejecutar paso a paso para lograr su cometido. Las reglas del algoritmo de Porter son las siguientes:

1. Tomar un documento como entrada.
2. Leer el documento línea a línea.
3. Tokenizar (cortar el texto en pedazos) la línea.
4. Derivar las palabras.
5. Devolver las palabras derivadas.

El algoritmo de Porter es conocido por su simplicidad, velocidad y su uso en ambientes de recuperación de información IR (Information Retrival en inglés) por la rápida recuperación y obtención de consultas de búsqueda.

Un ejemplo de cómo funciona el algoritmo de Porter se muestra en la Ilustración 3.8:

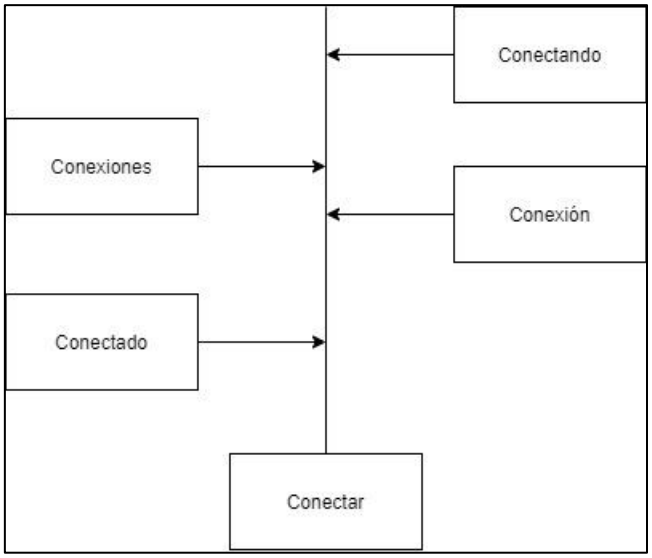


Ilustración 3.8 (Moreno, 2020)

El algoritmo de Porter es bastante bueno, pero no es perfecto y uno de los problemas que posee es que algunas veces puede generar derivaciones que no necesariamente son palabras coherentes o que pertenezcan al lenguaje con el que se está trabajando.

3.3.2 Lematización

Otra aproximación de PNL y de la minería de texto es la Lematización. Al igual que el algoritmo de Porter esta busca reducir las palabras a sus raíces, asegurándose que la raíz sea correcta o pertenezca al lenguaje que se está analizando.

Dentro de la lematización una palabra es llamada lema. Un lema es la forma canónica, de diccionario o en la que se encuentra citada una palabra dentro de un conjunto de palabras. Para conseguir este resultado se toma en consideración el análisis morfológico de las palabras, el cual se logra utilizando "diccionarios detallados" en los cuales un algoritmo encontrará la forma de asociar a las palabras con sus respectivas raíces (lema).

La lematización funciona de la siguiente manera:

- Agrupa a todas las variaciones de una palabra.
- Al igual que Porter se encarga de asociar varias palabras en una misma raíz.
- La "salida" o resultado de utilizar lematización son palabras que propiamente pertenecen al lenguaje humano.

Un ejemplo de cómo se comporta la lematización en inglés en la Ilustración 3.9:

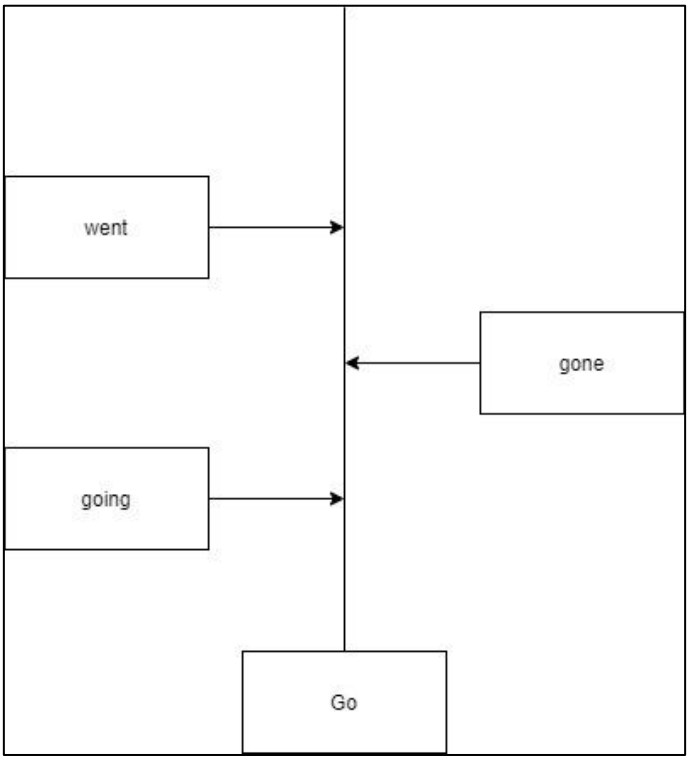


Ilustración 3.9 (Moreno, 2020)

A diferencia del Algoritmo de Porter la lematización produce o deriva palabras que son coherentes o pertenecen al lenguaje humano. En cuestiones de rendimiento el algoritmo de Porter es más rápido que la lematización, pero no siempre puede presentar resultados coherentes. La lematización en términos de rendimiento es mucho más lenta, pero produce mejores resultados.

3.4 Recolección de Datos Usando R

3.4.1 Preparación de herramientas en R Studio para la Etapa de Captura

Esta etapa corresponde a la **captura de datos** dentro del **modelo de análisis de redes sociales** y es seguramente la más importante debido a que es en este momento en el cual nos adentraremos por primera vez al indomable y salvaje océano de datos que posee internet.

Los datos que se recogerán corresponderán a tweets generados por usuarios dentro de la red social Twitter.

La herramienta que se ha escogido para realizar este proceso es el lenguaje estadístico **R** y el **IDE R Studio** debido a que ambos combinados cuentan con los paquetes y herramientas para realizar todo el proceso de análisis de redes sociales. Esto quiere decir que el **IDE R Studio** es capaz de asistir al investigador desde la etapa de captura hasta la etapa de visualización o presentación de resultados.

La recolección de tweets desde Twitter no es una característica que viene integrada por defecto dentro del IDE R Studio por lo que para poder obtenerlos debemos instalar el paquete **TwitteR** el cual nos permitirá obtener tweets desde la red social de manera directa utilizando un **API** proporcionado por la red social. Para poder hacer uso del **API** debemos contar primero con una cuenta común y corriente de Twitter y luego crear una de desarrollador para poder utilizar las herramientas que brinda la red social.

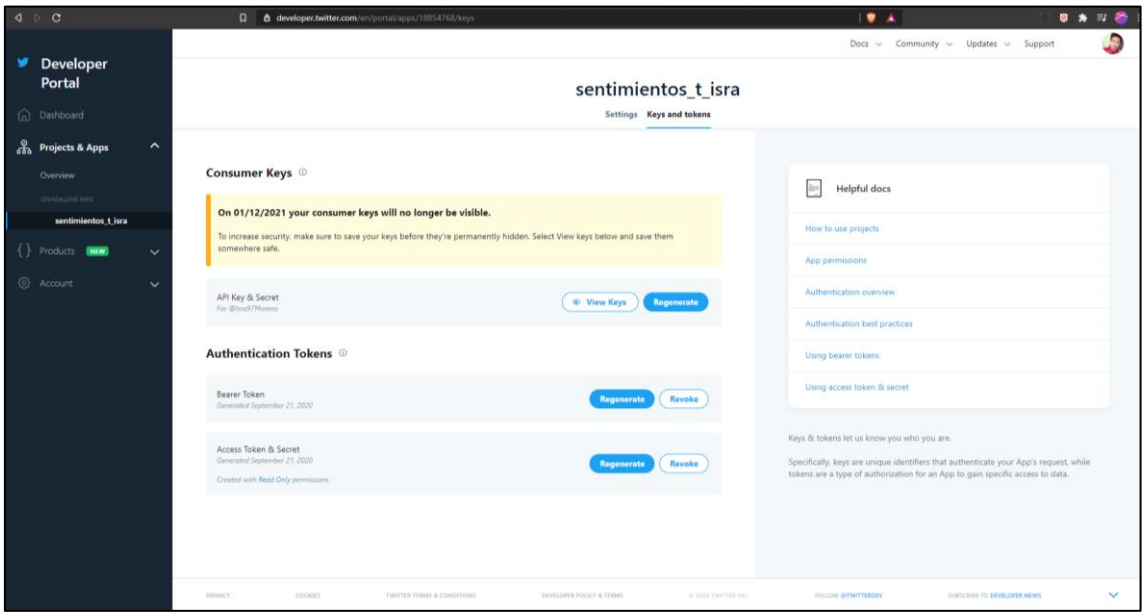


Ilustración 3.10 (Moreno, 2020)

La Ilustración 3.10 indica cómo debería verse el portal para desarrolladores de Twitter una vez obtenido la cuenta de desarrollador.

Una vez que tengamos nuestra cuenta de desarrollador deberemos pasar a crear el nombre de una aplicación para poder hacer uso de las llaves y tokens del API de Twitter para poder comenzar a descargar los tweets que necesitemos.

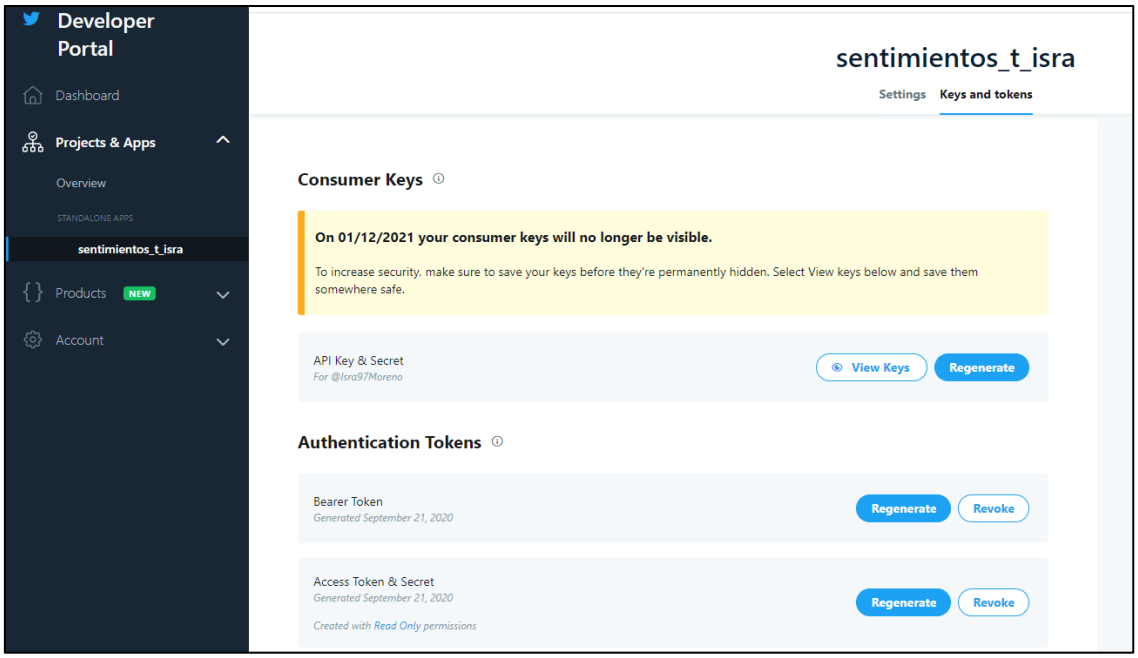


Ilustración 3.11 (Moreno, 2020)

La Ilustración 3.11 muestra que se han generado correctamente las llaves y tokens de acceso para el API de Twitter.

Dentro de R Studio una vez que hayamos realizado el proceso de instalación del paquete TwitteR, deberemos escribir un par de líneas de código para comenzar a recuperar los tweets que deseemos.

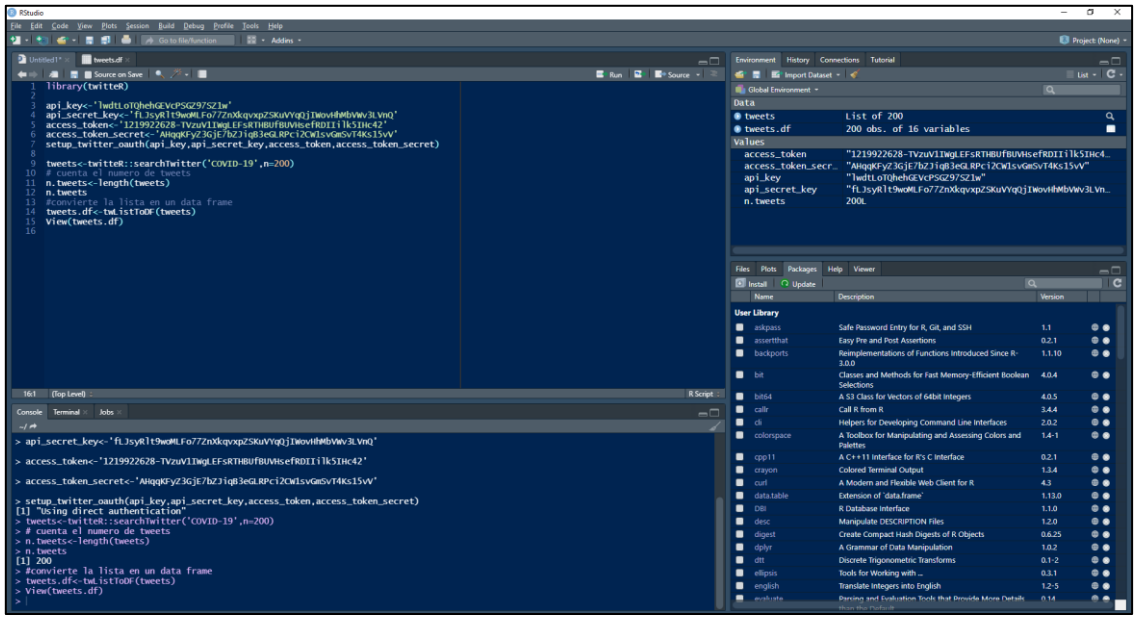


Ilustración 3.12 (Moreno, 2020)

A partir de la Ilustración 3.12 se comenzarán a recuperar todo tipo de tweets dependiendo de los parámetros que se ingresen dentro de la función searchTwitter().

La función searchTwitter() nos permite ingresar varios parámetros como pueden ser:

- Tema o cuenta a buscar.
- Cantidad de tweets a recuperar (tiene un límite de 32000).
- Idioma de los tweets (para configurar el idioma debemos seguir la normativa ISO 639-1).
- Rango de fechas de los tweets (desde-hasta), las fechas deben ser ingresadas en el formato YYYY-MM-DD.
- Por zona geográfica (se debe ingresar la latitud, longitud y radio en millas).
- Por popularidad.
- Tweets en tiempo real.

Todos estos parámetros pueden ser manipulados según la conveniencia del investigador para ampliar o reducir su campo de captura. Los únicos parámetros que son requeridos para funcionar son la cadena de texto que indique el tema o cuenta a buscar.

3.4.2 Preguntas a Nivel Internacional

Como se ha visto en capítulos previos, el análisis de redes sociales el cual se complementa con análisis de sentimientos para la obtención de información valiosa debe partir de una "línea base" en la cual debemos fundamentar nuestra investigación. Está línea base debe estar sustentada por una pregunta en base a la cual queremos trabajar y de la cual esperamos obtener respuestas.

Dentro de esta investigación se ha decidido realizar una pregunta a nivel internacional a la cual se le aplicará su respectivo análisis haciendo uso del lenguaje R y sus distintas herramientas de análisis de sentimientos y de datos.

La pregunta a nivel internacional en cuestión es:

¿Cuál es la opinión del pueblo americano acerca de los candidatos Donald Trump y Joe Biden?

El proceso de análisis de redes sociales nos debería arrojar los datos necesarios para resolver está pregunta.

3.4.3 Preguntas Nivel Nacional

Dentro del panorama nacional se ha decidido realizar una pregunta a la cual se le aplicará su respectivo análisis haciendo uso del lenguaje R y sus distintas herramientas de análisis de sentimientos y de datos.

Las preguntas a nivel nacional en cuestión son:

¿Es realmente Guillermo Lasso un candidato aceptado por el pueblo?

¿Cuál es la opinión de la gente sobre Guillermo Lasso y Yaku Pérez?

¿Es realmente Yaku Pérez un candidato aceptado por el pueblo?

El proceso de análisis de redes sociales nos debería arrojar los datos necesarios para resolver estas preguntas.

3.4.4 Etapa de Captura en inglés

Para empezar con la captura de datos en inglés debemos cargar los siguientes paquetes dentro de R Studio para comenzar a trabajar:

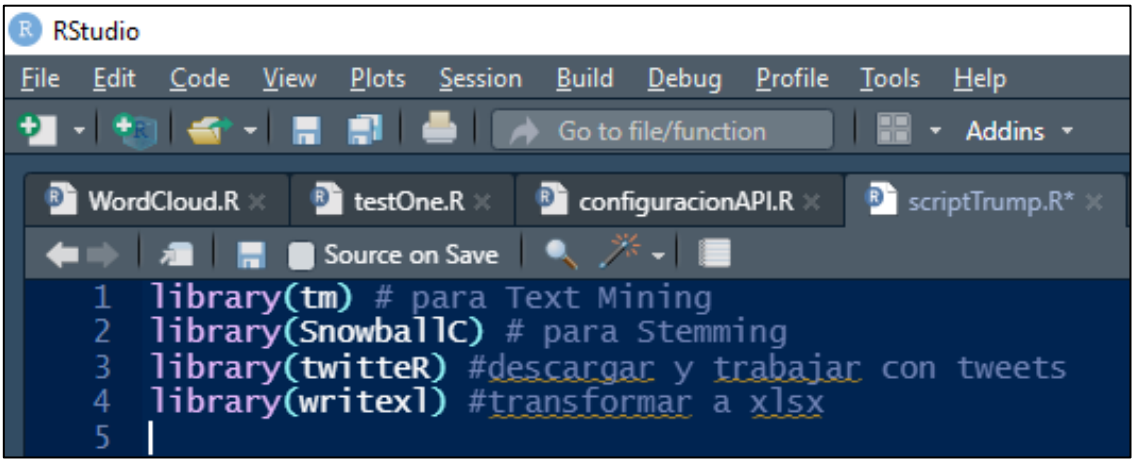


Ilustración 3.13 (Moreno, 2020)

La Ilustración 3.13 muestra las cuatro librerías que servirán para la etapa de captura en el idioma inglés.

Librerías/Paquetes:

El paquete/librería **tm** es un framework (marco de trabajo) para aplicaciones que involucren la minería de texto dentro de R. Actualmente el paquete se encuentra en la versión 0.7-7.

La librería **SnowballC** permite encontrar la raíz de una palabra (stemming en inglés) utilizando el **algoritmo de stemming de Porter**. Esto permite colapsar palabras a una raíz común con el objetivo de ayudar a la comparación de vocabulario. Actualmente se encuentra en la versión 0.7.0 y soporta idiomas cómo el danés, holandés, español, inglés, finés, alemán, francés, etc.

La librería **writexl** permite exportar un data frame desde R a la extensión xlsx sin necesidad de que el data frame sea procesado por otro lenguaje (cómo Java), acción que solía ser necesaria antes.

El proceso de captura empezará cuando se tengan las siguientes líneas de código:

```
13  #Etapa de Recoleccion o Captura
14  # extrayendo tweets relacionados con la cuenta oficial de Donald Trump
15  tweets<-twitteR::searchTwitter('realDonaldTrump',n=2000)
16  #conservamos solo tweets no retweets
17  tweets<-twitteR::strip_retweets(tweets)
18  # cuenta el numero de tweets
19  n.tweets<-length(tweets)
20  n.tweets
21  #convierte la lista en un data frame
22  tweets.df<-twListToDF(tweets)
23  #permite visualizar el data frame
24  View(tweets.df)
```

Ilustración 3.14 (Moreno, 2020)

La Ilustración 3.14 muestra la captura de tweets acerca de Donald Trump y su transformación a un data frame.

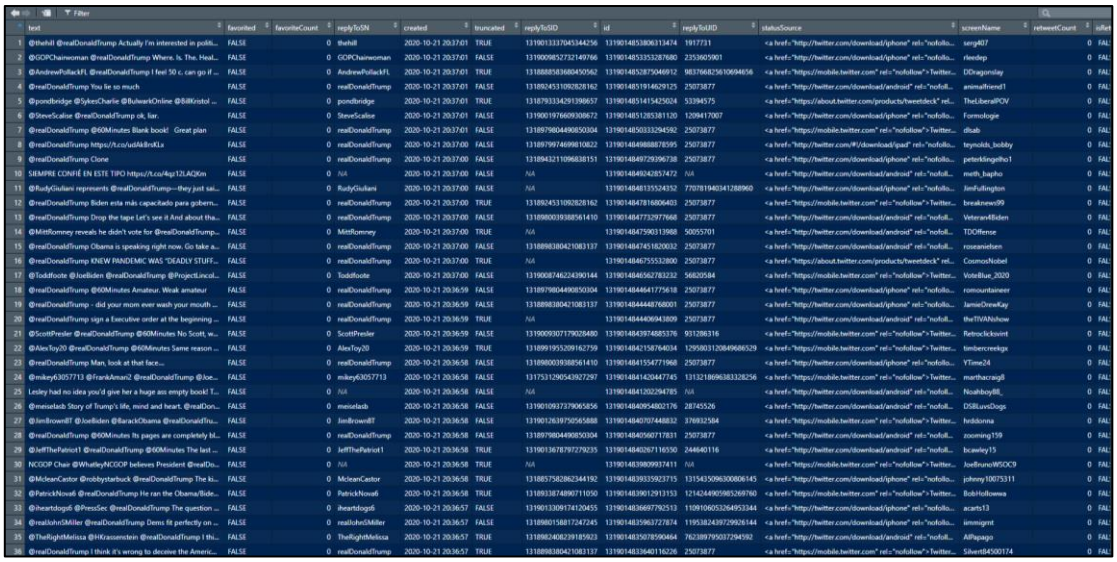
La línea N°15 se encarga de realizar el proceso de captura de los tweets que se encuentren relacionados con **la cuenta @realDonaldTrump**. Los tweets recolectados al momento de haber realizado la ejecución de la línea 15 corresponden a tweets tomados en tiempo real durante el siguiente rango de tiempo desde **2020-10-21 20:34:55** hasta **2020-10-21 20:37:01**. La primera fecha corresponde al último tweet recuperado y la segunda fecha al primer tweet recuperado. La hora en la cual se registran los tweets corresponde al tiempo **GMT (**Hora del Meridiano de Greenwich**) UTC+00.**

La línea 16 se encarga de remover los retweets que existan dentro de la variable tweets. Se remueven los retweets ya que buscamos obtener un set de datos que solo contenga tweets puros.

Las líneas 19 y 20 se encargan de contar la cantidad de tweets recuperados una vez ya removidos los retweets.

La línea 22 se encarga de convertir a la variable tweets en un **data frame.** Un data frame es una estructura de datos bidimensional que permite almacenar datos heterogéneos.

Utilizando la función View de la línea 24 podremos visualizar el data frame creado.



	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	statusSource	screenName	retweetCount	isRet
1	@thehill @realDonaldTrump Actually I'm interested in politi...	FALSE	0	thehill	2020-10-21 20:37:01	TRUE	1319013337045344256	1319014853806313474	1917731	<a href="http://twitter.com/download/iphone" rel="nofollo...	serg407	0	FAL
2	@GOPChairwoman @realDonaldTrump Where, Is The, Heal...	FALSE	0	GOPChairwoman	2020-10-21 20:37:01	FALSE	1319009852732149766	1319014853353287680	2353605901	<a href="http://twitter.com/download/iphone" rel="nofollo...	rleedep	0	FAL
3	@AndrewPollackFL @realDonaldTrump I feel 50 c. can go if ...	FALSE	0	AndrewPollackFL	2020-10-21 20:37:01	TRUE	1318888583680450562	1319014852875046912	983766825610694656	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	DDragonsley	0	FAL
4	@realDonaldTrump You lie so much	FALSE	0	realDonaldTrump	2020-10-21 20:37:01	FALSE	1318924531092828162	1319014851914629125	25073877	<a href="http://twitter.com/download/android" rel="nofoll...	animalfriend1	0	FAL
5	@pondbridge @SykesCharlie @BulwarkOnline @BillKristol ...	FALSE	0	pondbridge	2020-10-21 20:37:01	TRUE	1318793334291398657	1319014851415425024	53394575	<a href="https://about.twitter.com/products/tweetdeck" rel...	TheLiberalPOV	0	FAL
6	@SteveScalise @realDonaldTrump ok, liar.	FALSE	0	SteveScalise	2020-10-21 20:37:01	FALSE	1319001976609308672	1319014851285381120	1209417007	<a href="http://twitter.com/download/iphone" rel="nofollo...	Formologie	0	FAL
7	@realDonaldTrump @60Minutes Blank book! Great plan	FALSE	0	realDonaldTrump	2020-10-21 20:37:01	FALSE	1318979804490850304	1319014850333294592	25073877	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	dlsab	0	FAL
8	@realDonaldTrump https://t.co/udA8lnKLx	FALSE	0	realDonaldTrump	2020-10-21 20:37:00	FALSE	1318979974699810822	1319014849888878595	25073877	<a href="http://twitter.com/#!/download/ipad" rel="nofollo...	teynoldi_bobby	0	FAL
9	@realDonaldTrump Clone	FALSE	0	realDonaldTrump	2020-10-21 20:37:00	FALSE	1318943211096838151	1319014849729396738	25073877	<a href="http://twitter.com/download/iphone" rel="nofollo...	peterkdingelho1	0	FAL
10	SIEMPRE CONFÍE EN ESTE TIPO https://t.co/4qz12LAQKm	FALSE	0	NA	2020-10-21 20:37:00	FALSE	NA	1319014849242857472	NA	<a href="http://twitter.com/download/android" rel="nofoll...	meth_bapho	0	FAL
11	@RudyGiuliani represents @realDonaldTrump—they just sai...	FALSE	0	RudyGiuliani	2020-10-21 20:37:00	FALSE	NA	1319014848133524352	770781940341288960	<a href="http://twitter.com/download/iphone" rel="nofollo...	JimFullington	0	FAL
12	@realDonaldTrump Biden esta más capacitado para gobern...	FALSE	0	realDonaldTrump	2020-10-21 20:37:00	TRUE	1318924531092828162	1319014847816806403	25073877	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	breaknews99	0	FAL
13	@realDonaldTrump Drop the tape Let's see it And about tha...	FALSE	0	realDonaldTrump	2020-10-21 20:37:00	FALSE	1318980031938561410	1319014847732977668	25073877	<a href="http://twitter.com/download/android" rel="nofoll...	VeteranAtEden	0	FAL
14	@MittRomney reveals he didn't vote for @realDonaldTrump...	FALSE	0	MittRomney	2020-10-21 20:37:00	TRUE	NA	1319014847590313988	50055701	<a href="http://twitter.com/download/android" rel="nofoll...	TDOffense	0	FAL
15	@realDonaldTrump Obama is speaking right now. Go take a...	FALSE	0	realDonaldTrump	2020-10-21 20:37:00	FALSE	1318898380421083137	1319014847451620032	25073877	<a href="http://twitter.com/download/android" rel="nofoll...	roseanielsen	0	FAL
16	@realDonaldTrump KNEW PANDEMIC WAS "DEADLY STUFF...	FALSE	0	realDonaldTrump	2020-10-21 20:37:00	FALSE	NA	1319014846755532800	25073877	<a href="https://about.twitter.com/products/tweetdeck" rel...	CosmosNobel	0	FAL
17	@Toddfoote @JoeBiden @realDonaldTrump @ProjectLincol...	FALSE	0	Toddfoote	2020-10-21 20:37:00	FALSE	1319008746224390144	1319014846562783232	56820584	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	VoteBlue_2020	0	FAL
18	@realDonaldTrump @60Minutes Amateur. Weak amateur	FALSE	0	realDonaldTrump	2020-10-21 20:36:59	FALSE	1318979804490850304	1319014844641775618	25073877	<a href="http://twitter.com/download/iphone" rel="nofollo...	romountaineer	0	FAL
19	@realDonaldTrump - did your mom ever wash your mouth ...	FALSE	0	realDonaldTrump	2020-10-21 20:36:59	FALSE	1318898380421083137	1319014844487768001	25073877	<a href="http://twitter.com/download/iphone" rel="nofollo...	JamieOreeKay	0	FAL
20	@realDonaldTrump sign a Executive order at the beginning ...	FALSE	0	realDonaldTrump	2020-10-21 20:36:59	TRUE	NA	1319014844069343809	25073877	<a href="http://twitter.com/download/android" rel="nofoll...	theTIVANshow	0	FAL
21	@ScottPresler @realDonaldTrump @60Minutes No Scott, w...	FALSE	0	ScottPresler	2020-10-21 20:36:59	FALSE	1319009307179028480	1319014843974885376	931286316	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	Retrockicksvint	0	FAL
22	@AlexToy20 @realDonaldTrump @60Minutes Same reason ...	FALSE	0	AlexToy20	2020-10-21 20:36:59	TRUE	1318991955209162759	1319014842158764034	1295803120849686529	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	timbercreekgx	0	FAL
23	@realDonaldTrump Man, look at that face...	FALSE	0	realDonaldTrump	2020-10-21 20:36:58	FALSE	1318980031938561410	1319014841554771968	25073877	<a href="http://twitter.com/download/iphone" rel="nofollo...	YTime24	0	FAL
24	@mikey63057713 @FrankAmani2 @realDonaldTrump @Joe...	FALSE	0	mikey63057713	2020-10-21 20:36:58	FALSE	1317531290543927297	1319014841420447745	1313218696388328256	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	marthacraig8	0	FAL
25	Lesley had no idea you'd give her a huge ass empty book! T...	FALSE	0	NA	2020-10-21 20:36:58	FALSE	NA	1319014841202094785	NA	<a href="http://twitter.com/download/android" rel="nofoll...	Noahboy88_	0	FAL
26	@meiselasb Story of Trump's life, mind and heart. @realDon...	FALSE	0	meiselasb	2020-10-21 20:36:58	FALSE	1319010937379065856	1319014840954802176	28745526	<a href="http://twitter.com/download/android" rel="nofoll...	DSBLuvsDogs	0	FAL
27	@JimBrown87 @JoeBiden @BarackObama @realDonaldTru...	FALSE	0	JimBrown87	2020-10-21 20:36:58	FALSE	1319012639750565888	1319014840707448832	376932584	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	hrddonna	0	FAL
28	@realDonaldTrump @60Minutes Its pages are completely bl...	FALSE	0	realDonaldTrump	2020-10-21 20:36:58	FALSE	1318979804490850304	1319014840560717831	25073877	<a href="http://twitter.com/download/android" rel="nofoll...	zooming159	0	FAL
29	@JeffThePatriot1 @realDonaldTrump @60Minutes The last ...	FALSE	0	JeffThePatriot1	2020-10-21 20:36:58	TRUE	1319013678797279235	1319014840267116550	244640116	<a href="http://twitter.com/download/android" rel="nofoll...	bcawley15	0	FAL
30	NCGOP Chair @WhatleyNCGOP believes President @realDo...	FALSE	0	NA	2020-10-21 20:36:58	TRUE	NA	1319014839809937411	NA	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	JoeBrunoWSOC9	0	FAL
31	@McleanCastor @robbystarbuck @realDonaldTrump The ki...	FALSE	0	McleanCastor	2020-10-21 20:36:58	TRUE	1318857582862344192	1319014839335923715	1315435096300806145	<a href="http://twitter.com/download/android" rel="nofoll...	johnny10075311	0	FAL
32	@PatrickNova6 @realDonaldTrump He ran the Obama/Bide...	FALSE	0	PatrickNova6	2020-10-21 20:36:58	TRUE	1318933874890711050	1319014839012913153	1214244905985269760	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	BobHollowwa	0	FAL
33	@iheartdogs6 @PressSec @realDonaldTrump The question ...	FALSE	0	iheartdogs6	2020-10-21 20:36:57	FALSE	1319013309174120455	1319014838669792513	1109106053264953344	<a href="http://twitter.com/download/iphone" rel="nofollo...	acarts13	0	FAL
34	@realJohnSMiller @realDonaldTrump Dems fit perfectly on ...	FALSE	0	realJohnSMiller	2020-10-21 20:36:57	FALSE	1318980158817247245	1319014835963727874	1195382439729926144	<a href="http://twitter.com/download/iphone" rel="nofollo...	iimmgmt	0	FAL
35	@TheRightMelissa @HKrassenstein @realDonaldTrump I thi...	FALSE	0	TheRightMelissa	2020-10-21 20:36:57	TRUE	1318982406239185923	1319014835078590464	762389795037294592	<a href="http://twitter.com/download/android" rel="nofoll...	AlPapago	0	FAL
36	@realDonaldTrump I think it's wrong to deceive the Americ...	FALSE	0	realDonaldTrump	2020-10-21 20:36:57	TRUE	1318898380421083137	1319014833640116226	25073877	<a href="https://mobile.twitter.com" rel="nofollow">Twitter...	Silvert84500174	0	FAL

Ilustración 3.15 (Moreno, 2020)

La ilustración 3.15 se encarga de mostrar todos los tweets recuperados acerca de Donald Trump junto con otros parámetros.

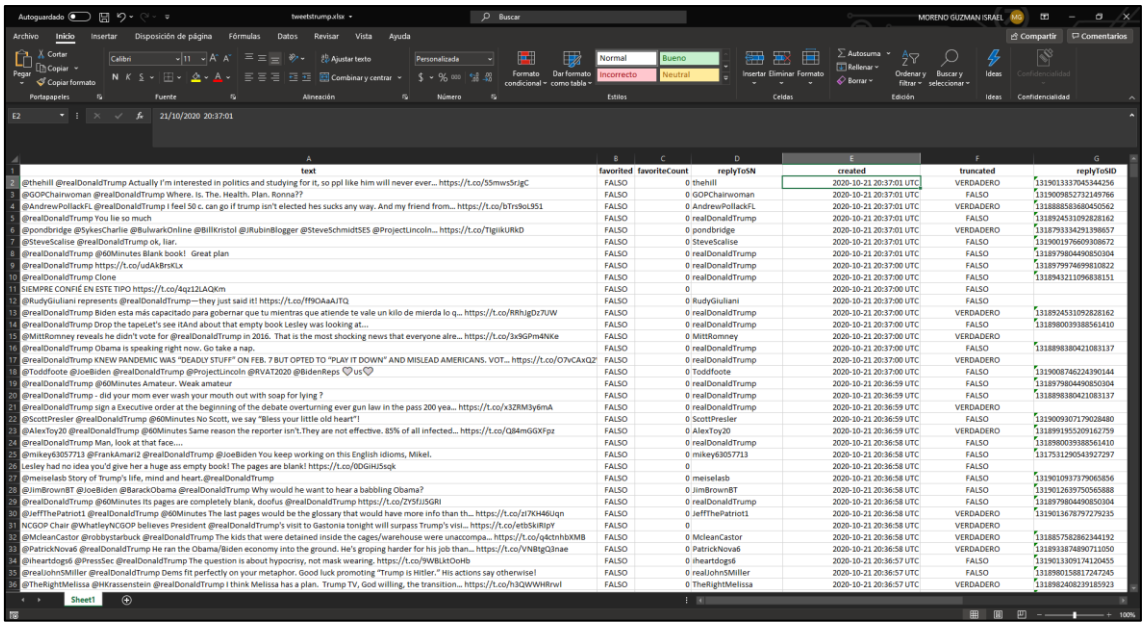
Dentro de todos estos datos nos centraremos en la **columna text** debido a que en esta se almacenan los tweets recuperados.

Para evitar perder la información recuperada durante la ejecución del método searchTwitter guardaremos el **data frame** como un archivo xlsx tal cual indica la línea 27 de la Ilustración 3.16.

```
13  #Etapa de Recoleccion o Captura
14  # extrayendo tweets relacionados con la cuenta oficial de Donald Trump
15  tweets<-twitteR::searchTwitter('realDonaldTrump',n=2000)
16  #conservamos solo tweets no retweets
17  tweets<-twitteR::strip_retweets(tweets)
18  # cuenta el numero de tweets
19  n.tweets<-length(tweets)
20  n.tweets
21  #convierte la lista en un data frame
22  tweets.df<-twListToDF(tweets)
23  #permite visualizar el data frame
24  View(tweets.df)
25  #Guardamos el data frame como archivo en formato xls
26  # convierte en xls
27  writexl::write_xlsx(tweets.df,"C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetstrump.xlsx")
```

Ilustración 3.16 (Moreno, 2020)

Una vez procesada la línea 27 podremos explorar el data frame desde la comodidad de Excel en caso de ser necesario.



	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID
2	@thehill @realDonaldTrump Actually I'm interested in politics and studying for it, so ppl like him will never ever... https://t.co/55mwsSrjgC	FALSO	0	thehill	2020-10-21 20:37:01 UTC	VERDADERO	1319013337045344256
3	@GOPChairwoman @realDonaldTrump Where. Is. The. Health. Plan. Ronna??	FALSO	0	GOPChairwoman	2020-10-21 20:37:01 UTC	FALSO	1319009852732149766
4	@AndrewPollackFL @realDonaldTrump I feel 50 c. can go if trump isn't elected hes sucks any way. And my friend from... https://t.co/bTrs9oL951	FALSO	0	AndrewPollackFL	2020-10-21 20:37:01 UTC	VERDADERO	1318888583680450562
5	@realDonaldTrump You lie so much	FALSO	0	realDonaldTrump	2020-10-21 20:37:01 UTC	FALSO	1318924531092828162
6	@pondbridge @SykesCharlie @BulwarkOnline @BillKristol @JRubinBlogger @SteveSchmidtSES @ProjectLincoln... https://t.co/TIgiiiKURkD	FALSO	0	pondbridge	2020-10-21 20:37:01 UTC	VERDADERO	1318793334291398657
7	@SteveScalise @realDonaldTrump ok, liar.	FALSO	0	SteveScalise	2020-10-21 20:37:01 UTC	FALSO	1319001976609308672
8	@realDonaldTrump @60Minutes Blank book! Great plan	FALSO	0	realDonaldTrump	2020-10-21 20:37:01 UTC	FALSO	1318979804490850304
9	@realDonaldTrump https://t.co/udAkBrsKLx	FALSO	0	realDonaldTrump	2020-10-21 20:37:00 UTC	FALSO	1318979974699810822
10	@realDonaldTrump Clone	FALSO	0	realDonaldTrump	2020-10-21 20:37:00 UTC	FALSO	1318943211096838151
11	SIEMPRE CONFÍE EN ESTE TIPO https://t.co/4qz12LAQKm	FALSO	0		2020-10-21 20:37:00 UTC	FALSO	
12	@RudyGiuliani represents @realDonaldTrump—they just said it! https://t.co/ff9OAaAJTQ	FALSO	0	RudyGiuliani	2020-10-21 20:37:00 UTC	FALSO	
13	@realDonaldTrump Biden esta más capacitado para gobernar que tu mientras que atiende te vale un kilo de mierda lo q... https://t.co/RRhJgDz7UW	FALSO	0	realDonaldTrump	2020-10-21 20:37:00 UTC	VERDADERO	1318924531092828162
14	@realDonaldTrump Drop the tapeLet's see itAnd about that empty book Lesley was looking at...	FALSO	0	realDonaldTrump	2020-10-21 20:37:00 UTC	FALSO	1318980039388561410
15	@MittRomney reveals he didn't vote for @realDonaldTrump in 2016. That is the most shocking news that everyone alre... https://t.co/3x9GPm4NKe	FALSO	0	MittRomney	2020-10-21 20:37:00 UTC	VERDADERO	
16	@realDonaldTrump Obama is speaking right now. Go take a nap.	FALSO	0	realDonaldTrump	2020-10-21 20:37:00 UTC	FALSO	1318898380421083137
17	@realDonaldTrump KNEW PANDEMIC WAS "DEADLY STUFF" ON FEB. 7 BUT OPTED TO "PLAY IT DOWN" AND MISLEAD AMERICANS. VOT... https://t.co/O7vCAxQ2'	FALSO	0	realDonaldTrump	2020-10-21 20:37:00 UTC	VERDADERO	
18	@Toddfoote @JoeBiden @realDonaldTrump @ProjectLincoln @RVAT2020 @BidenReps 🤍us🤍	FALSO	0	Toddfoote	2020-10-21 20:37:00 UTC	FALSO	1319008746224390144
19	@realDonaldTrump @60Minutes Amateur. Weak amateur	FALSO	0	realDonaldTrump	2020-10-21 20:36:59 UTC	FALSO	1318979804490850304
20	@realDonaldTrump - did your mom ever wash your mouth out with soap for lying ?	FALSO	0	realDonaldTrump	2020-10-21 20:36:59 UTC	FALSO	1318898380421083137
21	@realDonaldTrump sign a Executive order at the beginning of the debate overturning ever gun law in the pass 200 yea... https://t.co/x3ZRM3y6mA	FALSO	0	realDonaldTrump	2020-10-21 20:36:59 UTC	VERDADERO	
22	@ScottPresler @realDonaldTrump @60Minutes No Scott, we say "Bless your little old heart"!	FALSO	0	ScottPresler	2020-10-21 20:36:59 UTC	FALSO	1319009307179028480
23	@AlexToy20 @realDonaldTrump @60Minutes Same reason the reporter isn't.They are not effective. 85% of all infected... https://t.co/QB4mGGXFpz	FALSO	0	AlexToy20	2020-10-21 20:36:59 UTC	VERDADERO	1318991955209162759
24	@realDonaldTrump Man, look at that face....	FALSO	0	realDonaldTrump	2020-10-21 20:36:58 UTC	FALSO	1318980039388561410
25	@mikey63057713 @FrankAmari2 @realDonaldTrump @JoeBiden You keep working on this English idioms, Mikel.	FALSO	0	mikey63057713	2020-10-21 20:36:58 UTC	FALSO	1317531290543927297
26	Lesley had no idea you'd give her a huge ass empty book! The pages are blank! https://t.co/0DGiHJi5sqk	FALSO	0		2020-10-21 20:36:58 UTC	FALSO	
27	@meiselasb Story of Trump's life, mind and heart.@realDonaldTrump	FALSO	0	meiselasb	2020-10-21 20:36:58 UTC	FALSO	1319010937379065856
28	@JimBrownBT @JoeBiden @BarackObama @realDonaldTrump Why would he want to hear a babbling Obama?	FALSO	0	JimBrownBT	2020-10-21 20:36:58 UTC	FALSO	1319012639750565888
29	@realDonaldTrump @60Minutes Its pages are completely blank, doofus @realDonaldTrump https://t.co/ZY5fzJ5GRI	FALSO	0	realDonaldTrump	2020-10-21 20:36:58 UTC	FALSO	1318979804490850304
30	@JeffThePatriot1 @realDonaldTrump @60Minutes The last pages would be the glossary that would have more info than th... https://t.co/zI7KH46Uqn	FALSO	0	JeffThePatriot1	2020-10-21 20:36:58 UTC	VERDADERO	1319013678759279235
31	NCGOP Chair @WhatleyNCGOP believes President @realDonaldTrump's visit to Gastonia tonight will surpass Trump's visi... https://t.co/etb5kiRIpY	FALSO	0		2020-10-21 20:36:58 UTC	VERDADERO	
32	@McleanCastor @robbystarbuck @realDonaldTrump The kids that were detained inside the cages/warehouse were unaccompa... https://t.co/q4ctnhbXMB	FALSO	0	McleanCastor	2020-10-21 20:36:58 UTC	VERDADERO	1318857582862344192
33	@PatrickNova6 @realDonaldTrump He ran the Obama/Biden economy into the ground. He's groping harder for his job than... https://t.co/VNBtgQ3nae	FALSO	0	PatrickNova6	2020-10-21 20:36:58 UTC	VERDADERO	1318933874890711050
34	@iheartdogs6 @PressSec @realDonaldTrump The question is about hypocrisy, not mask wearing. https://t.co/9WBLktOoHb	FALSO	0	iheartdogs6	2020-10-21 20:36:57 UTC	FALSO	1319013309174120455
35	@realJohnSMiller @realDonaldTrump Dems fit perfectly on your metaphor. Good luck promoting "Trump is Hitler." His actions say otherwise!	FALSO	0	realJohnSMiller	2020-10-21 20:36:57 UTC	FALSO	1318980158817247245
36	@TheRightMelissa @HKrassenstein @realDonaldTrump I think Melissa has a plan. Trump TV, God willing, the transition... https://t.co/h3QiWWHRrwl	FALSO	0	TheRightMelissa	2020-10-21 20:36:57 UTC	VERDADERO	1318982408239185923

Ilustración 3.17 (Moreno, 2020)

La Ilustración 3.17 se encarga de mostrar al data frame de Donald Trump en forma de un archivo xlsx.

El mismo proceso que se ha realizado para la captura de tweets correspondientes a la cuenta @**realDonaldTrump** se realizará para la cuenta del candidato opositor a Trump **el demócrata** @**JoeBiden**.

```
#Etapa de Recoleccion o Captura
# extrayendo tweets relacionados con la cuenta oficial de Joe Biden
tweets<-twitteR::searchTwitter('JoeBiden',n=2000)
#conservamos solo tweets no retweets
tweets<-twitteR::strip_retweets(tweets)
# cuenta el numero de tweets
n.tweets<-length(tweets)
n.tweets
#convierte la lista en un data frame
tweets.df<-twListToDF(tweets)
#permite visualizar el data frame
View(tweets.df)
```

Ilustración 3.18 (Moreno, 2020)

La Ilustración 3.18 muestra la captura de tweets acerca de Joe Biden y su posterior transformación a un data frame.

Al revisar el data frame que contiene los tweets relacionados con Joe Biden podemos observar que contiene 888 tweets recuperados.

	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID
1	@9NEWSNANCY @JoeBiden Nice editing $400,000/ year y...	FALSE	0	9NEWSNANCY	2020-10-21 21:34:20	TRUE	1312130926277619712	1319029275664801792	45451379
2	@Fantine21 @browardpolitics @JoeBiden @realDonaldTru...	FALSE	0	Fantine21	2020-10-21 21:34:19	TRUE	1318887910414962688	1319029272460275714	5682572(
3	@djguillyd @Ramiii_B @MarkDice @JoeBiden 😂	FALSE	0	djguillyd	2020-10-21 21:34:18	FALSE	1319028408857153536	1319029269050306560	2149479:
4	@JoeBiden Absolutely—trump is committing genocide on a...	FALSE	0	JoeBiden	2020-10-21 21:34:18	FALSE	1319021893446492162	1319029268681285638	939091
5	@GMHikaru @BarackObama @AOC @JoeBiden @actblue 6...	FALSE	0	GMHikaru	2020-10-21 21:34:18	FALSE	1318722767458430977	1319029265803833344	6432454:
6	@JoeBiden should go. Turn Texas Blue. #TexasVotes https:...	FALSE	0	JoeBiden	2020-10-21 21:34:17	FALSE	NA	1319029265266999304	939091
7	While polls claim @JoeBiden has a lead over #Trump in Flori...	FALSE	0	NA	2020-10-21 21:34:17	TRUE	NA	1319029263996227586	NA
8	@SteveWCarlson @GlennKesslerWP @AndrewBatesNC @P...	FALSE	0	SteveWCarlson	2020-10-21 21:34:17	TRUE	1318971480475041792	1319029263828385795	1743695!
9	@SenRonJohnson @JoeBiden It's a scandal you're still a US ...	FALSE	0	SenRonJohnson	2020-10-21 21:34:17	FALSE	1319008566116745217	1319029263748706305	2337378!
10	I couldn't agree more! https://t.co/yOT1Wgz3BD	FALSE	0	NA	2020-10-21 21:34:16	FALSE	NA	1319029260904873985	NA
11	@FlyoverSagacity @JoeBiden Regarding the economy, the ...	FALSE	0	JustAManic2020	2020-10-21 21:34:16	TRUE	1319028611328937984	1319029258765860865	1267276(
12	@JoeBiden #can'tGoSoonEnough	FALSE	0	JoeBiden	2020-10-21 21:34:16	FALSE	1319021893446492162	1319029257675415552	939091
13	@SenRonJohnson @maxseddon @JoeBiden Won't be long ...	FALSE	0	SenRonJohnson	2020-10-21 21:34:14	TRUE	1318991608612851714	1319029251799134208	2337378!
14	way past time!! https://t.co/vzfnwjKPOq	FALSE	0	NA	2020-10-21 21:34:14	FALSE	NA	1319029250473639936	NA
15	@rexthetvterrier @JoeBiden Oh hope so for everyone's sake...	FALSE	0	rexthetvterrier	2020-10-21 21:34:13	FALSE	1318939098871140354	1319029248104038400	2384269!
16	@SteveHofstetter @JoeBiden Remember to stay hydrated, k...	FALSE	0	SteveHofstetter	2020-10-21 21:34:13	FALSE	1319027989506453504	1319029247743328257	1597824(
17	@BarnettforAZ @JoeBiden @BarackObama Someone who ...	FALSE	0	BarnettforAZ	2020-10-21 21:34:13	FALSE	1319022181511122945	1319029245943959553	1117898!
18	@Whatdoi36535534 @MagnaCarta121 @Anderology @Joe...	FALSE	0	Whatdoi36535534	2020-10-21 21:34:13	FALSE	1319028438720778240	1319029245830664194	1315151(
19	@JoeBiden Never forget this tweet. https://t.co/51sGTwSNIQ	FALSE	0	JoeBiden	2020-10-21 21:34:13	FALSE	1319021893446492162	1319029245444804608	939091

Showing 1 to 19 of 888 entries, 16 total columns

Ilustración 3.19 (Moreno, 2020)

La Ilustración 3.19 muestra el data frame que contiene los tweets acerca de Joe Biden.

El rango de tiempo en el que fueron recuperados los tweets corresponde al intervalo que va desde **2020-10-21 21:30:27** hasta **2020-10-21 21:34:20.** Este rango de tiempo también corresponde al tiempo **GMT UTC+00**.

Al igual que en el primer caso nos encargamos de guardar el data frame que contiene los tweets en un archivo xlsx.

```
#Etapa de Recoleccion o Captura
# extrayendo tweets relacionados con la cuenta oficial de Joe Biden
tweets<-twitteR::searchTwitter('JoeBiden',n=2000)
#conservamos solo tweets no retweets
tweets<-twitteR::strip_retweets(tweets)
# cuenta el numero de tweets
n.tweets<-length(tweets)
n.tweets
#convierte la lista en un data frame
tweets.df<-twListToDF(tweets)
#permite visualizar el data frame
View(tweets.df)
#Guardamos el data frame como archivo en formato xls
# convierte en xls
writexl::write_xlsx(tweets.df,"C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetsbiden.xlsx")
```

Ilustración 3.20 (Moreno, 2020)

La Ilustración 3.20 muestra cómo se almacena el data frame de Joe Biden en un archivo xlsx para su permanencia.

3.4.5 Etapa de Captura en español

La etapa de captura de tweets en español es prácticamente el mismo proceso que realizamos con anterioridad, la única diferencia es que aquí delimitaremos aún más la captura de los datos dentro la función searchTwitter debido a que buscamos tweets en español y dentro del Ecuador.

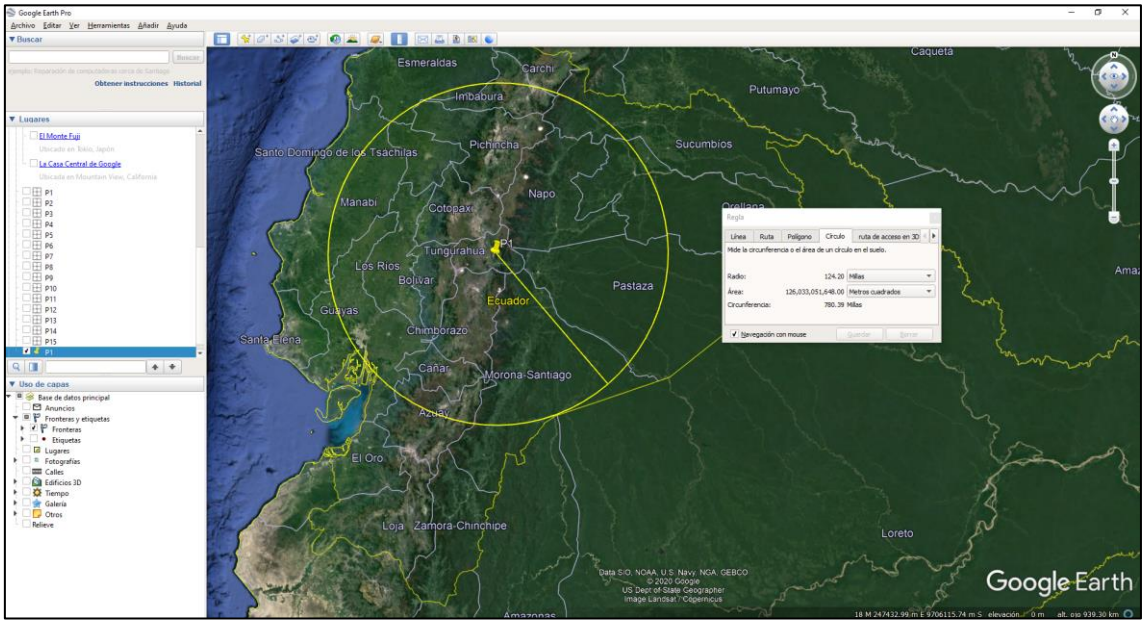


Ilustración 3.21 (Moreno, 2020)

Se ha decidido delimitar la búsqueda de Tweets dentro del Ecuador en el radio que indica la Ilustración 3.21. El radio de búsqueda de tweets estará dentro de las 124.20 Millas terrestres y abarca las siguientes provincias del Ecuador:

- Imbabura
- Pichincha
- Napo
- Sucumbíos
- Orellana
- Pastaza
- Morona-Santiago
- Azuay
- Cañar
- Guayas
- Chimborazo
- Los Ríos
- Bolívar
- Tungurahua
- Cotopaxi
- Manabí
- Santo Domingo de los Tsáchilas
- Esmeraldas
- Carchi

El radio abarca 19 de las 24 provincias del Ecuador, aquellas provincias que han quedado fuera del radio de búsqueda son:

- Loja
- El Oro
- Zamora
- Santa Elena
- Galápagos

Estás 5 provincias han quedado fuera del radio debido a que si se lo aumenta se podrían comenzar a recibir tweets de los vecinos países de Colombia y Perú.

Con las coordenadas del punto **P1** (de donde parte el radio) podemos configurar la función searchTwitter para nuestra captura de tweets.

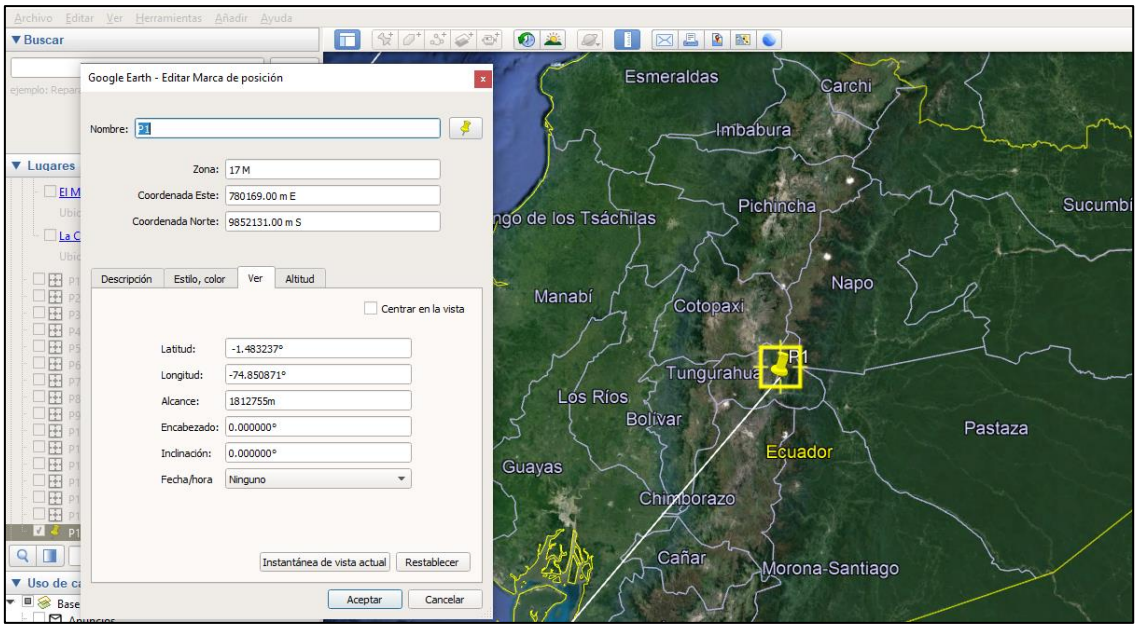


Ilustración 3.22 (Moreno, 2020)

La Ilustración 3.22 se encarga de mostrar el radio que se abarca desde el punto P1 en el mapa de Ecuador utilizando la herramienta Google Earth.

Las líneas de código para la captura de tweets de Guillermo Lasso se muestran en la Ilustración 3.23:

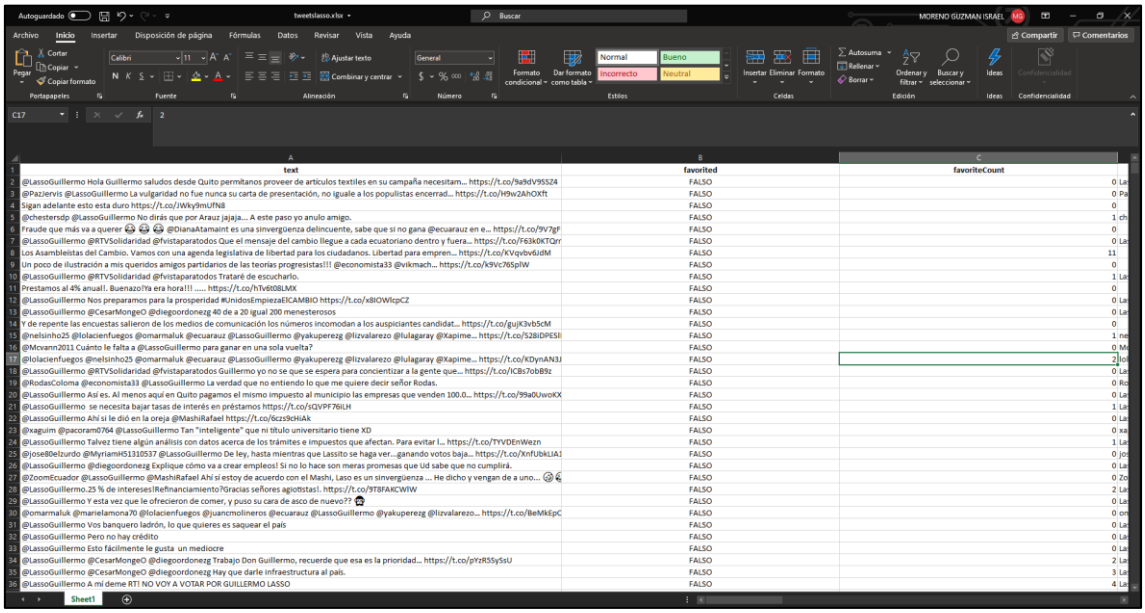
```
#Etapa de Recoleccion o Captura
# extrayendo tweets relacionados con la cuenta oficial de Guillermo Lasso
tweets<-twitteR::searchTwitter('LassoGuillermo',n=2000,lang='es',geocode='-1.483237,-74.850871,124.20mi')
#conservamos solo tweets no retweets
tweets<-twitteR::strip_retweets(tweets)
# cuenta el numero de tweets
n.tweets<-length(tweets)
n.tweets
#convierte la lista en un data frame
tweets.df<-twListToDF(tweets)
#permite visualizar el data frame
View(tweets.df)
#Guardamos el data frame como archivo en formato xls
# convierte en xls
writexl::write_xlsx(tweets.df,"C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_espanol\\tweetslasso.xlsx")
```

Ilustración 3.23 (Moreno, 2020)

Los nuevos parámetros dentro de la línea 15 en la Ilustración 3.23 incluyen:

- **lang** permite especificar el lenguaje en el que se quieren recuperar los tweets. En este caso utilizaremos la abreviatura '**es**' para recuperar tweets en español.
- **Geocode** permite establecer la latitud, longitud y radio de búsqueda de los tweets. Los datos ingresados son los que obtuvimos de Google Earth Pro.

Al igual que con los otros tweets estos también fueron almacenados en archivos de extensión xlsx tal cual indica la Ilustración 3.24.



text	favorited	favoriteCount
@LassoGuillermo Hola Guillermo saludos desde Quito permitanos proveer de artículos textiles en su campaña necesitam... https://t.co/9a9dV95SZ4	FALSO	0
@Pazlervis @LassoGuillermo La vulgaridad no fue nunca su carta de presentación, no iguale a los populistas encerrad... https://t.co/H9w2AhOXft	FALSO	0
Sigan adelante esto esta duro https://t.co/JWky9mUfN8	FALSO	0
@chestersdp @LassoGuillermo No dirás que por Arauz jajaja... A este paso yo anulo amigo.	FALSO	1
Fraude que más va a querer 😂 😂 😂 @DianaAtamaint es una sinvergüenza delincuente, sabe que si no gana @ecuarauz en e... https://t.co/9V7gfl	FALSO	0
@LassoGuillermo @RTVSolidaridad @fvistaparatodos Que el mensaje del cambio llegue a cada ecuatoriano dentro y fuera... https://t.co/F63k0KTQrr	FALSO	0
Los Asambleistas del Cambio, Vamos con una agenda legislativa de libertad para los ciudadanos. Libertad para empren... https://t.co/KVqv8vdJdM	FALSO	11
Un poco de ilustración a mis queridos amigos partidarios de las teorías progresistas!!! @economista33 @vikmach... https://t.co/K9Vc76SplW	FALSO	0
@LassoGuillermo @RTVSolidaridad @fvistaparatodos Trataré de escucharlo.	FALSO	1
Prestamos al 4% anual!. BuenazoIYa era hora!!! https://t.co/hTv6t0BLMX	FALSO	0
@LassoGuillermo Nos preparamos para la prosperidad #UnidosEmpiezaElCAMBIO https://t.co/x8IOWlcpCZ	FALSO	0
@LassoGuillermo @CesarMongeO @diegoordonezg 40 de a 20 igual 200 menesterosos	FALSO	0
Y de repente las encuestas salieron de los medios de comunicación los números incomodan a los auspiciantes candidat... https://t.co/gujK3vb5cM	FALSO	0
@nelsinho25 @lolacienfuegos @omarmaluk @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo @lulagaray @Xapime... https://t.co/52BIDPESIi	FALSO	1
@Mcvann2011 Cuánto le falta a @LassoGuillermo para ganar en una sola vuelta?	FALSO	0
@lolacienfuegos @nelsinho25 @omarmaluk @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo @lulagaray @Xapime... https://t.co/KDynAN3J	FALSO	2
@LassoGuillermo @RTVSolidaridad @fvistaparatodos Guillermo yo no se que se espera para concientizar a la gente que... https://t.co/ICBs7obB9z	FALSO	0
@RodasColoma @economista33 @LassoGuillermo La verdad que no entiendo lo que me quiere decir señor Rodas.	FALSO	0
@LassoGuillermo Así es. Al menos aquí en Quito pagamos el mismo impuesto al municipio las empresas que venden 100.0... https://t.co/99a0UwoKX	FALSO	0
@LassoGuillermo se necesita bajar tasas de interés en préstamos https://t.co/SQVPF76ILH	FALSO	1
@LassoGuillermo Ahí si le dió en la oreja @MashiRafael https://t.co/6czs9cHiAk	FALSO	0
@xaguim @pacoram0764 @LassoGuillermo Tan "inteligente" que ni título universitario tiene XD	FALSO	0
@LassoGuillermo Talvez tiene algún análisis con datos acerca de los trámites e impuestos que afectan. Para evitar l... https://t.co/TYVDEnWezn	FALSO	1
@Jose80elzurdo @MyriamH51310537 @LassoGuillermo De ley, hasta mientras que Lassito se haga ver...ganando votos baja... https://t.co/XnfUbkLIA1	FALSO	0
@LassoGuillermo @diegoordonezg Explique cómo va a crear empleos! Si no lo hace son meras promesas que Ud sabe que no cumplirá.	FALSO	0
@ZoomEcuador @LassoGuillermo @MashiRafael Ahí si estoy de acuerdo con el Mashi, Laso es un sinvergüenza ... He dicho y vengan de a uno...	FALSO	0
@LassoGuillermo.25 % de intereses!Refinanciamiento?Gracias señores agiottas1. https://t.co/9TBFAMCWIW	FALSO	2
@LassoGuillermo Y esta vez que le ofrecieron de comer, y puso su cara de asco de nuevo?? ☎️	FALSO	0
@omarmaluk @marielamona70 @lolacienfuegos @juancmolineros @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo... https://t.co/BeMkEpC	FALSO	0
@LassoGuillermo Vos banquero ladrón, lo que quieres es saquear el país	FALSO	0
@LassoGuillermo Pero no hay crédito	FALSO	0
@LassoGuillermo Esto fácilmente le gusta un mediocre	FALSO	0
@LassoGuillermo @CesarMongeO @diegoordonezg Trabajo Don Guillermo, recuerde que esa es la prioridad... https://t.co/pYzR55y5sU	FALSO	2
@LassoGuillermo @CesarMongeO @diegoordonezg Hay que darle infraestructura al país.	FALSO	3
@LassoGuillermo A mí deme RT! NO VOY A VOTAR POR GUILLERMO LASSO	FALSO	4

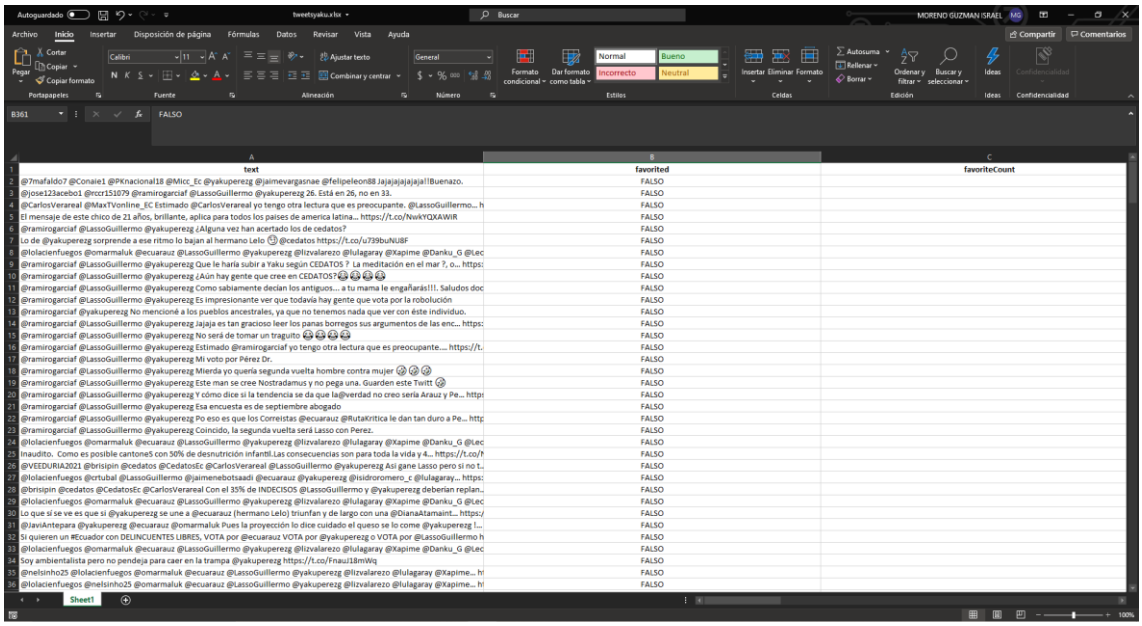
Ilustración 3.24 (Moreno, 2020)

Para el candidato Yaku Pérez se realizó el mismo proceso y las líneas de código para su recuperación de tweets se muestran en la Ilustración 3.25:

```
13  #Etapa de Recoleccion o Captura
14  # extrayendo tweets relacionados con la cuenta oficial de Yaku Perez
15  tweets<-twitteR::searchTwitter('yakuperezg',n=2000,lang='es',geocode='-1.483237,-74.850871,124.20mi')
16  #conservamos solo tweets no retweets
17  tweets<-twitteR::strip_retweets(tweets)
18  # cuenta el numero de tweets
19  n.tweets<-length(tweets)
20  n.tweets
21  #convierte la lista en un data frame
22  tweets.df<-twListToDF(tweets)
23  #permite visualizar el data frame
24  View(tweets.df)
25  #Guardamos el data frame como archivo en formato xls
26  # convierte en xls
27  writexl::write_xlsx(tweets.df,"C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_espanol\\tweetsyaku.xlsx")
```

Ilustración 3.25 (Moreno, 2020)

La cantidad de tweets recuperados para el candidato Yaku Pérez suman la cantidad de 360.



text	favorited	favoriteCount
@7mafaldo7 @Conaie1 @PKnacional18 @Micc_Ec @yakuperezg @jaimevargasnae @felipeleon88 Jajajajajajaja!!Buenazo.	FALSO	
@jose123acebo1 @rccr151079 @ramirogarciaf @LassoGuillermo @yakuperezg 26. Está en 26, no en 33.	FALSO	
@CarlosVerareal @MaxTVonline_EC Estimado @CarlosVerareal yo tengo otra lectura que es preocupante. @LassoGuillermo... h	FALSO	
El mensaje de este chico de 21 años, brillante, aplica para todos los paises de america latina... https://t.co/NwkYQXAWiR	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg ¿Alguna vez han acertado los de cedatos?	FALSO	
Lo de @yakuperezg sorprende a ese ritmo lo bajan al hermano Lelo 🤔 @cedatos https://t.co/u739buNU8F	FALSO	
@lolacienfuegos @omarmaluk @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo @lulagaray @Xapime @Danku_G @Lec	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Que le haría subir a Yaku según CEDATOS ? La meditación en el mar ?, o... https:	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg ¿Aún hay gente que cree en CEDATOS?😂😂😂😂	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Como sabiamente decían los antiguos... a tu mama le engañarás!!!. Saludos doc	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Es impresionante ver que todavía hay gente que vota por la robolución	FALSO	
@ramirogarciaf @yakuperezg No mencioné a los pueblos ancestrales, ya que no tenemos nada que ver con éste individuo.	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Jajaja es tan gracioso leer los panas borregos sus argumentos de las enc... https:	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg No será de tomar un traguito 😂😂😂😂	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Estimado @ramirogarciaf yo tengo otra lectura que es preocupante.... https://t.	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Mi voto por Pérez Dr.	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Mierda yo quería segunda vuelta hombre contra mujer 🤔🤔🤔	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Este man se cree Nostradamus y no pega una. Guarden este Twitt 🤔	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Y cómo dice si la tendencia se da que la@verdad no creo sería Arauz y Pe... https	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Esa encuesta es de septiembre abogado	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Po eso es que los Correistas @ecuarauz @RutaKritica le dan tan duro a Pe... http	FALSO	
@ramirogarciaf @LassoGuillermo @yakuperezg Coincido, la segunda vuelta será Lasso con Perez.	FALSO	
@lolacienfuegos @omarmaluk @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo @lulagaray @Xapime @Danku_G @Lec	FALSO	
Inaudito. Como es posible cantoneS con 50% de desnutrición infantil.Las consecuencias son para toda la vida y 4... https://t.co/h	FALSO	
@VEEDURIA2021 @brisipin @cedatos @CedatosEc @CarlosVerareal @LassoGuillermo @yakuperezg Asi gane Lasso pero si no t...	FALSO	
@lolacienfuegos @ortubal @LassoGuillermo @jaimenebotsaadi @ecuarauz @yakuperezg @isidroromero_c @lulagaray... https:	FALSO	
@brisipin @cedatos @CedatosEc @CarlosVerareal Con el 35% de INDECISOS @LassoGuillermo y @yakuperezg deberían replan...	FALSO	
@lolacienfuegos @omarmaluk @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo @lulagaray @Xapime @Danku_G @Lec	FALSO	
Lo que sí se ve es que si @yakuperezg se une a @ecuarauz (hermano Lelo) triunfan y de largo con una @DianaAtamaint... https:/	FALSO	
@JaviAntepara @yakuperezg @ecuarauz @omarmaluk Pues la proyección lo dice cuidado el queso se lo come @yakuperezg !...	FALSO	
Si quieren un #Ecuador con DELINCUENTES LIBRES, VOTA por @ecuarauz VOTA por @yakuperezg o VOTA por @LassoGuillermo h	FALSO	
@lolacienfuegos @omarmaluk @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo @lulagaray @Xapime @Danku_G @Lec	FALSO	
Soy ambientalista pero no pendeja para caer en la trampa @yakuperezg https://t.co/Fnauz18miWq	FALSO	
@nelsinho25 @lolacienfuegos @omarmaluk @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo @lulagaray @Xapime... ht	FALSO	
@lolacienfuegos @nelsinho25 @omarmaluk @ecuarauz @LassoGuillermo @yakuperezg @lizvalarezo @lulagaray @Xapime... ht	FALSO	

Ilustración 3.26 (Moreno, 2020)

La Ilustración 3.26 muestra todos los tweets recuperados para el candidato Yaku Pérez en formato xlsx.

Capítulo 4 Procesamiento de Datos

Durante esta etapa nos encargaremos de procesar solamente la columna que contiene el texto de los tweets dentro de los archivos xlsx. El procesamiento para cada archivo es el mismo pero los datos y resultados que obtengamos serán diferentes.

4.1 Transformación de los Tweets

Para obtener conclusiones y comenzar a identificar relaciones entre los tweets y aquello que expresan, estos deben ser limpiados y analizados a discreción del investigador. Este proceso de transformación se puede llegar a repetir las veces que sean necesarias.

4.1.1 Transformación de los Tweets a un Corpus

La **librería tm** permite la transformación de nuestros tweets a un corpus. Un corpus es una colección de documentos que contienen (lenguaje natural) texto. La agrupación de todos los tweets en un solo documento (corpus) permite optimizar las operaciones de procesamiento de estos.

La transformación de los tweets a un corpus se realiza de la siguiente forma, tal como indica la Ilustración 4.1:

```
3  library(xlsx) # para leer archivos xlsx
4
5  #leemos el archivo xlsx que contiene los tweets y lo transformamos en data frame
6  tweets.df<-read.xlsx("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
7                       sheetIndex = 1,
8                       startRow = 1,
9                       colIndex = 1)
10
11 # convirtiendo a todo en un solo corpus
12 myCorpus<-Corpus(VectorSource(tweets.df$text))
```

Ilustración 4.1 (Moreno, 2020)

La línea 6 del script permite leer el archivo tweetstrump.xlsx y transformarlo a un data frame para después convertirlo a un corpus en la línea 12.

Si inspeccionamos **el corpus myCorpus** podemos encontrar que todos los tweets han sido convertidos a corpus. Para la inspección de corpus utilizamos el método **inspect(myCorpus)**.

```
3  library(xlsx) # para leer archivos xlsx
4
5  #leemos el archivo xlsx que contiene los tweets y lo transformamos en data frame
6  tweets.df<-read.xlsx("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
7                       sheetIndex = 1,
8                       startRow = 1,
9                       colIndex = 1)
10
11 # convirtiendo a todo en un solo corpus
12 myCorpus<-Corpus(VectorSource(tweets.df$text))
13 #inspeccion del corpus
14 inspect(myCorpus)
```

Ilustración 4.2 (Moreno, 2020)

La Ilustración 4.2 muestra la transformación del archivo xlsx original a un corpus.

Una vez ejecutado **inspect(myCorpus)** visualizaremos al corpus desde la consola de R.

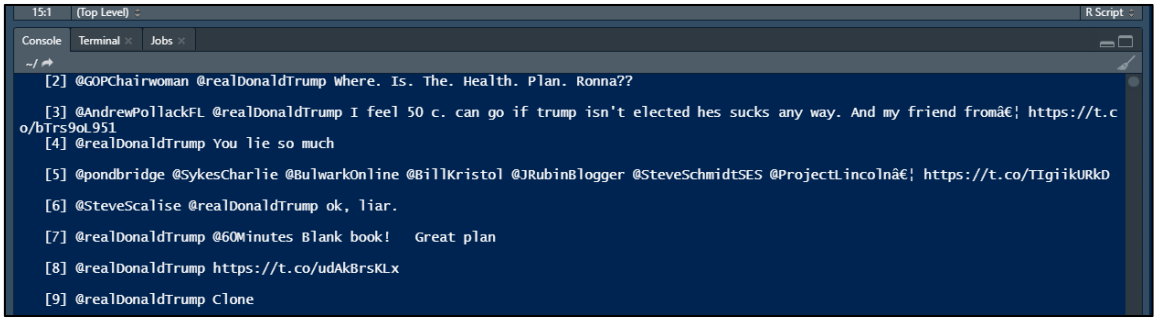


Ilustración 4.3 (Moreno, 2020)

La Ilustración 4.3 muestra como luce un corpus al momento de utilizar el método inspect().

Este proceso se realizará con todos los archivos xlsx que guardamos durante la etapa de captura.

4.1.2 Limpieza y Normalización del Corpus

Seguramente está sea la parte más importante dentro de todo el proceso ya que a partir de aquí podremos comenzar a construir relaciones para la construcción del conocimiento y la obtención de los resultados finales.

La limpieza de los tweets que componen el corpus es bastante importante debido a que de esta manera seremos capaces de remover todo el “ruido” que viene dentro del corpus.

```
10 #convirtiendo el data frame en un corpus para su normalización
11 myCorpus<-Corpus(VectorSource(tweets.df$text))
12 #inspección del corpus
13 inspect(myCorpus)
14 #limpieza y normalización del contenido del corpus
15 corpus_limpio<-tm_map(myCorpus,tolower)#convierte todo el texto en minúsculas
16 #removemos urls innecesarias dentrol del corpus
17 remove_url<- function(x) gsub("http[^[:space:]]*","",x)
18 corpus_limpio<- tm_map(corpus_limpio,content_transformer(remove_url))
19 #removiendo los número dentro del corpus
20 corpus_limpio<-tm_map(corpus_limpio,removeNumbers)
21 #removiendo signos de puntuación
22 corpus_limpio<-tm_map(corpus_limpio,removePunctuation)
23 #removiendo emojis
24 removeEmojo<-function(x)gsub("[^\x01-\x7F]","",x)
25 corpus_limpio<-tm_map(corpus_limpio,content_transformer(removeEmojo))
26 #removiendo palabras de parada
27 corpus_limpio<-tm_map(corpus_limpio,removeWords,stopwords(kind = "en"))
28 #encontrnado las raices de las palabras
29 corpus_limpio<-tm_map(corpus_limpio,stemDocument)#se utiliza el algoritmo de Porter
```

Ilustración 4.4 (Moreno, 2020)

De la línea 10 a la línea 29 ocurre toda la transformación, depuración y normalización del corpus. La Ilustración 4.4 se encarga de mostrar el proceso.

El proceso de limpieza del corpus corresponde a los siguientes pasos:

- Transformar todo el texto a minúsculas para facilitar el tratamiento de todas las palabras dentro del corpus. Esto ocurre en la línea 15.

- Remover URL innecesarias dentro los tweets, debido a que no aportan valor al análisis. Las líneas 17y 18 se encargan de realizar esta tarea.
- Se remueven todos los números del corpus en la línea 20. Lo que nos interesan son las palabras.
- Remover los signos de puntuación. Estos no aportan valor al momento de realizar un análisis. Son removidos en la línea 22.
- Los emoticones se encargan de expresar emociones, pero no son muy útiles al momento de realizar minería de texto. Todos los emoticones son removidos en las líneas 24 y 25.
- Las palabras de parada (artículos, pronombres) son removidos en la línea 27.
- En la línea 29 se buscan las raíces de todas las palabras dentro del corpus utilizando el algoritmo de Porter.

El algoritmo de Porter se encarga de encontrar la raíz de las palabras que residen en el corpus creado. Está técnica es realizada una vez el texto ha sido normalizado (depurado y preparado para análisis).

Capítulo 5 Aplicación del Análisis de Sentimientos

Durante este capítulo se procesará al corpus ya normalizado de los cuatro scripts que tenemos (dos en inglés y español) con el objetivo de responder a las preguntas planteadas durante los capítulos 3.3.2 y 3.3.3 utilizando librerías y herramientas presentes en el IDE R Studio. Además, se aplicará el análisis de sentimiento a nivel de oración y palabra.

5.1 Identificación y Clasificación de Sentimientos

5.1.1 Nubes de Palabras

Una de las formas más sencillas de identificar o realizar un análisis de sentimientos es utilizar una nube de palabras, debido a que dentro de este gráfico podremos identificar cuáles son las más usadas dentro de un texto o corpus en concreto. Esto nos puede permitir tener una idea bastante clara sobre la polaridad de este.

Dentro R Studio la construcción de una nube de palabras lo hacemos con la ayuda de la **librería worldcloud (versión 2.6)** la cual cuenta con varias herramientas que nos permitirán visualizar los sentimientos dentro de nuestro texto.

La construcción de la nube de palabras definitiva no se logrará a la primera debido a que seguramente se filtraran varias palabras que no aporten valor a nuestro análisis.

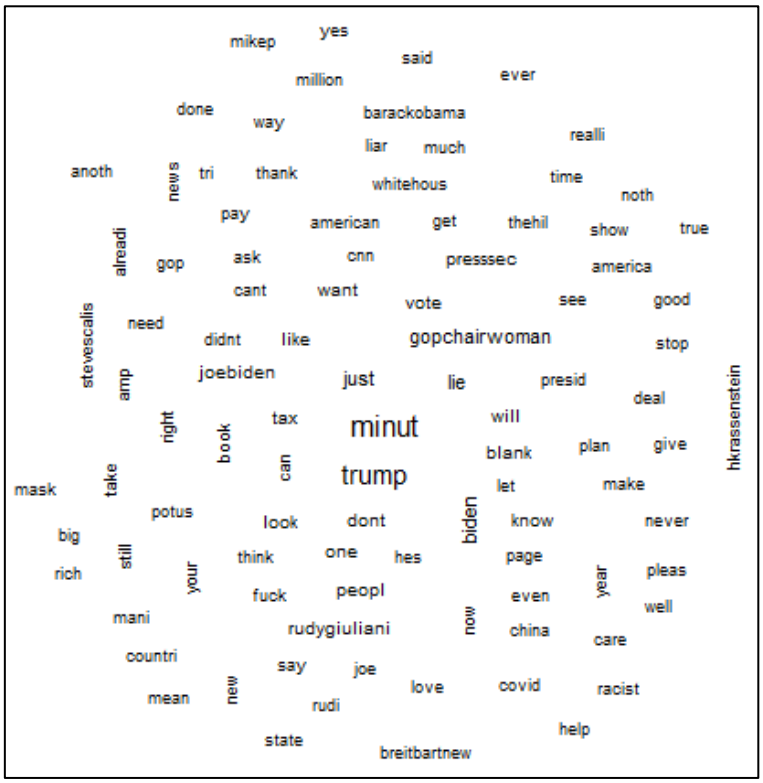


Ilustración 5.1 (Moreno, 2020)

La Ilustración 5.1 nos muestra un primer resultado de nuestra nube (correspondiente a los tweets de Donald Trump) de palabras. Dicha ilustración no es el resultado final del análisis, pero si nos permite tener una primera aproximación a lo que puede ser nuestra nube de palabras.

Una nube de palabras que se encargue de indicar la realidad frente a la cual se encuentra enfrentado el personaje al cual estamos analizando debe pasar por un proceso de depuración el cual ocurre durante la limpieza de nuestro corpus. Aquello que permite tener una nube de palabras bastante depurada es el algoritmo de Porter y los parámetros **removewords** y **stopwords(kind = "en")** dentro de la función **tm_map** la cual es parte de la librería **tm.** Aun así, eso no nos garantiza que todo aquello que se encuentre dentro de la nube de palabras sea algo que aporte valor analítico. Cuando ello ocurre deberemos crear un arreglo que contenga a todas aquellas palabras que no aportan algún valor a la nube de palabras final.

```
#arreglo de stopwords
corpus_limpio<-tm_map(corpus_limpio,removeWords,c('realdonaldtrump','itsjefftiedrich','gopchairwoman','trump',
                                                  'therightmelissa','rudygiuliani','breitbartnew',
                                                  'page','that','minut','rudi','potus','joebiden',
                                                  'joe','biden','thehil','get','book','tri','let','gop','your',
                                                  'even','just','noth','obama','amp','presssec','mani','anoth',
                                                  'one','look','hkrassenstein','barackobama',
                                                  'thing','new','stevescalis'))
```

Ilustración 5.2 (Moreno, 2020)

```
corpus_limpio<-tm_map(corpus_limpio,removeWords,c('joebiden','barackobama','kamalaharri','senronjohnson','trump',
                                                  'biden','steveschmidts','joe','realdonaldtrump','ananavarro',
                                                  'tri','rvat','projectlincoln','drbiden','aoc','one',
                                                  'kayleighmcenani','obama','doesnt','toddfoot','gabbygifford',
                                                  'educ','that','youramerican','man',
                                                  'amp'))
```

Ilustración 5.3 (Moreno, 2020)

La Ilustración 5.2 nos muestra el arreglo creado para la filtración de palabras correspondientes a la nube de Donald Trump y la Ilustración 5.3 para la nube de Joe Biden.

La construcción de cualquiera de las dos nubes de palabras se llevó de la siguiente manera:

```
40  #construcción de una Nube de Palabras
41  wordcloud(corpus_limpio,min.freq =10,random.order = F)
42  wordcloud(corpus_limpio,max.words = 75,colors = brewer.pal(8,"Set2"),random.order = F,rot.per = .10)
```

Ilustración 5.4 (Moreno, 2020)

- La línea N°41 permite visualizar a nuestra nube de palabras de la forma más básica posible. La Ilustración 5.4 corresponde a esta línea de código.

- La línea N°42 nos permite visualizar a nuestra nube de palabras mucho más estilizada debido a que las palabras se encuentran representadas por colores con ayuda de la librería **RColorBrewer** la cual es llamada al momento de cargar la librería **wordcloud**.

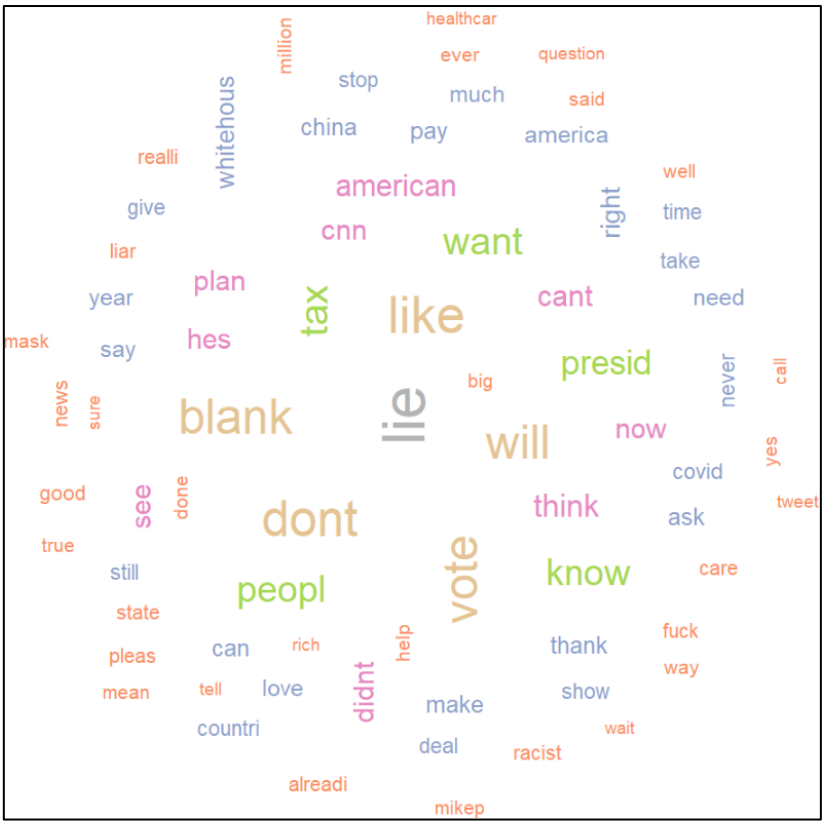


Ilustración 5.5 (Moreno, 2020)

La Ilustración 5.5 nos muestra una nube de palabras mucho más clara y limpia. Debido a que ya comienza a indicarnos el tipo de sentimientos generales de la gente sobre Donald Trump. Aun así, no representa la versión final de la misma.

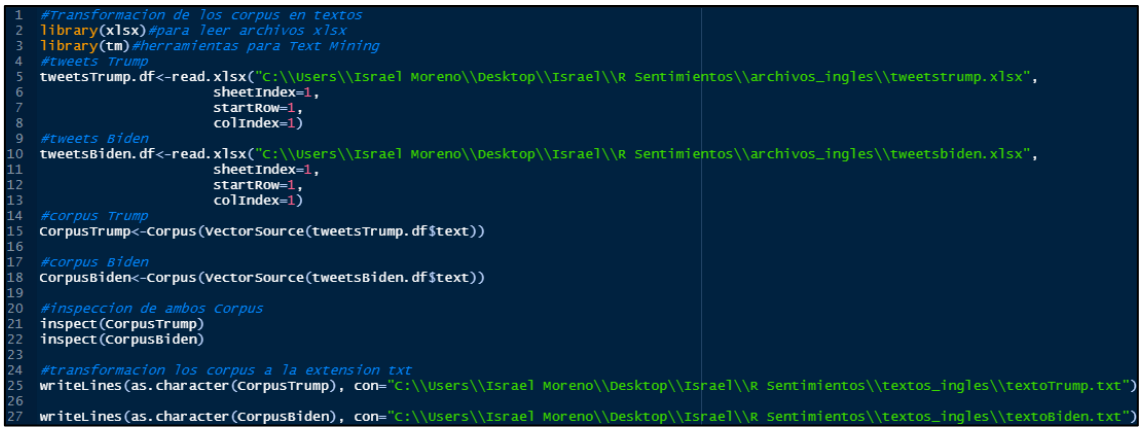
5.1.2 Comparando Nubes de Palabras

Una sola nube de palabras es capaz de proveernos suficientes ideas acerca de un tema, texto, opinión, etc. Aun así, el verdadero potencial de las nubes de palabras se puede explotar aún más al comparar dos de estas, debido a que podemos analizar dos puntos de vista diferentes sobre un mismo tema, punto de vista a lo largo del tiempo, o la evolución de un discurso político al inicio y final de un período.

El comparar las nubes de palabras de Trump y Biden nos permitirán tener una perspectiva general del público sobre ambos durante el período de campaña política en Estados Unidos. De la misma manera dentro del panorama nacional podemos hacer lo mismo comparando las nubes de Guillermo Lasso y Yaku Pérez.

El proceso de comparación de nubes se realizó de la siguiente forma:

- Transformación de los archivos XLSX donde se almacenaron los tweets a .TXT para posteriormente reubicarlos en una misma carpeta (para facilitar la construcción de la comparación de nubes).



```
#Transformacion de los corpus en textos
library(xlsx)#para leer archivos xlsx
library(tm)#herramientas para Text Mining
#tweets Trump
tweetsTrump.df<-read.xlsx("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
                          sheetIndex=1,
                          startRow=1,
                          colIndex=1)
#tweets Biden
tweetsBiden.df<-read.xlsx("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetsbiden.xlsx",
                          sheetIndex=1,
                          startRow=1,
                          colIndex=1)
#corpus Trump
CorpusTrump<-Corpus(VectorSource(tweetsTrump.df$text))

#corpus Biden
CorpusBiden<-Corpus(VectorSource(tweetsBiden.df$text))

#inspeccion de ambos corpus
inspect(CorpusTrump)
inspect(CorpusBiden)

#transformacion los corpus a la extension txt
writeLines(as.character(CorpusTrump), con="C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\textos_ingles\\textoTrump.txt")

writeLines(as.character(CorpusBiden), con="C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\textos_ingles\\textoBiden.txt")
```

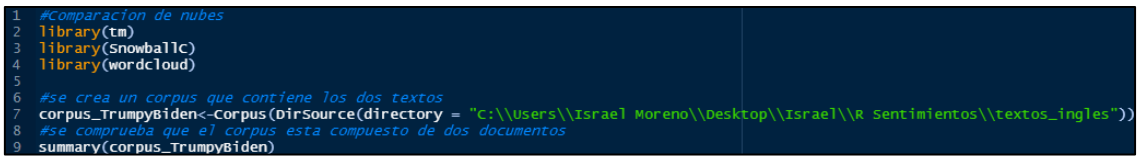
Ilustración 5.6 (Moreno, 2020)

Debido a que R no cuenta con una librería que transforma de manera directa un archivo XLSX a TXT se transformó a ambos archivos primero a un corpus que solo tomaba el campo texto. Esto ocurre entre las líneas 5-18 de la Ilustración 5.6.

Con el objetivo de verificar que la transformación fue exitosa y se está recuperando únicamente la información que se desea se inspeccionan ambos corpus entre las líneas 15 y 18.

R permite transformar a un corpus a la extensión TXT con el método **writeLines()** siempre y cuando se le indique con el método **as.character()** que aquello que se va escribir en el TXT son caracteres. De preferencia como se indicó con anterioridad se deben almacenar ambos archivos en un mismo directorio.

- Posteriormente se creó otro script en R Studio dentro del cual se procesarán ambos archivos TXT para convertirlos en nubes de palabras.



```
#Comparacion de nubes
library(tm)
library(SnowballC)
library(wordcloud)

#se crea un corpus que contiene los dos textos
corpus_TrumpyBiden<-Corpus(DirSource(directory = "C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\textos_ingles"))
#se comprueba que el corpus esta compuesto de dos documentos
summary(corpus_TrumpyBiden)
```

Ilustración 5.7 (Moreno, 2020)

Para crear la nube de palabras comparativa se importan las mismas librerías con las que hemos estado trabajando en las nubes de palabras individuales. Posteriormente creamos un corpus que contiene a los dos archivos TXT creados en el anterior script.

Para asegurarnos que el proceso ha sido exitoso utilizamos el método **summary()** el cual nos permite revisar que el corpus está compuesto de dos archivos tal cual indica la línea 9 de la Ilustración 5.7.

```
11  #Limpieza de los textos y el corpus
12  corpus_limpio<-tm_map(corpus_TrumpyBiden,tolower)#convierte todo el texto en minúsculas
13  #removemos urls innecesarias dentrol del corpus
14  remove_url<- function(x) gsub("http[^[:space:]]*","",x)
15  corpus_limpio<- tm_map(corpus_limpio,content_transformer(remove_url))
16  #removiendo los numeros dentro del corpus
17  corpus_limpio<-tm_map(corpus_limpio,removeNumbers)
18  #removiendo signos de puntuacion
19  corpus_limpio<-tm_map(corpus_limpio,removePunctuation)
20  #removiendo emojis
21  removeEmojo<-function(x)gsub("[^\x01-\x7F]","",x)
22  corpus_limpio<-tm_map(corpus_limpio,content_transformer(removeEmojo))
23  #removiendo espacios
24  corpus_limpio<-tm_map(corpus_limpio,stripWhitespace)
25  #removiendo palabras de parada
26  corpus_limpio<-tm_map(corpus_limpio,removeWords,stopwords(kind = "en"))
27  #encontrnado las raíces de las palabras
28  corpus_limpio<-tm_map(corpus_limpio,stemDocument)#se utiliza el algoritmo de Porter
```

Ilustración 5.8 (Moreno, 2020)

Desde la línea 11 a la 28 tal cual indica la Ilustración 5.8 repetimos el mismo proceso que se ha realizado con los corpus previamente procesados en los cuales nos encargamos de limpiar y depurar lo más que se puede a estos.

- La parte más exhaustiva del proceso es remover todas aquellas palabras que no aportan nada. Estas pueden ser adjetivos, palabras que no tienen ningún significado en particular (arrojadas principalmente por el algoritmo de Porter).

```
30  corpus_limpio<-tm_map(corpus_limpio,removeWords,c('joebiden','realdonaldtrump','barackobama','steveschmidts','gopchairwoman',
31                                                     'minut','therightmelissa','kamalaharri','trump','senronjohnson','rudygiuliani',
32                                                     'joe','stevescalis','book','thehil','aoc','ananavarro','hkrassenstein',
33                                                     'kayleighmcenani','gop','empti','itsjefftiedrich','page','biden','ivankatrump',
34                                                     'rvat','projectlincoln','breitbartnew','drbiden','toddfoot','hunter',
35                                                     'mikep','devo','email','today','potus','that','gabbygifford','caslernoel',
36                                                     'sachabaroncohen','walshfreedom','amp','thedemcoalit','unleashthetea',
37                                                     'youramerican','nataliegwint','dcexamin','shes','juli',
38                                                     'scottpresl','week','markdic','ines','donald','proudsocialist','blank',
39                                                     'mmpadellan','ilhanmnbeliev','ronna','keitholbermann','choic','call',
40                                                     'anoth','ilhanmn','impeach','point','bidenrep','markmeadow','youv','stahl',
41                                                     'lesley','shadi','paper','wolfblitz','ericbool','mean','ericbol',
42                                                     'secpompeo','seen','cri','howardmortman','sarahdauterman','get','espn',
43                                                     'bidenharri','nwhile','use','got','repdougcollin','checkmatest','ign',
44                                                     'mani','presssec','son','cnnsitroom','asianamerican','ron','cent',
45                                                     'republican','yes','parti','pleas','seem','question','look',
46                                                     'whitehouse','even','whitehous','say','russianron','makeamericadecentagain',
47                                                     'pennsylvania','traffick','florida','come','big'))
```

Ilustración 5.9 (Moreno, 2020)

De la línea 30 a la 47 de la Ilustración 5.9 podemos observar todas las palabras que se han removido de la nube de palabras comparativas debido a que no aportaban algún valor al análisis. La cantidad de palabras que se remueve no es una cantidad fija y tampoco se la realiza con alguna fórmula en específico, está tarea depende netamente del investigador o científico de datos.

Al momento de realizar la depuración de palabras se busca que el investigador tenga una posición neutral y muestre los resultados cómo son. En esta etapa se debe evitar cualquier sesgo.

- La construcción de la nube comparativa es bastante similar a la de una sola nube. Lo importante en esta etapa es etiquetar adecuadamente las nubes de palabras y organizar a nuestros datos en matrices.

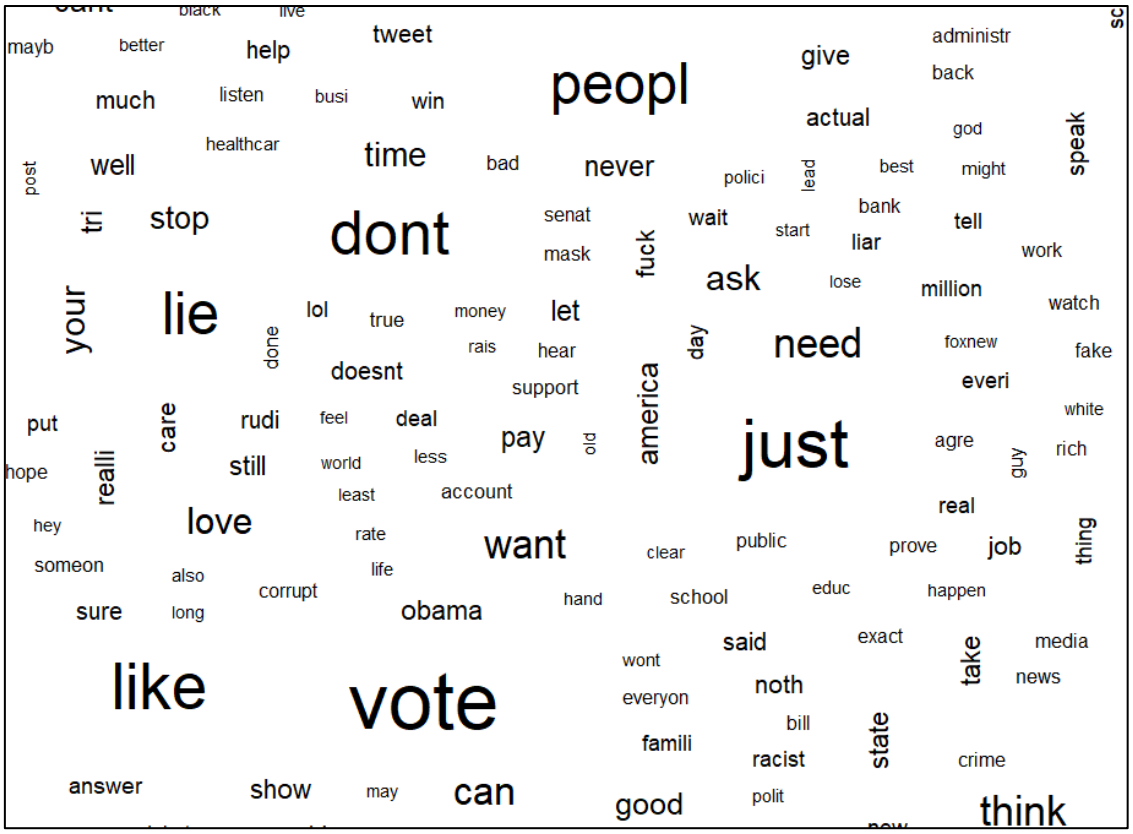


Ilustración 5.10 (Moreno, 2020)

La Ilustración 5.10 nos muestra cómo es la expresión más simple de una nube de datos comparativa. En esta forma la nube no aporta ningún tipo de valor a la investigación aún.

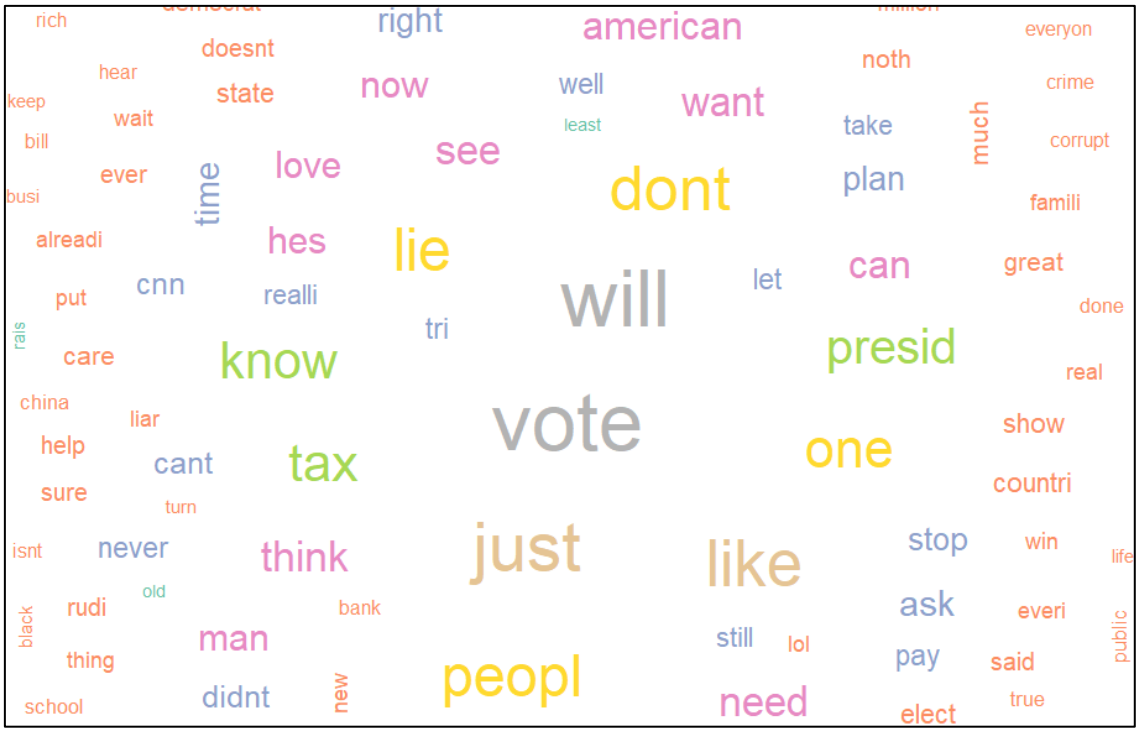


Ilustración 5.11 (Moreno, 2020)

La Ilustración 5.11 nos muestra una nube un poco más estilizada, pero que aun así no se encuentra debidamente etiquetada.

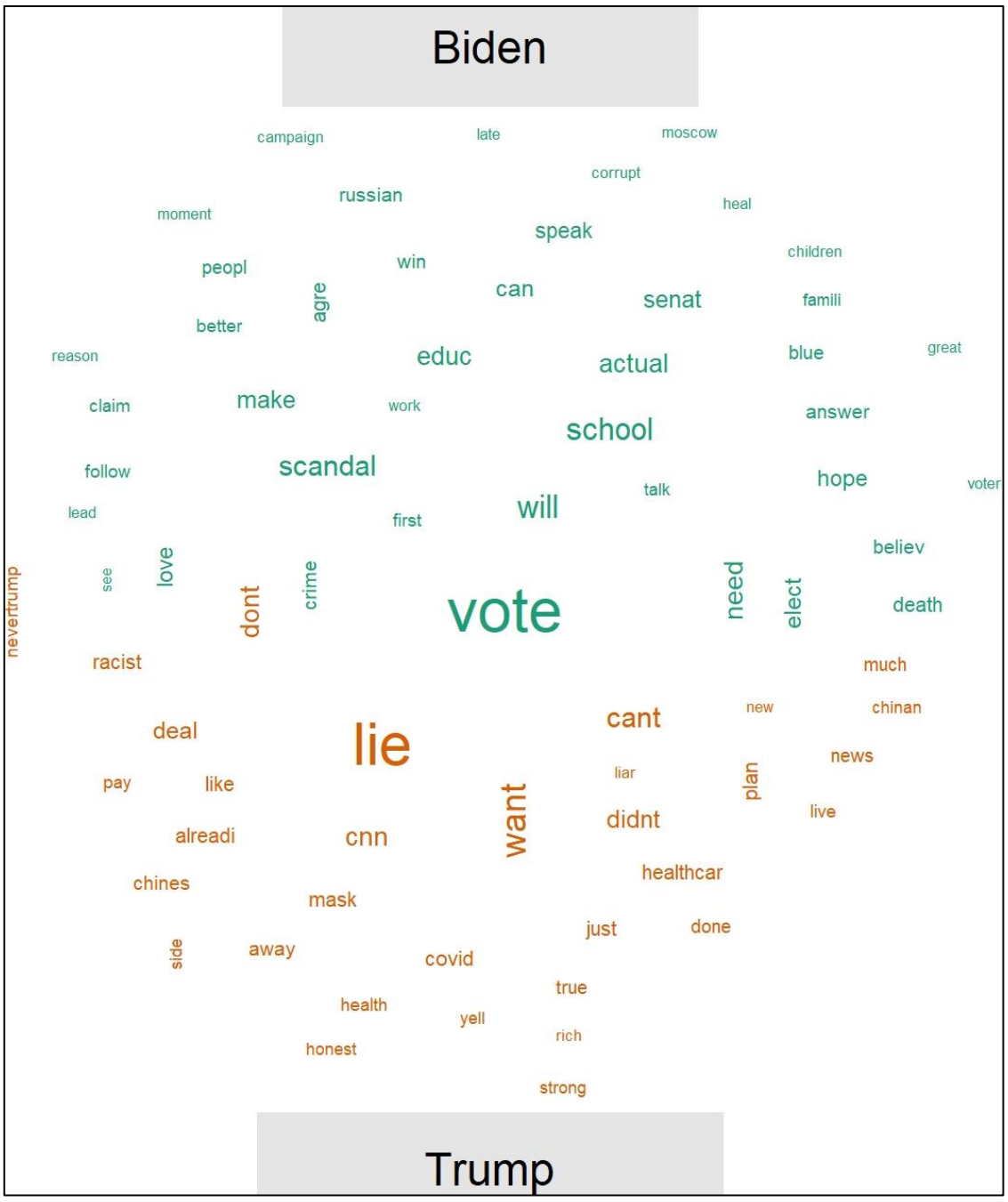


Ilustración 5.12 (Moreno, 2020)

La Ilustración 5.12 nos muestra la nube comparativa final en la cual ya podemos ver bien diferenciadas y etiquetadas a las palabras que corresponden a cada corpus.

```
48 wordcloud(corpus_limpio,min.freq = 10)
49 wordcloud(corpus_limpio,min.freq =10,colors = brewer.pal(8,"Set2"),random.order = F,rot.per = .10)
50
51 #Creando la Nube de Comparación
52 matrix_de_terminos<-TermDocumentMatrix(corpus_limpio) #crea una matriz de terminos es un objeto propio de r
53 matrix_de_terminos<-as.matrix(matrix_de_terminos)# la transforma en una matriz nxm
54 colnames(matrix_de_terminos)<-c('Biden','Trump')# nombramos a la nubes
55 comparison.cloud(matrix_de_terminos,max.words =75,random.order = F,rot.per = .10) #crea la nube de comparacion
```

Ilustración 5.13 (Moreno, 2020)

La construcción de la nube definitiva se llevó a cabo desde la línea 48 a la 55 tal cual indica la Ilustración 5.13. Las líneas 48 y 49 construyeron las nubes que corresponden a las ilustraciones 5.10 y 5.11 dentro de las cuales aún no identificábamos las palabras que pertenecían a cada corpus.

Las líneas 52 a 55 se encargan de organizar y etiquetar a cada una de las palabras en donde corresponden.

La línea 52 nos permite crear una matriz de términos, donde las columnas son las palabras presentes dentro del texto y las filas indican la presencia de dicha palabra por documento analizado.

Ejemplo:

1	Análisis de texto
2	Minería de texto
3	Análisis de sentimiento
4	Análisis de redes sociales

Tabla 5.1 (Moreno, 2020)

Nuestros documentos TXT al momento de ser transformados en corpus lucen cómo la Tabla 5.1. Cada línea del archivo de texto es identificada como un “documento”.

Al momento en el que creamos una matriz de términos nuestro corpus se transforma en la siguiente estructura:

	Análisis	de	texto	Minería	sentimiento	redes	sociales
Análisis de texto	1	1	1	0	0	0	0
Minería de texto	0	1	1	0	0	0	0
Análisis de sentimiento	1	1	0	0	1	0	0
Análisis de redes sociales	1	1	0	0	0	1	1

Tabla 5.2 (Moreno, 2020)

La matriz representada en la Tabla 5.2 se encarga de mostrarnos la presencia de una palabra dentro de un documento con una notación binaria.

Este proceso se realiza para ambos corpus y luego se los transforma en una matriz **nxm** en la cual se suman la cantidad de incidencias de las palabras de cada uno de los corpus. De esta manera obtenemos los totales de cada palabra que se encuentra dentro de nuestro corpus. Finalmente agregamos a quien

pertenece el total de palabras de cada matriz en la línea 54 y graficamos en la línea 55 la nube de palabras de la Ilustración 5.12.

5.2 Análisis a Nivel de Sentimiento

Dentro del capítulo 2.3.1 de esta disertación se mencionan los distintos tipos de análisis de sentimiento que existen y cómo funcionan. En esta sección se indicará de qué manera se realizó la aplicación de la teoría y que herramientas se utilizaron para lograr dicho cometido.

Las herramientas que se han seleccionado para el análisis de sentimientos son las librerías **sentimentr (v 2.7.1)**, **tidyverse (v 1.3.0)** y **lexicón (v 1.2.1)**. Este conjunto permite realizar el respectivo análisis de sentimientos valiéndonos de métodos ya preparados, diccionarios y gráficos.

```
2  library(sentimentr)#Paquete completo para análisis de sentimientos
3  library(tidyverse) #Herramientas para Data Science
4  library(lexicon) #Lexicons disponibles
```

Ilustración 5.14 (Moreno, 2020)

La ilustración 5.14 muestra las tres librerías fundamentales para el Análisis a Nivel de Sentimiento en el lenguaje R.

```
7  tweets_trump.df<-read.xlsx("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
8                             sheetIndex=1,
9                             startRow=1,
10                            colIndex=1)
11
12 #Obteniendo sentimientos
13 tweets_trump.df %>%
14   get_sentences() %>%
15   sentiment()->tweets_trump_sentimiento #obtiene el sentimiento de manera individual
16
17 #Obteniendo sentimientos promedio
18 tweets_trump.df %>%
19   get_sentences() %>%
20   sentiment_by()->tweets_trump_sentimiento_promedio #obtiene el sentimiento por grupos
21
22 #Obteniendo la polaridad de los sentimientos, positivo negativo, neutral
23 tweets_trump.df %>%
24   get_sentences() %>%
25   extract_sentiment_terms()->tweets_trump_terminos_sentimiento
```

Ilustración 5.15 (Moreno, 2020)

Para comenzar con el análisis de sentimientos cargamos un archivo XLSX que contenga los tweets recuperados con anterioridad y seleccionamos solo la columna que contiene el texto de estos. Esto permitirá crear un **data frame** llamado **tweets_trump.df** tal cual indica la Ilustración 5.15. Esto ocurre en la línea 7 del script.

A partir del **data frame** llamado **tweets_trump.df** extraeremos los sentimientos de los tweets. Para optimizar el poder computacional del que se dispone, la documentación de la librería **sentimentr** indica que debemos utilizar el método **get_sentences()** con el objetivo de no sobrecargar la memoria de la computadora sobre todo cuando analizamos grandes cantidades de datos (miles o más). El método **get_senteces()** optimizará la lectura de cada uno de los tweets dentro de nuestro **data frame** al reconocerlos como cadenas de caracteres donde cada una es independiente de la otra.

Se crearán tres data frames distintos con el objetivo de facilitar el análisis de sentimiento. Los tres data frames son:

Nombre del data frame	Descripción	Método de sentimentr utilizado
tweets_trump_sentimiento en la línea 15.	Contiene el valor individual de cada palabra dentro de un tweet.	sentiment()
tweets_trump_sentimiento_promedio en la línea 20.	Contiene el valor promedio de cada tweet.	sentiment_by()
tweets_trump_terminos_sentimiento en la línea 25.	Contiene todas las palabras que denoten algún tipo de sentimiento dentro de los tweets.	extract_sentiment_terms()

Tabla 5.3 (Moreno, 2020)

Debido a que los data frames por sí solos no aportan gran valor a la investigación estos deben ser analizados y posteriormente resumidos de forma visual para obtener una mejor perspectiva del análisis.

La librería **tidyverse** incluye a una librería gráfica llamada **ggplot2 (v 3.3.3)** la cual permite crear, editar y personalizar gráficos que se adapten a nuestras necesidades y datos.

Probablemente la librería más importante dentro de este proceso sea **lexicon** debido a que permite cargar y utilizar hasta 10 diccionarios para aplicar la lematización a nuestros datos.

Los 10 diccionarios son:

Número de Diccionario	Nombre del Diccionario
1	**lexicon::hash_sentiment_jockers_rinker**
2	**lexicon::hash_sentiment_jockers**
3	**lexicon::emojis_sentiment**
4	**lexicon::hash_sentiment_huliu**
5	**lexicon::hash_sentiment_loughran_mcdonald**
6	**lexicon::hash_sentiment_nrc**

7	**lexicon::hash_sentiment_senticnet**
8	**lexicon::hash_sentiment_sentiword**
9	**lexicon::hash_sentiment_slangsd**
10	**lexicon::hash_sentiment_socal_google**

Tabla 5.4 (Moreno, 2020)

El diccionario que viene por defecto dentro de **sentimentr** es el número 1 desde la **versión 1.0.0** de la librería. Cualquier análisis que se haya realizado utilizando **sentimentr** previamente hacía uso del diccionario 4. En esta investigación haremos uso del diccionario que viene por defecto debido a que es el más reciente (año 2017).

Al complementar las tres librerías como se ha indicado hasta el momento obtendremos resultados iguales a los presentados en la Ilustración 5.15.

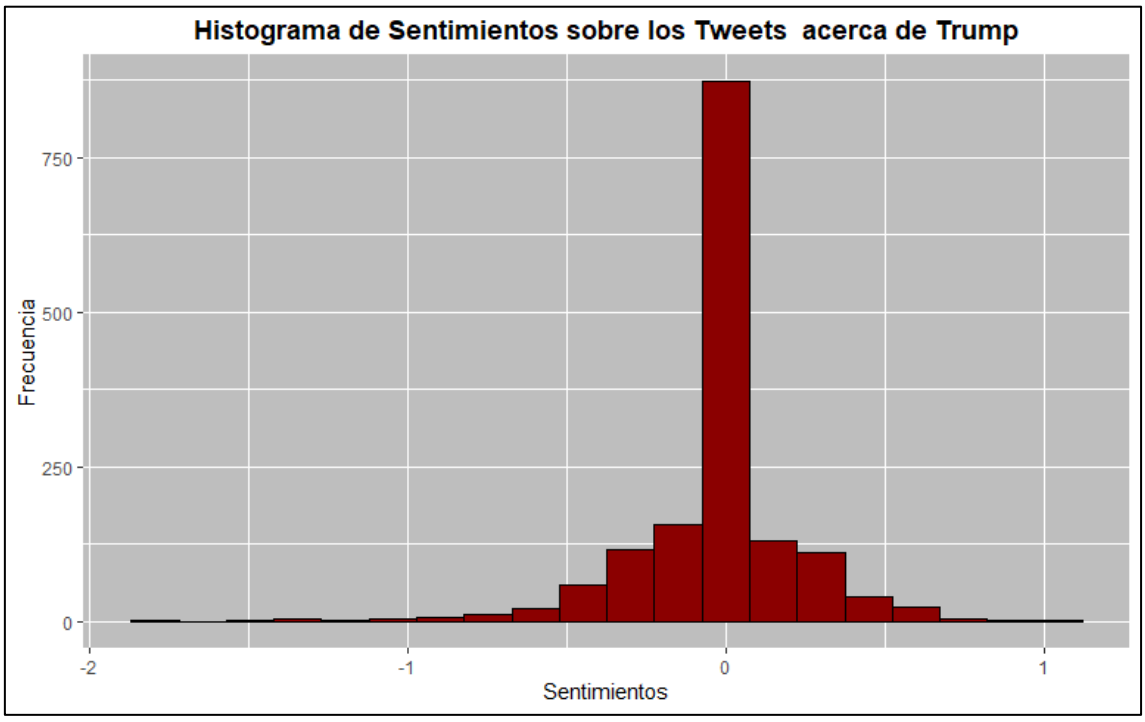


Ilustración 5.16 (Moreno, 2020)

La Ilustración 5.16 muestra la combinación de las tres librerías después de haber analizado al **data frame tweets_trump_sentimiento** y muestra la distribución del sentimiento que se tiene en los tweets de Trump.

5.2.1 Los Sentimientos

Para conocer el valor o sentimiento de una palabra necesitamos hacer uso de la lematización y un diccionario. Estos dos elementos nos permiten regresar o asociar a una palabra con su raíz mientras le asignamos un valor (positivo o negativo). Los valores de los sentimientos **pueden tomar valores entre -1 y 1** pero en caso de existir **transformadores de valencia** junto a una palabra el valor del sentimiento puede incrementar o decrementar, por esta razón podrían existir **valores que vayan más allá del -1 o 1**.

Con el objetivo de realizar un análisis a nivel de palabra se creó el data frame **tweets_trump_terminos_sentimiento.** Este data frame contiene todas las palabras lematizadas y clasificadas entre positivo, negativo y neutral.

negative	neutral	positive
character(0)	c("thehill", "realdonaldtrump", "actually", "i", "m", "in", "politi...	c("interested", "like")
character(0)	c("gopchairwoman", "realdonaldtrump", "where")	character(0)
character(0)	is	character(0)
character(0)	the	character(0)
character(0)	health	character(0)
character(0)	plan	character(0)
character(0)	ronna	character(0)
c("trump", "sucks")	c("andrewpollackfl", "realdonaldtrump", "i", "feel", "c", "can",...	character(0)
character(0)	c("and", "my", "from", "https", ":", "t", "co", "btrs", "ol")	friend
lie	c("realdonaldtrump", "you", "so", "much")	character(0)
character(0)	c("pondbridge", "sykescharlie", "bulwarkonline", "billkristol", ...	character(0)
liar	c("stevescalise", "realdonaldtrump", "ok")	character(0)
character(0)	c("realdonaldtrump", "minutes", "blank", "book")	character(0)
character(0)	plan	great
character(0)	c("realdonaldtrump", "https", ":", "t", "co", "udakbrsklx")	character(0)
character(0)	c("realdonaldtrump", "clone")	character(0)
character(0)	c("siempre", "confi", "en", "este", "tipo", "https", ":", "t", "co", ...	character(0)
character(0)	c("rudygiuliani", "represents", "realdonaldtrump", "they", "ju...	character(0)
character(0)	c("https", ":", "t", "co", "ff", "oaaajtq")	character(0)
character(0)	c("realdonaldtrump", "biden", "esta", "m", "s", "capacitado", ...	character(0)
empty	c("realdonaldtrump", "drop", "the", "tape", "let's", "see", "it", ...	character(0)
character(0)	c("mittromney", "reveals", "he", "didn't", "vote", "for", "reald...	character(0)
shocking	c("that", "that", "is", "the", "most", "news", "everyone", "alre",...	character(0)
character(0)	c("realdonaldtrump", "obama", "is", "speaking", "now")	right
character(0)	c("go", "take", "a")	nap
c("pandemic", "deadly", "mislead")	c("realdonaldtrump", "knew", "was", "stuff", "on", "feb", "but...	character(0)

Ilustración 5.17 (Moreno, 2020)

La ilustración 5.15 muestra el contenido del data frame **tweets_trump_terminos_sentimiento**.

Este data frame debe ser procesado debido a que necesitamos solamente a las palabras y su valencia. Es por ello que utilizamos el método **attributes**() y su atributo **$counts** tal como lo indica la Ilustración 5.18 para extraer la suma de todas las palabras más su valencia.

```
41  #Obteniendo terminos
42  terminos<-attributes(tweets_trump_terminos_sentimiento)$counts #cuenta los terminos y les da su respectiva polidaridad
43  #Obteniendo solo los terminos positivos
44  terminos_positivos<-terminos[polarity>0,]
45  #Obteniendo solo los terminos negativos
46  terminos_negativos<-terminos[polarity<0,]
```

Ilustración 5.18 (Moreno, 2020)

El data frame **términos** de la Ilustración 5.19 luce de la siguiente forma una vez ejecutada la línea 42.

terminos

Filter

	words	polarity	n
1	please	1.00	13
2	care	1.00	10
3	understand	1.00	4
4	truth	1.00	3
5	aid	1.00	2
6	appropriate	1.00	2
7	comrade	1.00	2
8	cares	1.00	2
9	accomplishments	1.00	2
10	wonder	1.00	2

Ilustración 5.19 (Moreno, 2020)

El data frame de la ilustración 5.18 contiene tanto a las palabras positivas cómo a las negativas dentro. Las palabras neutras son excluidas del data frame al momento de utilizar el método **attributes** debido a que su valor es cero. Es bastante probable que las palabras neutras sean nombres propios, conectores de oración o signos de puntuación.

Con el objetivo de realizar un estudio a nivel de palabra en nuestro data frame de términos dividiremos al mismo en dos creando otros data frames, uno contendrá los términos negativos y el otro los positivos. Esta división se puede apreciar en la Ilustración 5.20.

```
43  #Obteniendo solo los terminos positivos
44  terminos_positivos<-terminos[polarity>0,]
45  #Obteniendo solo los terminos negativos
46  terminos_negativos<-terminos[polarity<0,]
47  #Revisando los terminos positivos y negativos
48  head(terminos_positivos)
49  head(terminos_negativos)
```

Ilustración 5.20 (Moreno, 2020)

Si revisamos a estos dos data frames encontremos que cada uno tiene una cantidad diferente de elementos, esto se debe a que los sentimientos no se distribuyen de una forma equitativa. Es bastante raro o poco común el encontrar textos donde la distribución de los sentimientos sea de 50/50, aquello que altera

la distribución de los sentimientos en un texto es la presencia de los transformadores de valencia.

Una vez que utilizamos el método **head()** para revisar nuestros data frames encontraremos lo siguiente:

```
> head(terminos_positivos,10)
             words polarity  n
 1:         please        1 13
 2:           care        1 10
 3:     understand        1  4
 4:          truth        1  3
 5:            aid        1  2
 6:    appropriate        1  2
 7:        comrade        1  2
 8:          cares        1  2
 9: accomplishments        1  2
10:         wonder        1  2
```

Ilustración 5.21 (Moreno, 2020)

```
> head(terminos_negativos,10)
      words polarity  n
 1:   trump    -0.10 96
 2:     pay    -0.10 15
 3: fucking    -0.15  4
 4:   empty    -0.25 10
 5:     tax    -0.25  8
 6:    wait    -0.25  6
 7:  enough    -0.25  5
 8:   black    -0.25  4
 9:    hard    -0.25  4
10:  police    -0.25  4
```

Ilustración 5.22 (Moreno, 2020)

En esta ocasión tanto en la Ilustración 5.21 como en la 5.22 se ha decidido mostrar los 10 primeros términos de cada uno de los data frames. Con esta "vista previa" podemos apreciar de una forma bastante gráfica la diferencia entre ambos data frames y el predominio de un sentimiento sobre otro. Aun así, todavía falta un poco de análisis y el gráfico final.

De ambos data frames se obtendrán los términos con más incidencias para presentarlos en un gráfico de barras y posteriormente comparar los resultados.

Los pasos para obtener los términos con más incidencias serán los siguientes:

1. Analizar ambos data frames y ordenarlos de mayor a menor en base a la columna **n.**

2. Una vez encontradas las 10 primeras incidencias crear dos nuevos data frames que guarden las mismas.

3. Utilizando la librería **ggplot2** graficar los resultados.

4. Comparar los resultados.

La ejecución de los 4 pasos se detalla a continuación en las siguientes ilustraciones:

Filter

	words	polarity	n
1	trump	-0.10	96
269	lies	-1.00	23
2	pay	-0.10	15
64	lie	-0.50	15
52	stop	-0.40	12
270	racist	-1.00	12
271	liar	-1.00	11
4	empty	-0.25	10
5	tax	-0.25	8
168	bad	-0.75	8

Ilustración 5.23 (Moreno, 2020)

La Ilustración 5.23 muestra el data frame de términos positivos ordenado de mayor a menor en base a la columna **n**.

Una vez identificadas las incidencias de cada palabra se procede a crear los dos nuevos data frames con los 10 términos más frecuentes, tal cual indica la Ilustración 5.24.

```
104  #Top terminos Positivos
105  top_terminos_positivos<-terminos_positivos[n>=10,]
106
107  #Top terminos negativos
108  top_terminos_negativos<-terminos_negativos[n>=8,]
109
110  #Obteniendo solo los 10 primeros
111  top_terminos_positivos<-top_terminos_positivos[1:10,]
112
113  top_terminos_negativos<-top_terminos_negativos[1:10,]
```

Ilustración 5.24 (Moreno, 2020)

Una vez más utilizamos la librería **ggplot2** para la visualización de los datos después de su respectivo análisis.

```
115  #GRAFICOS
116  ggplot(top_terminos_positivos,aes(x=words,y=n))+geom_col()
117  ggplot(top_terminos_positivos,aes(x=words,y=n,fill=words))+geom_col()->barras_positivo
118  #Etiquetas y estilos
119  barras_positivo+labs(title = "Presencia de Palabras Positivas",x="Palabras Positivas",y="Presencia",fill="Palabras")->barras_positivo
120  barras_positivo+theme(panel.background = element_rect(fill = "grey"))->barras_positivo
121  barras_positivo+theme(plot.title = element_text(hjust = 0.5, face = "bold"))->barras_positivo
122  #Grafico de Barras final
123  barras_positivo
```

Ilustración 5.25 (Moreno, 2020)

En las líneas 116-117 de la Ilustración 5.25 configuramos los ejes y geometría de nuestro gráfico de barras que muestre los 10 términos positivos más utilizados en los tweets de Donald Trump. Una vez que hemos establecido la configuración de nuestro gráfico guardamos dicha configuración en un objeto llamado **barras_positivo**.

Entre las líneas 119-121 configuramos las etiquetas de nuestro gráfico, el color del fondo y el título. Finalmente, en la línea 123 obtendremos el objeto final **barras_positivo** que contiene el gráfico final, si ejecutamos está línea obtendremos la Ilustración 5.26.

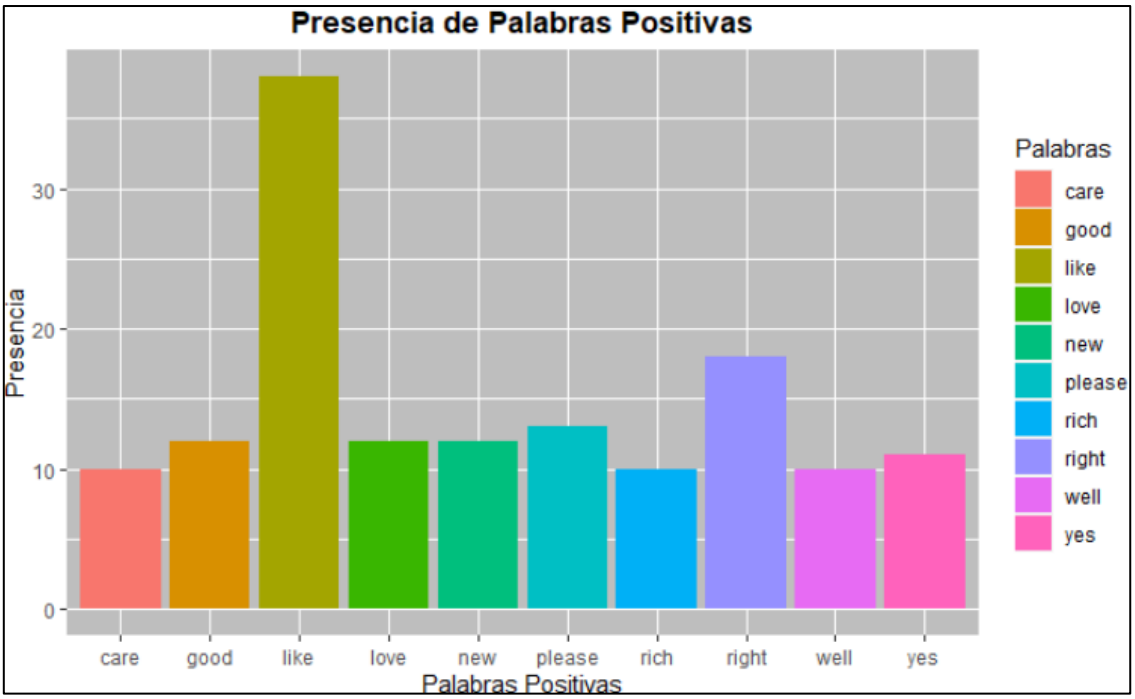


Ilustración 5.26 (Moreno, 2020)

El mismo procedimiento que se mostró en la Ilustración 5.26 se aplicará al data frame que contiene los términos negativos. Una vez que se tiene ambos gráficos listos se procederá a compararlos.

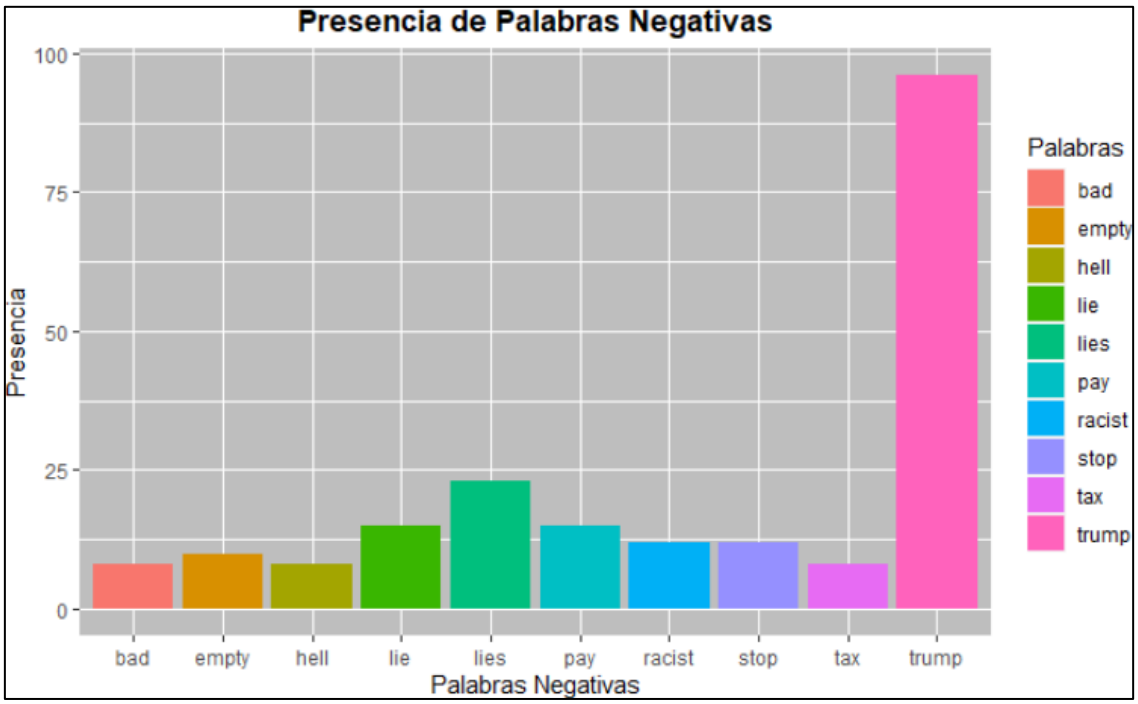


Ilustración 5.27 (Moreno, 2020)

La Ilustración 5.27 se encarga de mostrar los términos negativos encontrados dentro de los tweets de Donald Trump.

Al momento de comparar ambos gráficos podemos observar que la presencia de palabras negativas es mayor que las positivas y que el termino **Trump** es el más repetido entre ambos gráficos.

La constante aparición de la palabra **Trump** se debe a que sus tweets se escogieron para realizar este ejemplo, aunque realmente lo curioso de este asunto es que la palabra dentro del **diccionario 1** tenga una **valencia negativa de -0.10** tal cual se muestra en la ilustración 5.23. Esto probablemente se deba a que el autor del diccionario no tenía simpatía con Donald Trump o que el nombre en tiempos recientes se asocie con negatividad.

5.3 Análisis a Nivel de Emociones

El análisis de emociones es mucho más complejo que el análisis de sentimientos debido a que una misma palabra puede caer dentro de varias categorías dependiendo de su contexto (presencia de transformadores de valencia). **El valor de la emotividad de una palabra se encontrará entre 0-1,** esto quiere decir que un texto puede ser bastante emotivo como nada emotivo.

El proceso de obtener emociones es bastante similar al de sentimientos debido a que el análisis de emociones también se encuentra presente dentro la librería **sentimentr**. En este proceso también nos valdremos de las librerías **tiddyverse** y **lexicon** para la elaboración de los gráficos**.**

Aquí también usaremos un archivo XLSX que ya contenga los tweets obtenidos previamente.

```
7  tweets_trump.df<-read.xlsx("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
8                             sheetIndex=1,
9                             startRow=1,
10                            colIndex=1)
11
12 #Extracción de Emociones
13 tweets_trump.df %>%
14   get_sentences() %>%
15   emotion()->tweets_trump_emociones
16
17 #Emociones Promedio
18 tweets_trump.df %>%
19   get_sentences() %>%
20   emotion_by()->tweets_trump_emociones_promedio
21
22 #Emociones Terminos
23 tweets_trump.df %>%
24   get_sentences() %>%
25   extract_emotion_terms()->tweets_trump_terminos_emociones
```

Ilustración 5.28 (Moreno, 2020)

La Ilustración 5.28 muestra las líneas de código que se necesitan para extraer emociones a partir de tweets.

Aquí extraeremos:

- Las emociones en la línea 15.
- El promedio de emociones por documento en la línea 20.
- Los términos presentes en los documentos con su respectivo valor y emoción en la línea 25.

Cada uno de los resultados será almacenado en un data frame diferente cómo se aprecia en la siguiente tabla:

Nombre del data frame	Descripción	Método de sentimentr utilizado
tweets_trump_emociones en la línea 15.	Contiene cada una de las emociones por tweet.	emotion()
tweets_trump_emociones_promedio en la línea 20.	Contiene el valor promedio de emociones por cada tweet.	emotion_by()
tweets_trump_terminos_emociones en la línea 25.	Contiene todas las palabras que denoten algún tipo de emoción.	extract_emotion_terms

Tabla 5.5 (Moreno, 2020)

Cada data frame tendrá su respectivo tratamiento y se analizará su estructura para extraer las relaciones que se encuentran presentes entre los datos que los componen.

5.3.1 Las Emociones

Como se mencionó con anterioridad las emociones en términos cuantitativos y cualitativos tienen mayor variedad que los sentimientos. Esta "variedad" se reduce o amplía según el método (de la **librería sentimentr**) que utilicemos.

Por ejemplo los métodos **emotion()** y **emotion_by()** son capaces de clasificar a las palabras en 16 categorías de emociones diferentes, mientas que el método **extract_emotion_terms()** las clasifica en 8.

Las categorías para los dos primeros métodos se describen a continuación:

Número de Emoción	Tipo de Emoción	Traducción
1	anger	ira
2	anger_negated	ira negada
3	anticipation	anticipación
4	anticipation_negated	anticipación negada
5	disgust	disgusto
6	disgust_negated	disgusto negado
7	fear	miedo
8	fear_negated	miedo negado
9	joy	alegría
10	joy_negated	alegría negada
11	sadness	tristeza
12	sadness_negated	tristeza negada
13	surprise	sorpresa

14	surprise_negated	sorpresa negada
15	trust	confianza
16	trust_negated	confianza negada

Tabla 5.6 (Moreno, 2020)

Es probable que una sola palabra caiga en varias categorías debido a los transformadores de valencia y que el valor que se le otorgue sea diferente en cada una de ellas. Por esta razón siempre es recomendable complementar el análisis de emociones con el de sentimientos.

La tabla 5.7 nos muestra cómo funciona la clasificación de emociones utilizando el método **extract_emotion_terms()**.

Número de Emoción	Tipo de Emoción	Traducción
1	anger	ira
2	anticipation	anticipación
3	disgust	disgusto
4	fear	miedo
5	joy	alegría
6	sadness	tristeza
7	surprise	sorpresa
8	trust	confianza

Tabla 5.7 (Moreno, 2020)

El data frame que contiene los términos extraídos con el método de la tabla 5.7 califica con 0 a cualquier termino que considere neutro o que no se encuentre dentro del diccionario, esto se puede apreciar en la Ilustración 5.29.

	words	emotion_type	emotion	n
680		NA	0	1310
681	realdonaldtrump	NA	0	951
682	t	NA	0	643
683	co	NA	0	554
684	https	NA	0	554
685	the	NA	0	326
686	you	NA	0	261
687	to	NA	0	205
688	a	NA	0	196
689	is	NA	0	173

Ilustración 5.29 (Moreno, 2020)

La afirmación de que una misma palabra puede ser encasillada dentro varias categorías emocionales dependiendo de su contexto se puede apreciar en la Ilustración 5.30 sobre todo en la palabra “vote”, como vemos está se encuentra en 6 de las 8 categorías.

	words	emotion_type	emotion	n
1	trump	surprise	1	95
2	vote	anger	1	28
3	vote	anticipation	1	28
4	vote	joy	1	28
5	vote	sadness	1	28
6	vote	surprise	1	28
7	vote	trust	1	28
8	president	trust	1	23
9	don	trust	1	22
10	plan	anticipation	1	18

Ilustración 5.30 (Moreno, 2020)

El solamente echar un vistazo a los data frames nos puede aportar perspectivas y conclusiones acerca de las emociones que la gente tiene sobre Donald Trump, aun así, siempre es bueno complementar a estos análisis con gráficos debido a que el uso de colores y figuras nos ayuda a tener una mejor percepción del panorama.

Es probable que el encontrar más de una palabra en varias categorías nos haga pensar que la distribución de las emociones a lo largo de todos tweets no tenga sentido o no indique algún patrón en absoluto. Con el objetivo de responder a dicha duda se realizó el siguiente gráfico de dispersión:

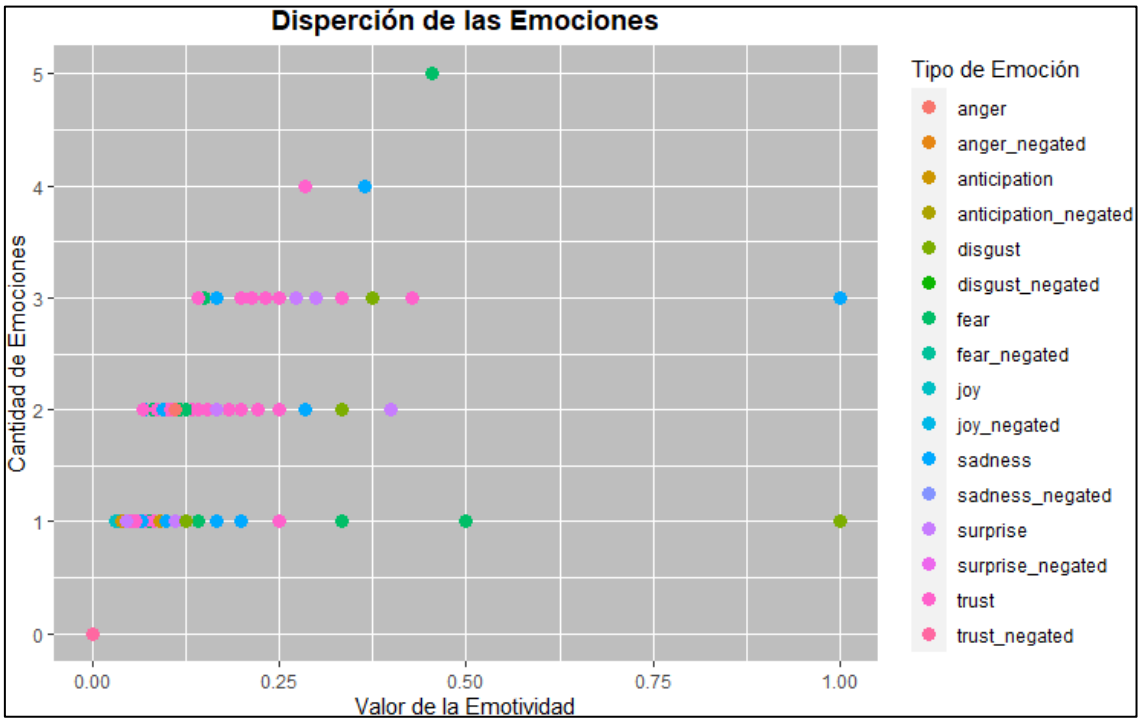


Ilustración 5.31 (Moreno, 2020)

Revisando la Ilustración 5.31 nos podemos dar cuenta que los tweets acerca de Donald Trump son poco emotivos en su mayoría, pero que cuando estos presentan el mayor grado de emotividad (valor de 1) son emociones “negativas” (tristeza y disgusto).

Lo dicho con anterioridad concuerda y coincide con los resultados obtenidos dentro del análisis de sentimientos realizado previamente. Aun así, solo estamos viendo una parte dentro de todo el panorama. Para complementar el panorama podemos revisar cuales son las emociones más predominantes dentro de los tweets de Donald Trump. A partir de lo visto en la ilustración 5.31 se podría llegar a concluir que la emoción más presente dentro de los tweets son el “disgusto”, la “tristeza” y el “miedo”. Esta perspectiva podría llegar a cambiar mirando la Ilustración 5.32.

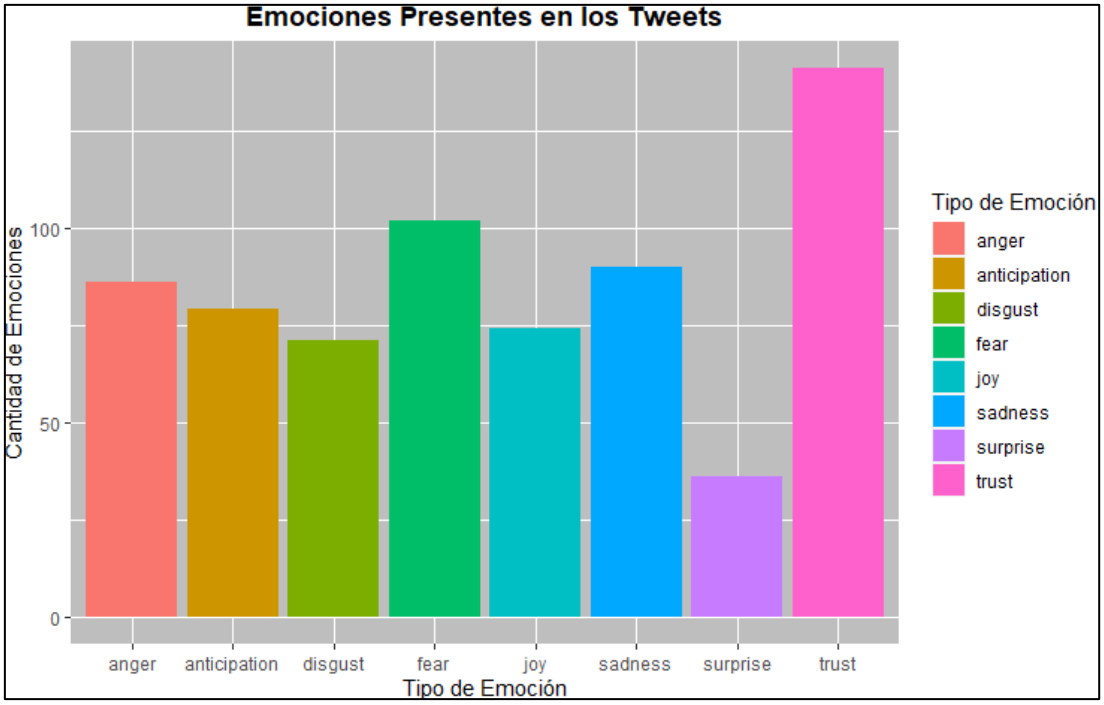


Ilustración 5.32 (Moreno, 2020)

La Ilustración 5.32 podría llegar contradecir al análisis de sentimientos y a la perspectiva inicial que tuvimos apreciando la ilustración 5.31 debido a que parecería indicar que los tweets que se hacen sobre Donald Trump inspiran “confianza”. Este “desacuerdo” se puede llegar a explicar comparando las Ilustraciones 5.31 y 5.32.

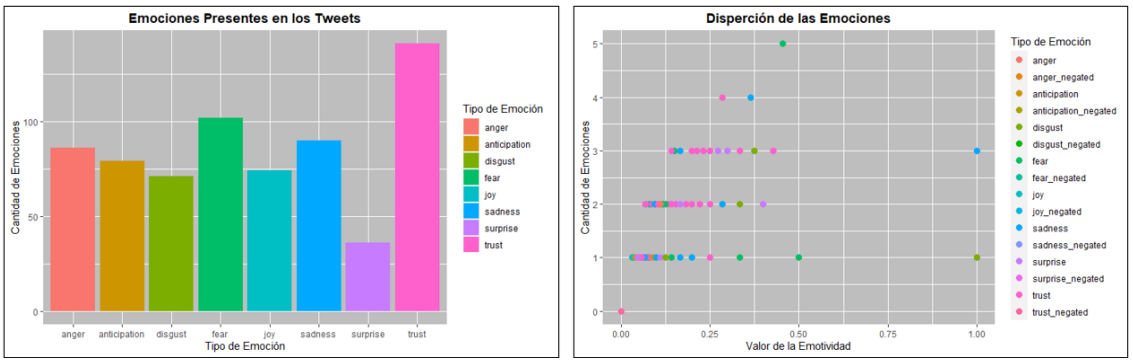


Ilustración 5.33 (Moreno, 2020)

La presencia dominante de la emoción “confianza” se puede explicar revisando la Ilustración 5.31 en la cual podemos apreciar que los términos cargados con está emoción son una mayoría, aun así, son poco emotivos (bastante cerca al 0.4). Las dos siguientes emociones más presentes son “miedo” y “tristeza”. Esto también se puede comprobar observando la ilustración 5.31. A diferencia de la “confianza” estas dos emociones tienen mayor valor emocional en el documento 0.50 para el miedo y 1 para la tristeza.

Lo visto en la Ilustración 5.33 se complementa con el análisis de sentimientos y nos muestra una perspectiva general bastante clara sobre Donald Trump hasta el momento.

5.4 Aplicando Porter y la Lematización

A partir del algoritmo de Porter obtuvimos las nubes de palabras (análisis a nivel de palabras) y en base a la lematización tuvimos un análisis bastante exhaustivo (a nivel de documento, oración y polaridad). Estas dos técnicas por sí solas nos han permitido realizar todas las aproximaciones del análisis de sentimientos.

En este capítulo se aplicarán ambas técnicas con el objetivo de conocer que otro tipo de resultados podemos obtener combinando las técnicas y aproximaciones de ambos métodos.

Debido a que ambos procesos son bastante complejos y aplicarlos a todos los candidatos nos podría tomar más tiempo del que se dispone solo se realizará a Donald Trump la combinación de Porter-Lematización.

Uno de los principales problemas de Porter que ya se discutió con anterioridad fue la inconsistencia de algunos resultados que arroja al momento de ejecutarse el algoritmo produciendo palabras que no tienen significado (más frecuente en lenguajes ajenos al inglés). Esto obliga al investigador a “crear” un arreglo que remueva las palabras inconsistentes dentro de la construcción de la nube de palabras como fue durante la presente investigación. Este problema se puede llegar a solucionar aplicando la Lematización.

El proceso de combinar ambos métodos se puede apreciar en la siguiente ilustración:

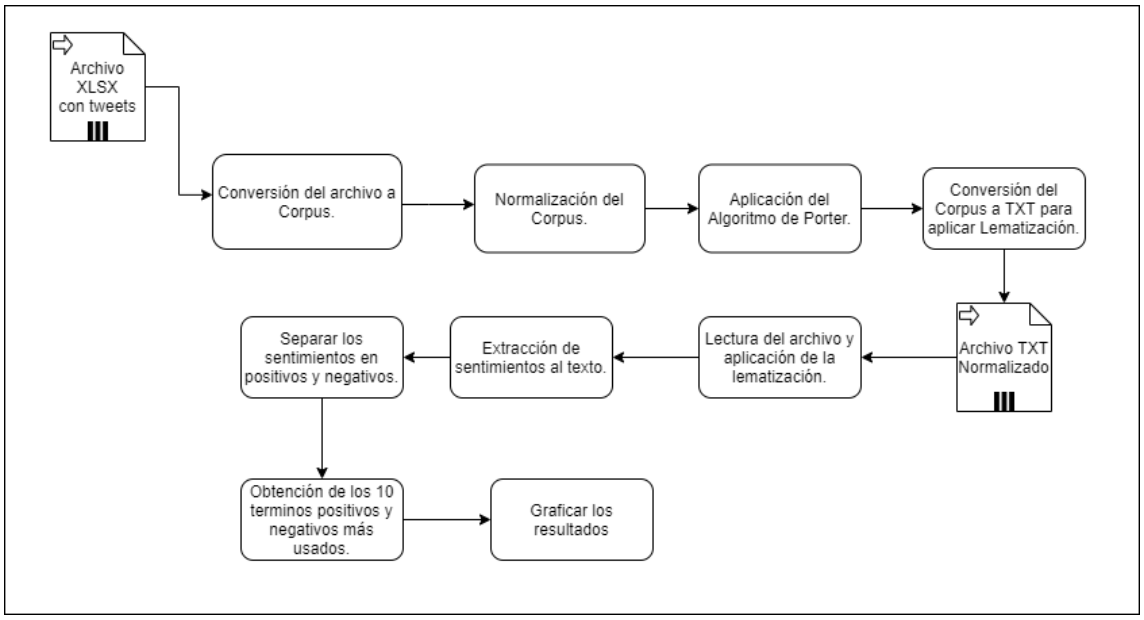


Ilustración 5.34 (Moreno, 2020)

La realización del proceso descrito en la Ilustración 5.34 debería omitir la construcción de un arreglo que contenga aquellos términos que no aportan nada a la investigación. Otra ventaja de combinar ambas técnicas es el obtener un documento normalizado en la etapa de lematización. Esto reduce el tiempo de clasificación de términos y reduce la búsqueda debido a que algunas palabras de parada han sido eliminadas junto con signos de puntuación y números.

El proceso de normalización y de lematización es el mismo con el que se ha venido trabajando en los otros scripts, esto también aplica para las librerías.

```
1  library(xlsx)#para leer archivos xlsx
2  library(sentimentr)#Paquete completo para análisis de sentimientos
3  library(tidyverse) #Herramientas para Data Science
4  library(lexicon) #Lexicons disponibles
5  library(tm) #mineria de texto
6  library(SnowballC) #algoritmo de porter
```

Ilustración 5.35 (Moreno, 2020)

La ilustración 5.35 muestra la carga de las librerías necesarias para aplicar tanto Porter cómo Lematización.

Una vez cargadas nuestras librerías procedemos a leer el archivo XLSX donde se encuentran almacenados los tweets de Donald Trump.

```
tweets_trump.df<-read.xlsx("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
                           sheetIndex=1,
                           startRow=1,
                           colIndex=1)
```

Ilustración 5.36 (Moreno, 2020)

La Ilustración 5.36 muestra la lectura del archivo xlsx que contiene los tweets de Donald Trump.

El siguiente paso es convertir en un corpus la columna texto del data frame y normalizarlo.

```
#Transformacion a Corpus Normalizacion
myCorpus<-Corpus(VectorSource(tweets_trump.df$text))

#Inspeccionando el corpus
inspect(myCorpus)

#Normalizando el Corpus
#limpieza y normalización del contenido del corpus
corpus_limpio<-tm_map(myCorpus,tolower)#convierte todo el texto en minúsculas
#removemos urls innecesarias dentrol del corpus
remove_url<- function(x) gsub("http[^[:space:]]*","",x)
corpus_limpio<- tm_map(corpus_limpio,content_transformer(remove_url))
#removiendo los número dentro del corpus
corpus_limpio<-tm_map(corpus_limpio,removeNumbers)
#removiendo signos de puntuación
corpus_limpio<-tm_map(corpus_limpio,removePunctuation)
#removiendo emojis
removeEmojo<-function(x)gsub("[^\x01-\x7F]","",x)
corpus_limpio<-tm_map(corpus_limpio,content_transformer(removeEmojo))
#removiendo palabras de parada
corpus_limpio<-tm_map(corpus_limpio,removeWords,stopwords(kind = "en"))
#encontrnado las raíces de las palabras
corpus_limpio<-tm_map(corpus_limpio,stemDocument)#se utiliza el algoritmo de Porter
```

Ilustración 5.37 (Moreno, 2020)

El proceso de normalización ocurre desde la línea 22 hasta la línea 34 de la Ilustración 5.37 y se aplica el algoritmo de Porter en la línea 36.

Debido a que el método **get_sentences()** no extrae los sentimientos a partir de un corpus debemos transformar a este en un archivo de texto para luego leerlo (revisar la Ilustración 5.38).

```
#Transformando el Corpus a texto
writeLines(as.character(corpus_limpio), con="C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\transformacion_corpus\\textoTrump.txt")

#Importando el corpus para que se lea como texto
texto<-readLines("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R Sentimientos\\transformacion_corpus\\textoTrump.txt")
```

Ilustración 5.38 (Moreno, 2020)

Después de almacenar el archivo TXT en la variable texto procedemos a extraer los sentimientos de dicha variable en la Ilustración 5.39

```
#Obteniendo sentimientos y aplicando lematizacion
texto%>%
  get_sentences() %>%
  #sentiment(polarity_dt=lexicon::hash_sentiment_sentiword)->tweets_trump_sentimiento #obtiene el sentimiento de manera individual
  sentiment()->tweets_trump_sentimiento

#Obteniendo la polaridad de los sentimientos, positivo negativo, neutral
texto%>%
  get_sentences() %>%
  extract_sentiment_terms()->tweets_trump_terminos_sentimiento
```

Ilustración 5.39 (Moreno, 2020)

En la línea 48 creamos **el data frame tweets_trump_sentimiento** en donde almacenamos a los tweets y su sentimiento. En el caso de la línea 53 almacenamos a los documentos junto con sus términos positivos, negativos y neutrales en **el data frame tweets_trump_trump_terminos_sentimiento**.

Nuestro siguiente paso será crear un data frame que solo contenga los términos positivos, negativos y neutrales con sus respectivas valencias. A este data frame lo vamos a separar en dos (términos positivos y negativos).

```
55  #Summary
56  terminos<-attributes(tweets_trump_terminos_sentimiento)$counts
57
58  #Obteniendo terminos positivos y negativos
59
60  #Obteniendo solo los terminos positivos
61  terminos_positivos<-terminos[polarity>0,]
62  #Obteniendo solo los terminos negativos
63  terminos_negativos<-terminos[polarity<0,]
```

Ilustración 5.40 (Moreno, 2020)

La Ilustración 5.40 muestra la separación de los términos según su polaridad.

Una vez que tenemos separados nuestros términos buscamos solo los 10 más repetidos o usados dentro de los tweets, para ello revisamos los data frames y ordenamos a ambos de mayor a menos según la columna n.

Untitled1* | terminos_positivos

Filter

	words	polarity	n
73	like	0.50	42
17	right	0.80	18
74	thank	0.50	16
29	love	0.75	15
129	deal	0.25	13
1	care	1.00	12
18	new	0.80	12
30	good	0.75	12
19	well	0.80	11
20	yes	0.80	11

Ilustración 5.41 (Moreno, 2020)

La Ilustración 5.41 muestra los términos positivos más repetidos dentro de los tweets de Donald Trump.

Una vez identificados los 10 primeros términos de cada data frame pasamos a crear dos nuevos data frames que los almacenarán.

```
#Obteniendo terminos positivos y negativos

#Obteniendo solo los terminos positivos
terminos_positivos<-terminos[polarity>0,]
#Obteniendo solo los terminos negativos
terminos_negativos<-terminos[polarity<0,]

#Revisando los data frames
head(terminos_positivos)
head(terminos_negativos)

#Obteniendo el top 10
#Top terminos Positivos
top_terminos_positivos<-terminos_positivos[n>=11,]

#Top terminos negativos
top_terminos_negativos<-terminos_negativos[n>=10,]

#Obteniendo solo los 10 primeros
top_terminos_positivos<-top_terminos_positivos[1:10,]

top_terminos_negativos<-top_terminos_negativos[1:10,]
```

Ilustración 5.42 (Moreno, 2020)

La Ilustración 5.42 muestra la creación de los data frames que contendrán a los 10 términos positivos y negativos más utilizados.

Ahora que todos los datos ya se encuentran separados, filtrados y procesados podemos pasar realizar los gráficos correspondientes e identificar si existe algún cambio o diferencia al combinar ambas técnicas.

Estos resultados serán mostrados y comparados en el capítulo 6 el cual está dedicado a la discusión de los resultados obtenidos.

Capítulo 6 Resultados

Durante este capítulo se mostrarán los resultados en forma de nubes de palabras para inglés y español obtenidos a partir del algoritmo de Porter. También se revisarán los resultados obtenidos a partir del análisis por lematización realizados a nivel de sentimientos y emociones solo para el idioma inglés. Finalmente se analizan los resultados obtenidos de combinar el algoritmo de Porter con la Lematización.

6.1 Despliegue de Resultados

6.1.1 Resultados Finales de las Nubes de Palabras en inglés

Las nubes de palabras se caracterizan por ser útiles al momento de resumir y enfatizar la frecuencia de las palabras que se encuentran dentro de un texto. Esta retroalimentación nos permite tener un vistazo rápido acerca del panorama que estamos estudiando sobre todo si este se encuentra compuesto de cientos o miles de datos.

Los resultados obtenidos buscan que los datos procesados y recuperados de Twitter se muestren de una manera bastante sencilla de entender y comprender.

La primera nube que vamos a visualizar corresponde a Donald Trump:

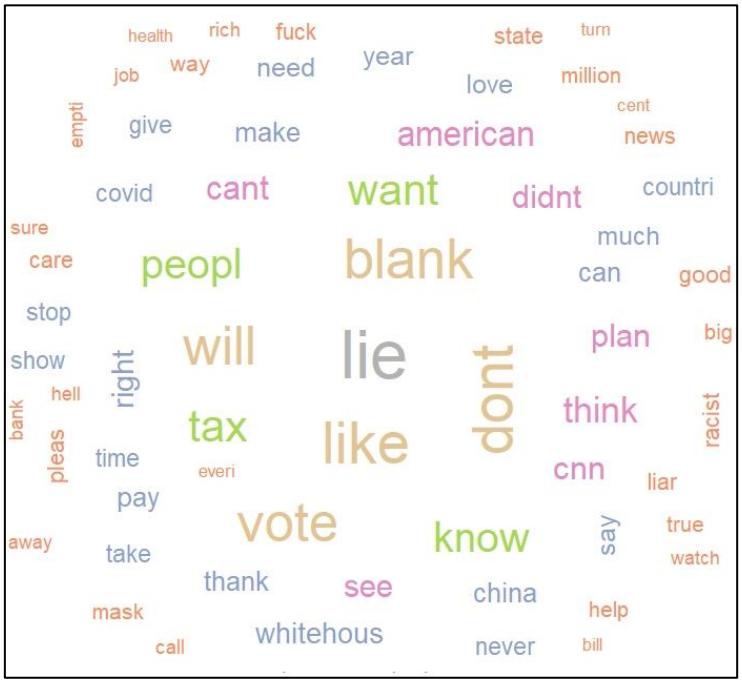


Ilustración 6.1 (Moreno, 2020)

Las mentiras durante su campaña presidencial, la supuesta evasión de impuestos en su periodo, el covid-19 y los problemas con China han sido los temas que han aquejado y mermado la popularidad de Donald Trump en tiempos recientes. Dichos temas los podemos ver plasmados en la Ilustración 6.1 y se encargan de mostrar el descontento del pueblo americano hacia el actual presidente.

El candidato Donald Trump que llegó a la casa blanca en el año 2017 como presidente de los Estados Unidos revolucionando los medios y la política norteamericana, debido a su poca o nula intervención en asuntos relacionados

con la política del país. Amado por aparentemente pocos y odiado por muchos el candidato no era más que una figura del entretenimiento y de los negocios en Estados Unidos que incursiono en la política de manera sorpresiva. Aquello le paso factura y los escándalos por acoso, manipulación, extorción y abuso de poder en su etapa como celebridad de los negocios y presidente han acabado con su imagen y lo ponen en riesgo para las nuevas elecciones.

La nube correspondiente a Joe Biden se muestra en la Ilustración 6.2:

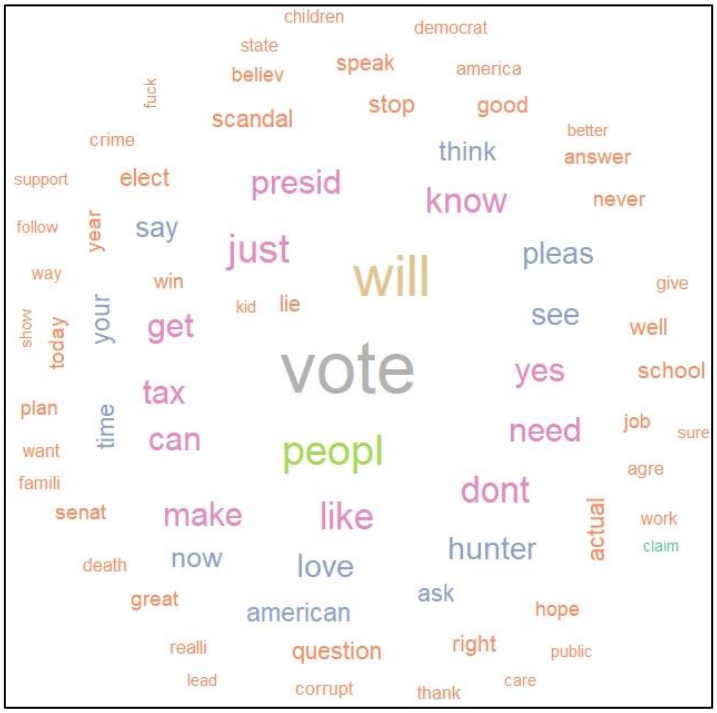


Ilustración 6.2 (Moreno, 2020)

El candidato Joe Biden exvicepresidente de la administración Obama en Estados Unidos se perfila como el cambio, la esperanza y la verdadera solución a los problemas que la administración Trump ha dejado en este tiempo. Con un discurso enfocado en los problemas del Estados Unidos moderno y con la promesa de mejorar los sistemas de salud, la educación y combatir la pandemia parece ser la esperanza del pueblo americano.

La preparación académica y experiencia de Joe Biden en la política americana lo ha llevado a compartir escenario con figuras políticas de renombre como fueron Ronald Reagan, Jimmy Carter y en tiempos más recientes Barack Obama. La inclusión social, el cambio climático, la salud y la educación más una dura historia de superación personal muestran a Joe Biden como la antítesis de Trump en un Estados Unidos conflictivo que necesita un verdadero cambio.

Ambas ilustraciones son capaces de mostrarnos las diferentes perspectivas que se tienen de ambos candidatos por separado. Esto por si solo es capaz de brindarnos conclusiones.

Si combinamos ambas nubes de palabras en un sola podremos identificar que tienen en común ambos candidatos y que tan polarizados se encuentran los sentimientos hacía cada uno de ellos.

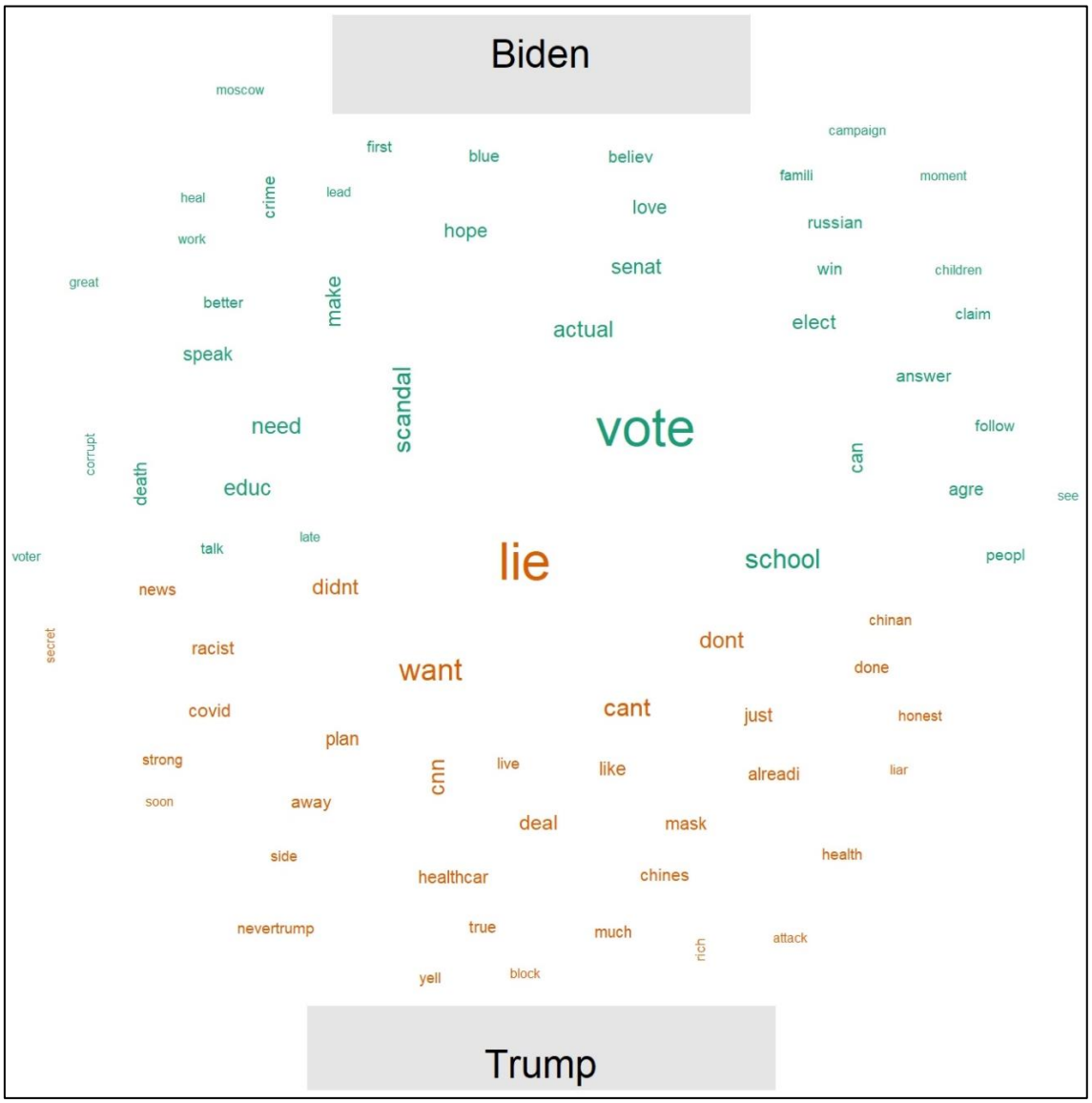


Ilustración 6.3 (Moreno, 2020)

La Ilustración 6.3 es bastante clara en los puntos de vista que se tiene sobre ambos candidatos a la presidencia de Estados Unidos. Por un lado, Joe Biden se alza como un candidato carismático e ideal para miles de americanos que están cansados de las constantes negligencias y escándalos que envuelven al actual presidente Donald Trump. En contraste si analizamos al actual presidente podemos ver que el descontento, las mentiras, el racismo y los problemas evidenciados en el sistema de salud americano parecen mermar su última esperanza de reelección.

6.1.2 Resultados finales de las nubes de palabras en español

Las ilustraciones que se presentan a continuación corresponden a la construcción de las nubes de palabras que contienen términos en español. Estas nubes buscan responder las cuestiones del capítulo 3.3.3.

El proceso de elaboración de estas nubes es similar a las anteriores con la excepción de que aquí no se llegó a utilizar el algoritmo de Porter. Esto se debió a que el algoritmo presentaba demasiadas inconsistencias en los resultados finales, a pesar de que la documentación de este indica que soporta al español como uno de sus lenguajes.

El resultado de la Ilustración 6.4 que corresponde al candidato Guillermo Lasso nos muestra lo siguiente:

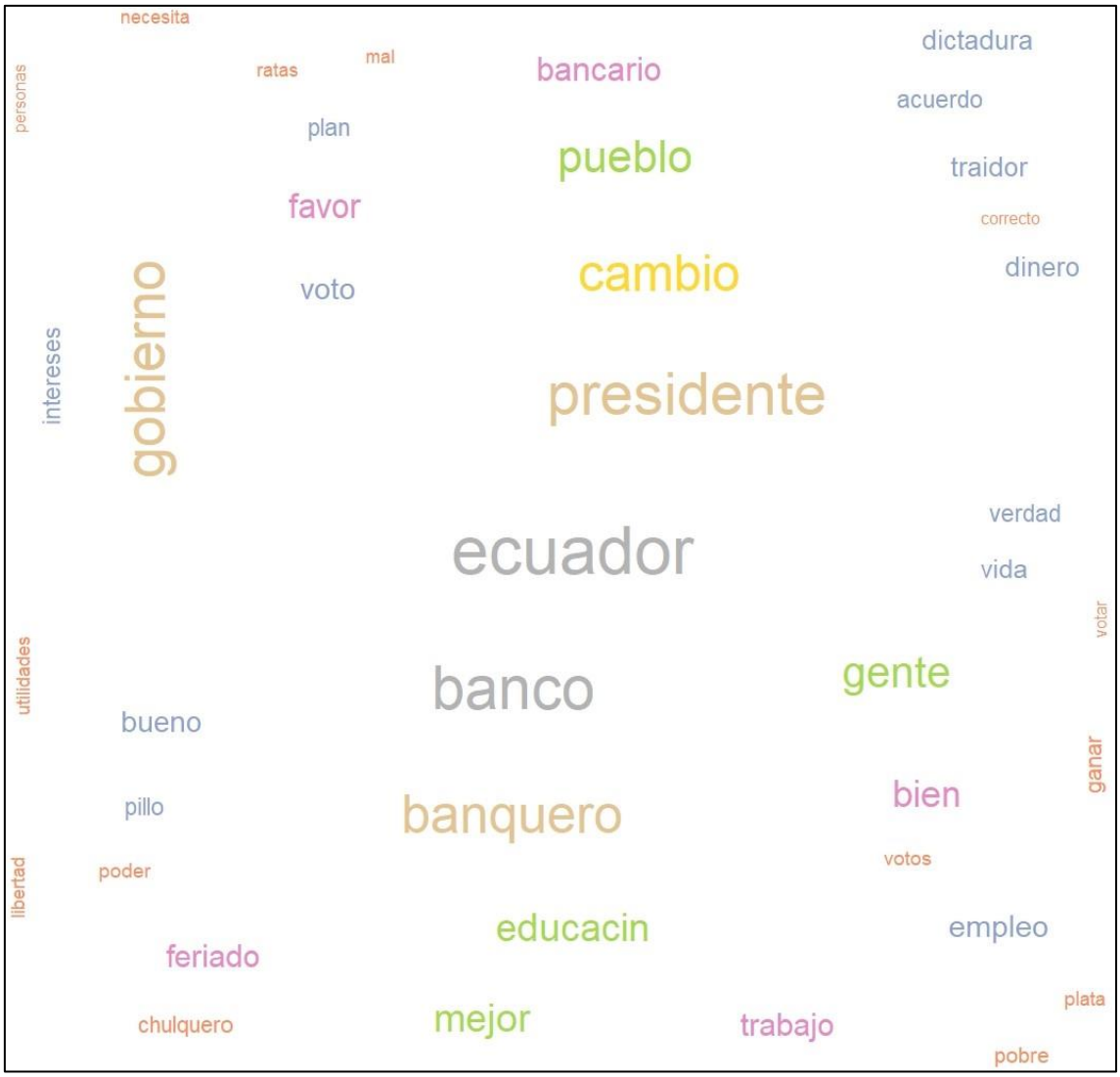


Ilustración 6.4 (Moreno, 2020)

El candidato Guillermo Lasso se presenta como el favorito entre la clase media y alta de la sociedad ecuatoriana ya que ofrece componer la economía del país y generar empleo. Aun así, su pasado donde obtuvo un supuesto beneficio (con el cual se supone fundo el Banco de Guayaquil en un futuro) durante el feriado bancario que atravesó el país durante el inicio del nuevo milenio lo persigue

hasta la actualidad. El haber sido el director ejecutivo del Banco de Guayaquil y ser su mayor accionista actual le ha generado rechazo de las clases más bajas del país. Debido a esto se le ha llamado con sobrenombres como “chulquero”.

A pesar de tener ciertas palabras que evocan rechazo o negatividad algunas de las propuestas que tiene el candidato se hacen evidentes y son representadas por las palabras “trabajo”, “empleo”, “libertad” y “cambio”.

La nube que corresponde a la Ilustración 6.5 pertenece al candidato Yaku Pérez:

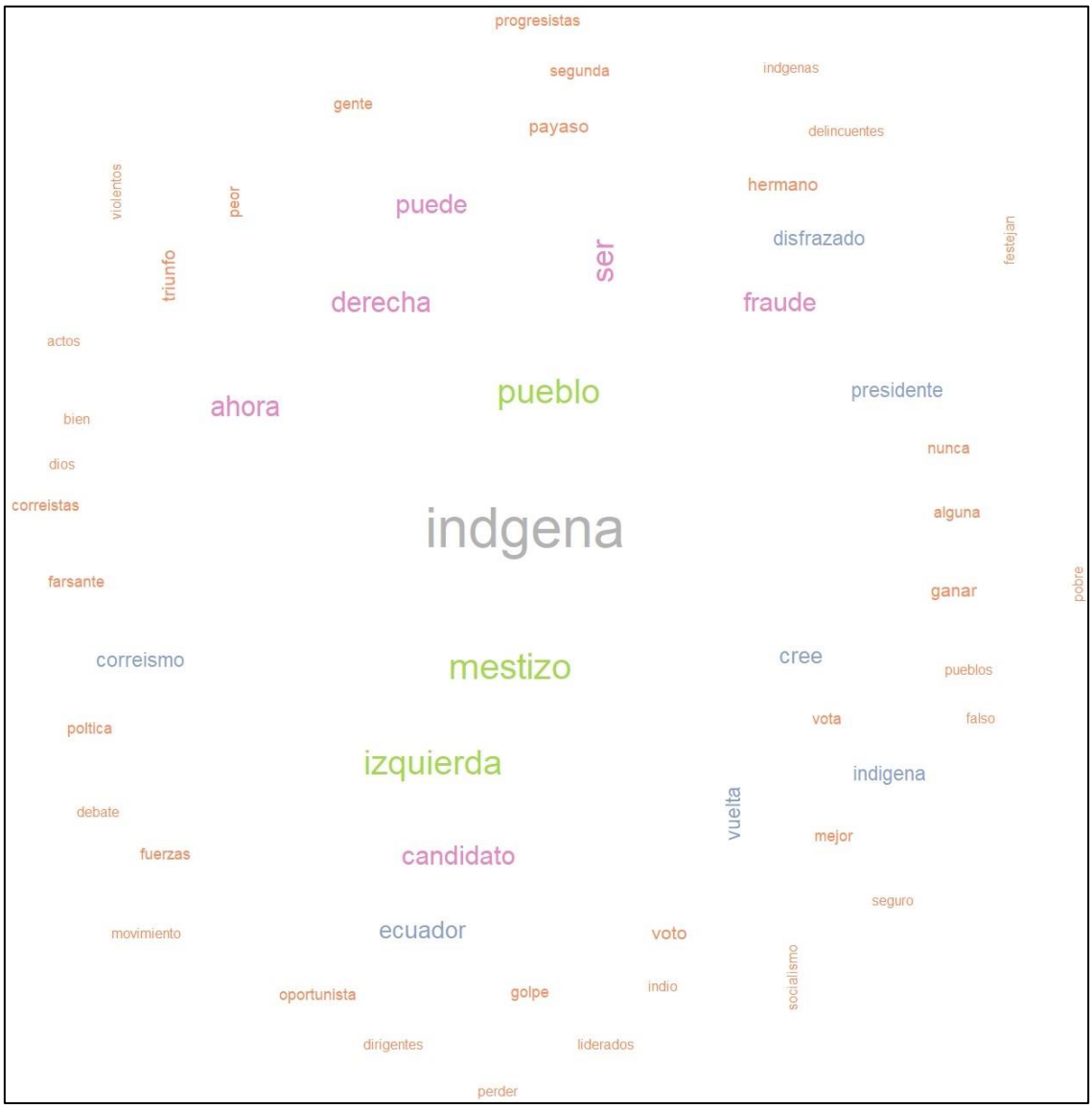


Ilustración 6.5 (Moreno, 2020)

El candidato Yaku Pérez se caracteriza por tener bastante simpatía con el “pueblo llano”, el indigenado y la población de la provincia del Azuay en general debido a que fue prefecto de esta. Su campaña respaldada por el indigenado, ambientalistas, alguna fracción de la clase media y las clases más bajas puede mostrar alguna sorpresa durante las elecciones. Aun así, algunos votantes lo rechazan por su asociación y respaldo del indigenado debido al paro nacional que provocaron los mismos durante octubre del año 2019.

Los términos cómo “indígena”, “pueblo” y “mestizo” parecen indicar la simpatía de dichas etnias con el candidato, aun así, se puede notar el amplio rechazo que se tiene por el mismo en ciertos sectores de la población cuando vemos términos cómo payaso, delincuentes (seguramente refiriéndose al paro nacional) y farsante.

El panorama global de ambos candidatos comparados se puede apreciar en la siguiente ilustración:

Ilustración 6.6 (Moreno, 2020)

La polarización de términos y opiniones es evidente mirando a la Ilustración 6.6 debido a que prácticamente ninguna llega a coincidir o relacionarse entre sí.

En conclusión, la nube comparativa solo nos muestra lo opuesto que ambos candidatos se encuentran, y la alta impopularidad que estos tienen en los distintos sectores de la población del Ecuador. La elección del uno o del otro solo dependerá de la conciencia y atención que el pueblo ecuatoriano ponga a las

propuestas de estos candidatos (las cuales no se hacen muy evidentes dentro de las nubes de palabras).

6.2 Resultados del Análisis de Sentimientos

El análisis de sentimientos utilizando lematización fue mucho más efectivo y rápido en algunas situaciones. A partir de esta técnica podemos derivar conclusiones a nivel de palabra, oración y documento.

6.2.1 Resultados del Análisis de Sentimiento a Trump

Una vez aplicados los métodos para extraer sentimientos de la librería **sentimentr** y haber divido los mismos en data frames se construyeron conclusiones que a continuación se presentan cómo gráficos estadísticos.

En el capítulo 6.1.1 se observó que la mayoría de las palabras dentro de la nube de Trump presentan una carga negativa. Su nube de palabras al igual que la de Joe Biden fueron depuradas con totalidad neutralidad, esto con el objetivo de retratar la realidad. Aun así, un análisis por lematización es el complemento perfecto para lo visto con anterioridad. El proceso de lematización permitió construir gráficos y conclusiones más elaboradas.

La Ilustración 6.7 nos muestra la distribución que siguen los sentimientos de los tweets acerca de Donald Trump.

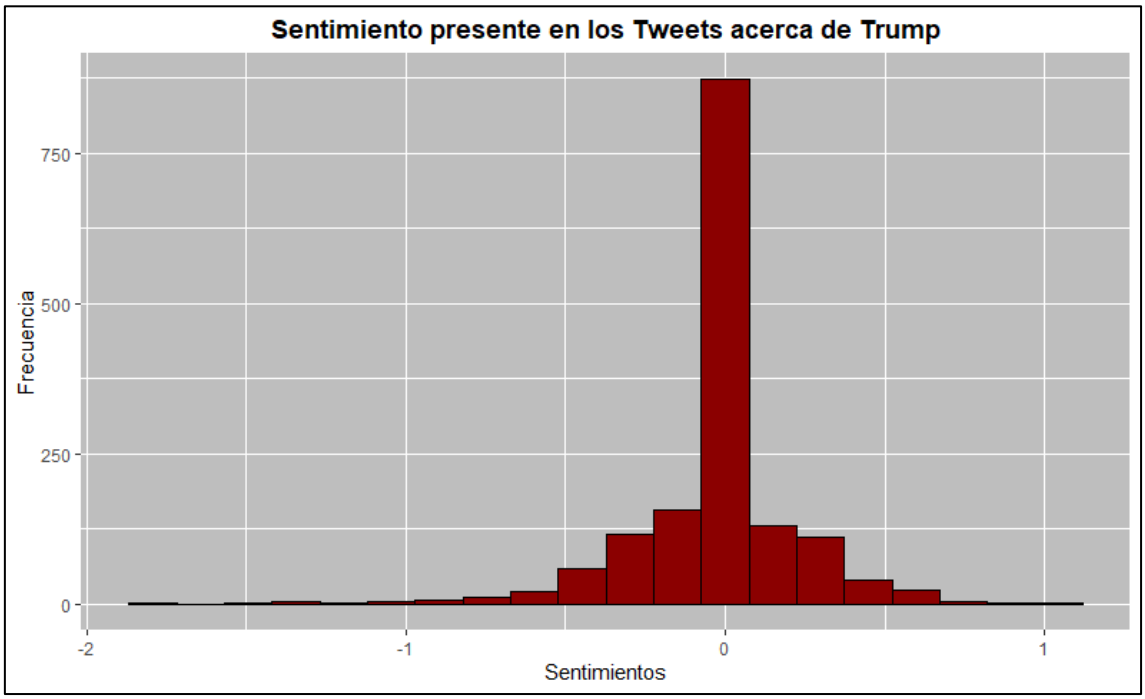


Ilustración 6.7 (Moreno, 2020)

La presencia de los modificadores de valencia, aumentan o disminuyen la carga de una palabra, esto explica que la distribución de los sentimientos del lado negativo llegue hasta casi -2.

La tendencia negativa de los documentos (tweets) también se puede apreciar en el siguiente gráfico.

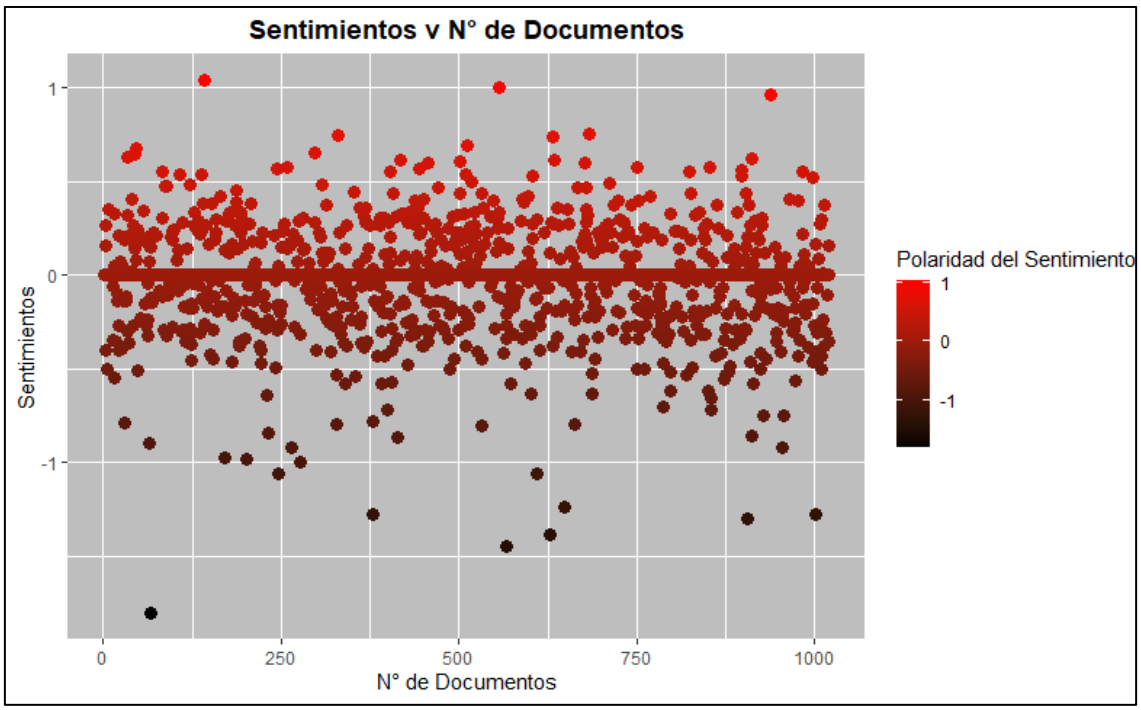


Ilustración 6.8 (Moreno, 2020)

En el gráfico de dispersión (Ilustración 6.8) podemos apreciar que la mayoría de los documentos tienden a ser negativos. La cantidad de documentos **extremadamente positivos son solo 3**, mientas que aquellos que apuntan a ser extremadamente negativos son más de 10. Esto se confirma obteniendo cuales son las palabras positivas y negativas más usadas.

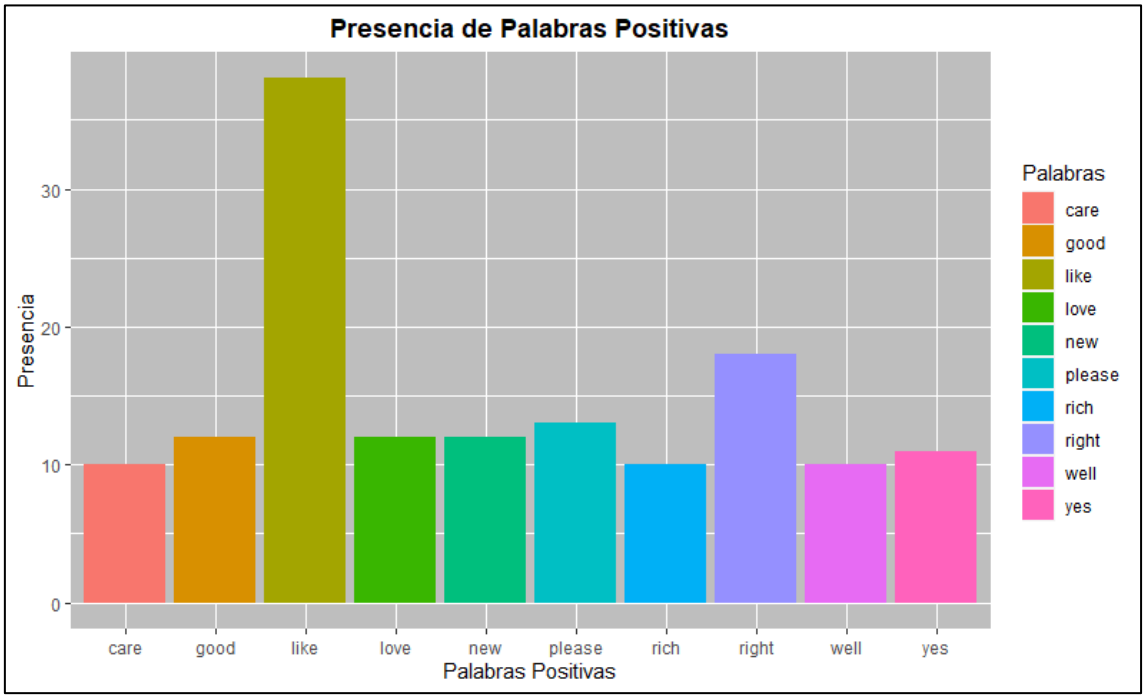


Ilustración 6.9(Moreno, 2020)

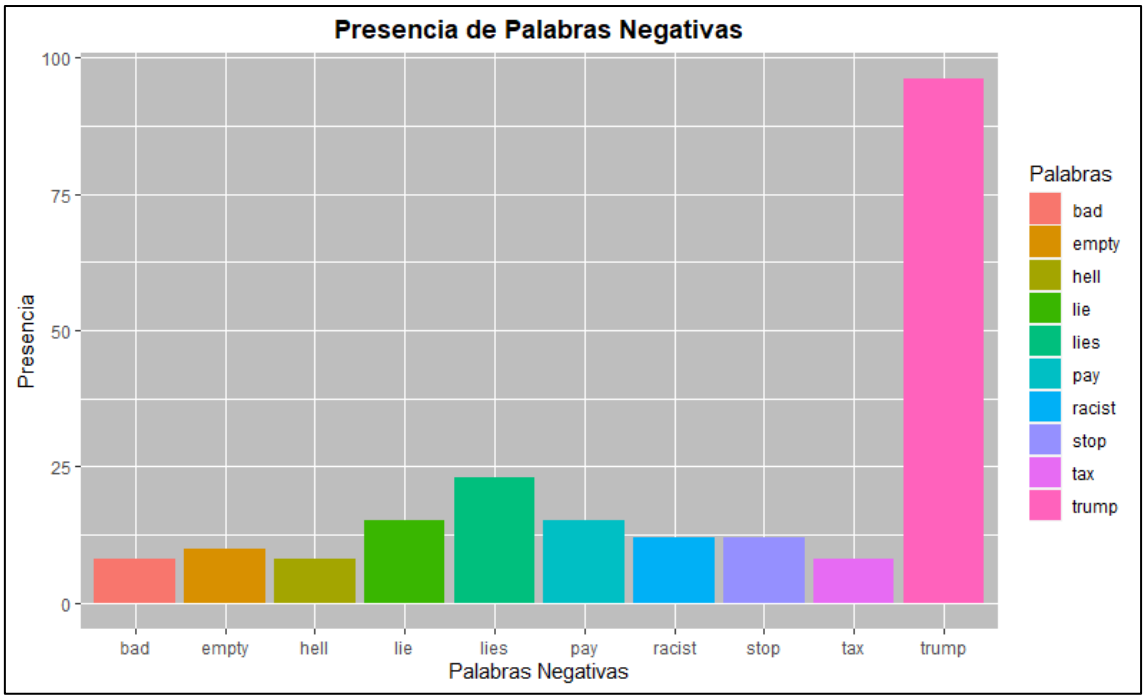


Ilustración 6.10 (Moreno, 2020)

Ambas Ilustraciones (6.9 y 6.10) nos muestran las 10 palabras positivas y negativas más usadas dentro de todos los tweets. De ambas ilustraciones podemos concluir que las palabras más usadas en los tweets son negativas y que la más presente entre ellas es “Trump”.

6.2.2 Resultados del Análisis de Sentimientos Joe Biden

Los resultados del análisis de sentimiento a Joe Biden nos ayudaran a confirmar las perspectivas ya obtenidas a partir de las nubes de palabras.

La siguiente ilustración nos muestra la distribución de los sentimientos dentro de los tweets referentes a Joe Biden:

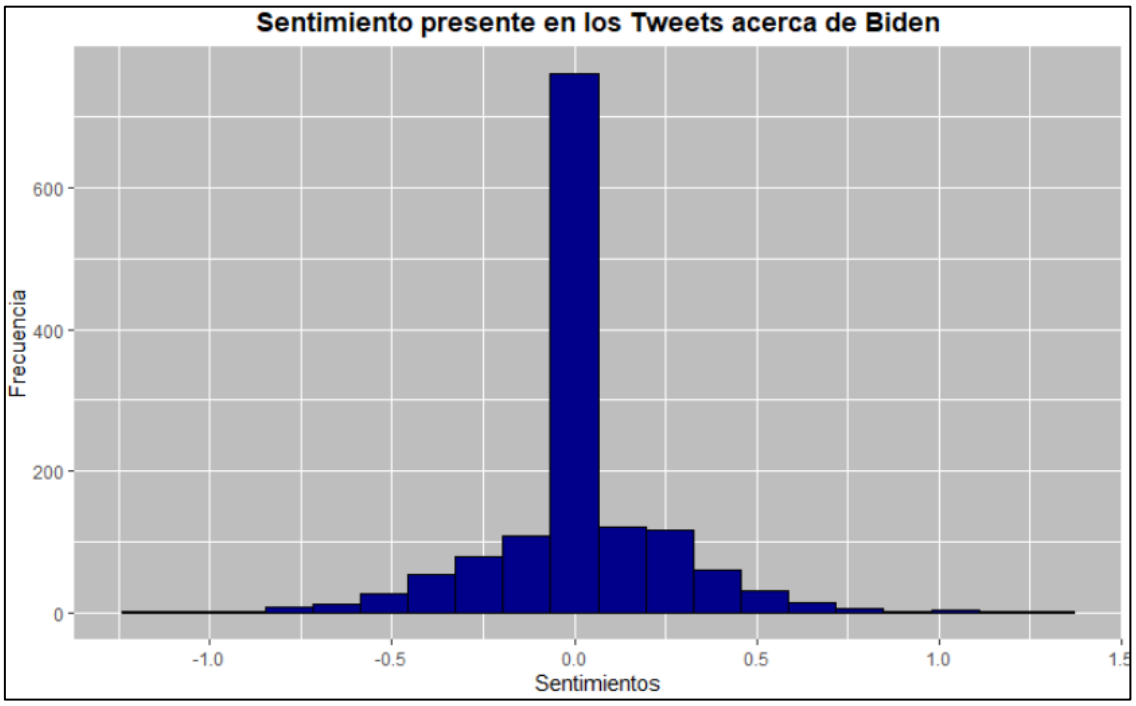


Ilustración 6.11 (Moreno, 2020)

La Ilustración 6.11 muestra la distribución de los sentimientos presentes en los tweets de Joe Biden.

Aparentemente la distribución de los sentimientos en los tweets de Joe Biden es simétrica, pero si nos fijamos detenidamente es posible darnos cuenta de que el gráfico tiene sesgo positivo y se acerca bastante al valor de 1.5. Esto indica que la mayor parte de los sentimientos son positivos dentro de los tweets acerca del candidato.

La aparente simetría se debe a que los sentimientos negativos están cerca de un -1.25 aproximadamente y que la cantidad de términos negativos y positivos parecen ser los mismos entre -0.5 y 0.5. La mejor forma de complementar está aproximación inicial es echa un vistazo al gráfico de dispersión que se encuentra en la Ilustración 6.12.

En la siguiente ilustración se revisará que tan dispersos se encuentran los sentimientos presentes en los tweets acerca de Joe Biden:

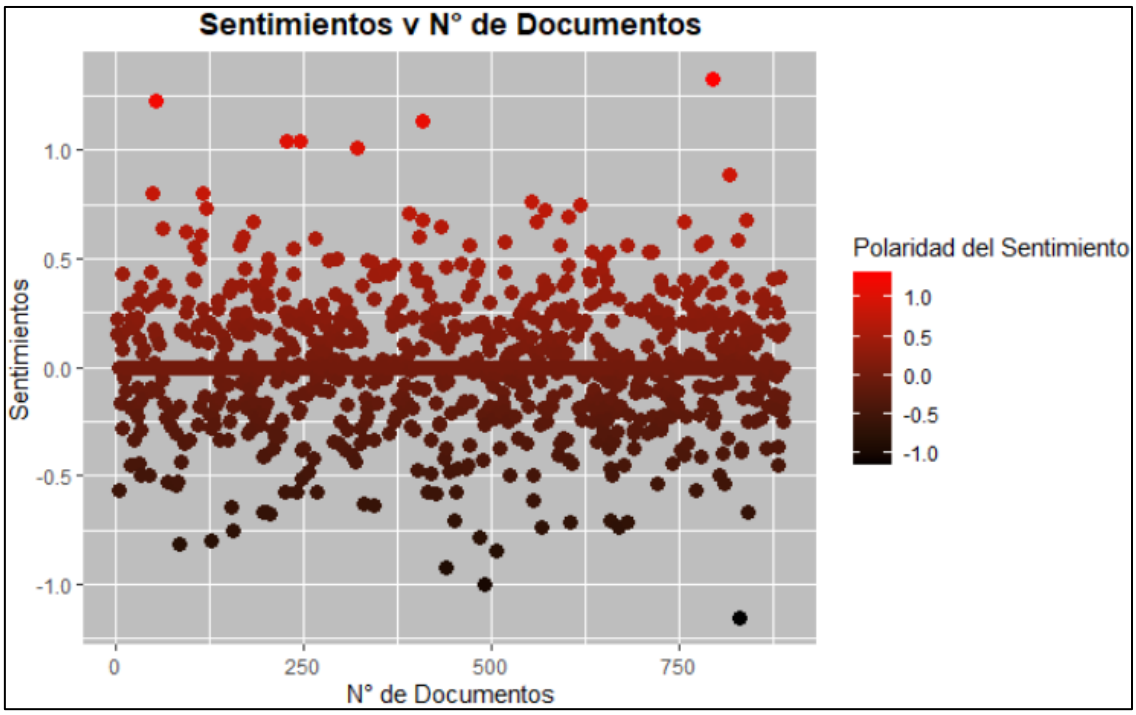


Ilustración 6.12 (Moreno, 2020)

La mayoría de los sentimientos se encuentran entre -0.5 y 0.5 tal cual se apreció en la Ilustración anterior (6.11), pero la cantidad de sentimientos positivos (entre 0.5-1) es mayor a la cantidad de sentimientos negativos (entre -0.5 y -1) dentro de los documentos.

La cantidad de términos “altamente positivos” (seis en total) es mayor a la cantidad de “altamente negativos” (solamente dos). Esto parece indicar que la tendencia general de los sentimientos es positiva cuando se trata de Joe Biden. Sin embargo, la distribución casi equitativa de los sentimientos no le da una victoria contundente frente a Trump en esté análisis.

Las siguientes ilustraciones buscan generar un panorama completo para obtener una conclusión.

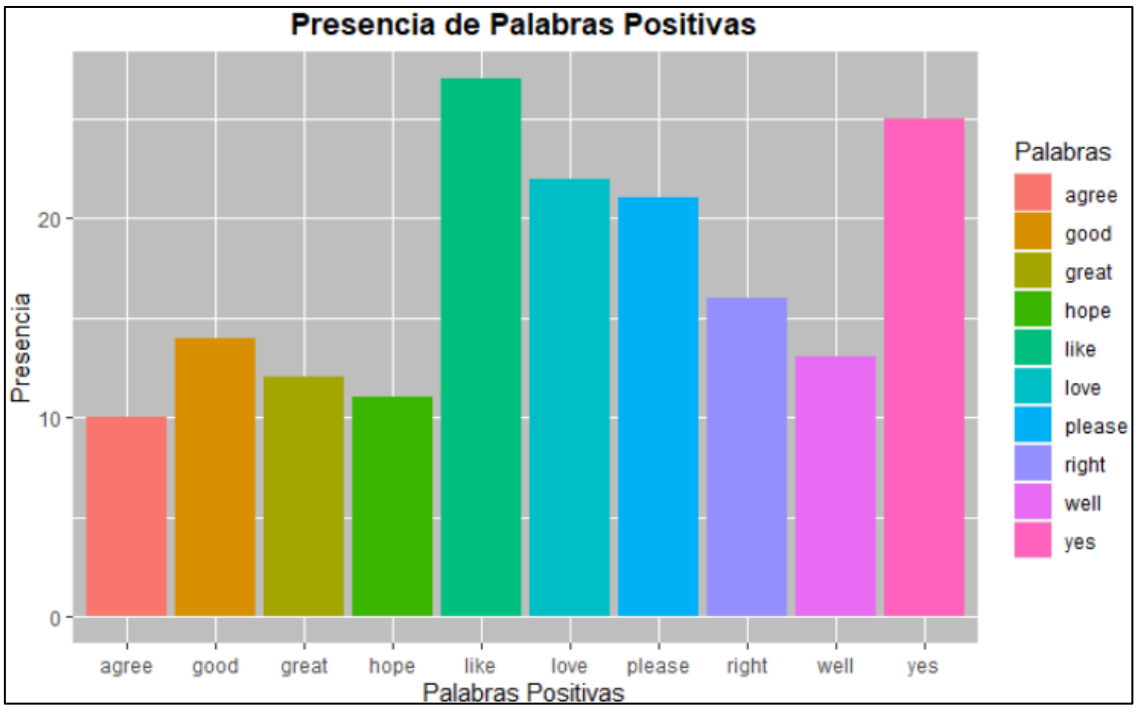


Ilustración 6.13 (Moreno, 2020)

La Ilustración 6.13 muestra un gráfico de barras correspondiente a las palabras positivas más frecuentes dentro de los tweets de Joe Biden.

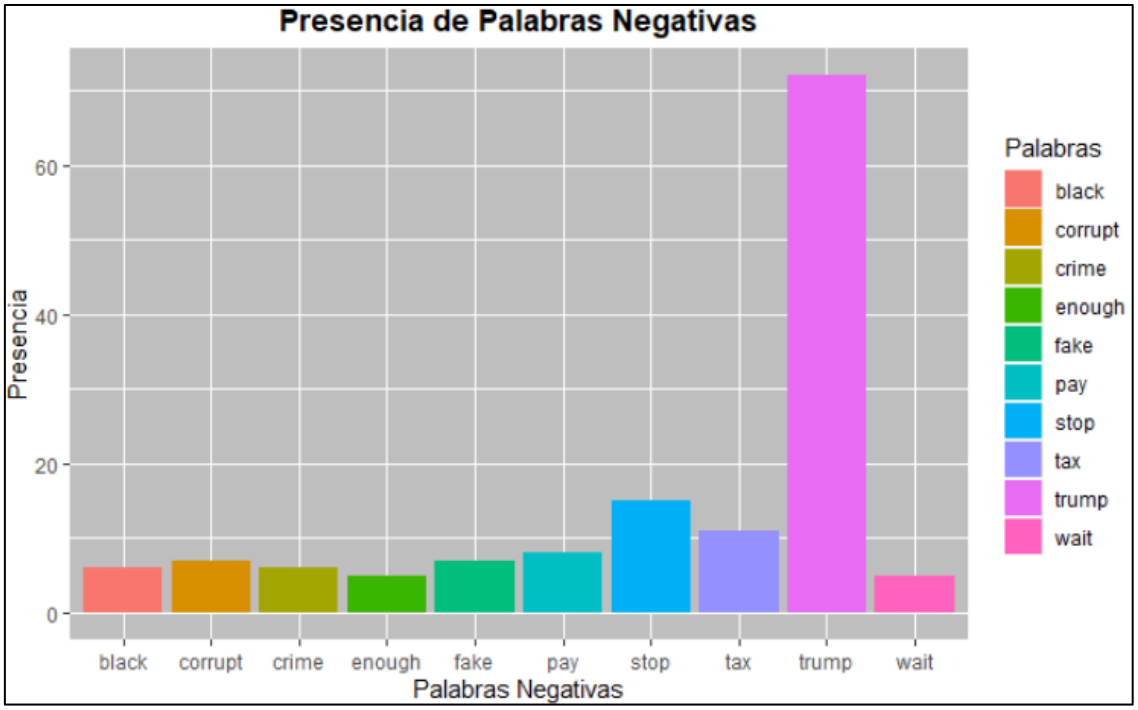


Ilustración 6.14 (Moreno, 2020)

La Ilustración 6.14 muestra un gráfico de barras correspondiente a las palabras negativas más frecuentes dentro de los tweets de Joe Biden.

Los sentimientos generales por Joe Biden son positivos y superan de manera contundente a los negativos. La palabra más negativa al igual que en el anterior análisis vuelve a ser “Trump”. Esto podría explicar el aparente balance entre sentimientos vistos en el gráfico de dispersión e histograma, debido a la gran cantidad de veces que dicho termino se repite a lo largo de varios documentos.

6.3 Resultados del Análisis de Emociones

La emotividad es un poco más complicada de analizar debido a que dentro de las categorías del análisis de sentimiento esta entra en el nivel de palabra el cual es uno de los complejos. Esto obliga a complementarla con otros análisis. En los capítulos previos se han realizado análisis a nivel de oración, documento y polaridad los cuales serán claves en esta etapa.

6.3.1 Resultados del Análisis de Emociones a Donald Trump

El análisis a nivel sentimiento nos mostraba que la mayoría de los documentos (tweets) tendían a ser negativos, otra cosa que también conocemos de ello son las palabras más negativas de los mismos. Lo que no conocíamos hasta entonces era que tipo de emociones estos representaban.

En la siguiente ilustración podremos apreciar la dispersión de las emociones en los tweets de Trump.

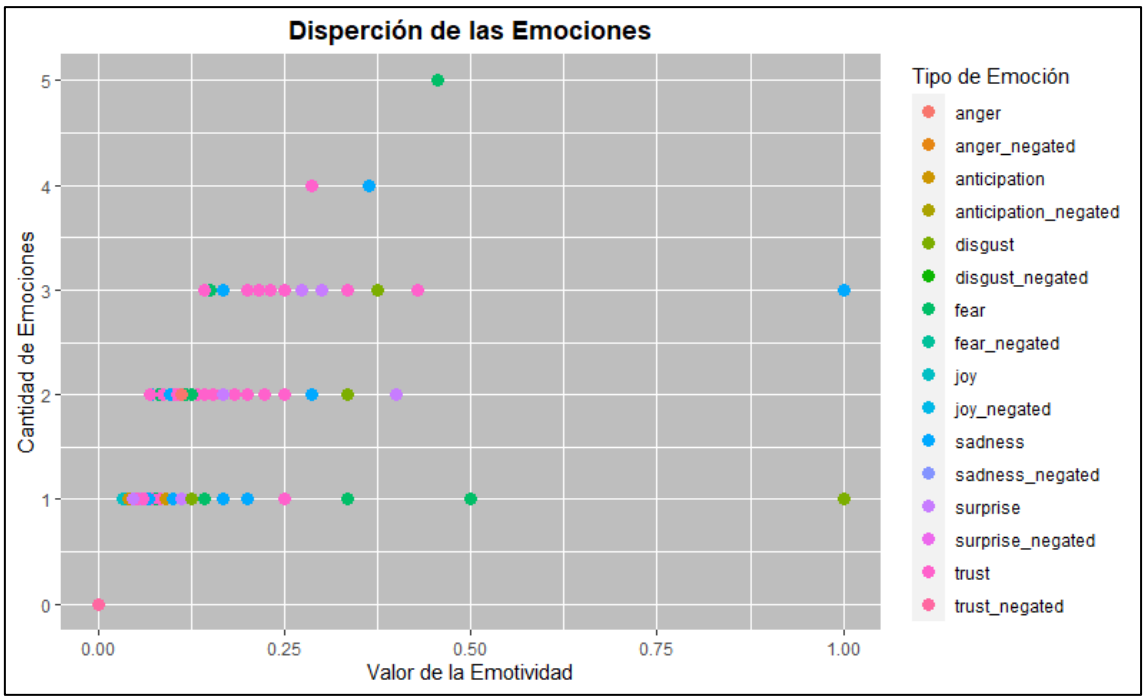


Ilustración 6.15 (Moreno, 2020)

La Ilustración 6.15 nos muestra que la mayor cantidad de emociones por documento es de 5 y que la mayoría de estas se encuentran entre 0 y 0.50. Esto indica que los tweets son poco emotivos pero que la mayor cantidad de emociones presentes en dicho rango son aquellas que evocan confianza, negación de la confianza, miedo y disgusto. De igual manera las palabras más emotivas son aquellas que pertenecen a la categoría de tristeza y disgusto.

La Ilustración 6.16 muestra cual es la categoría emotiva que más palabras acumula dentro de los tweets de Donald Trump.

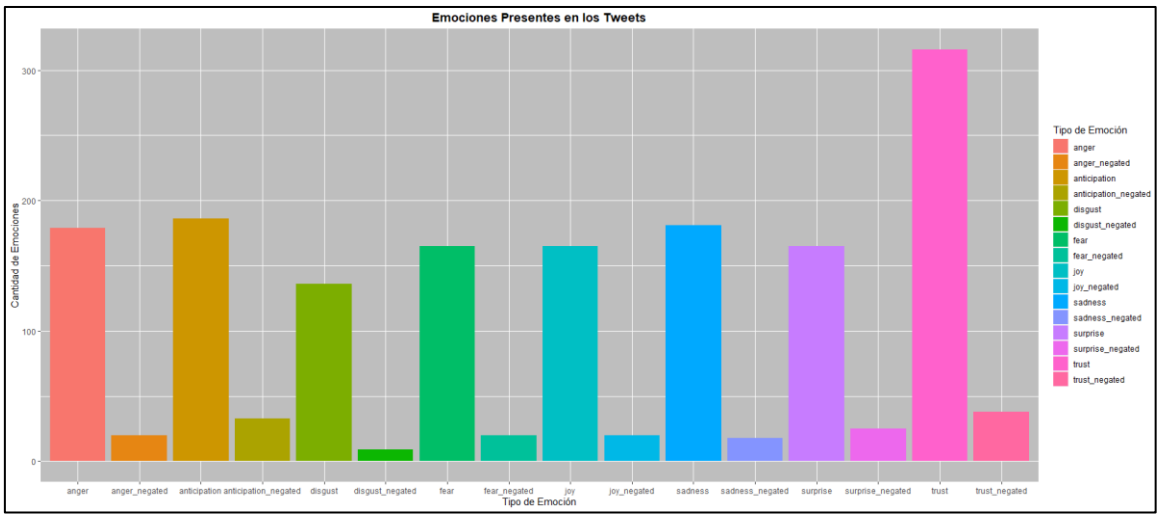


Ilustración 6.16 (Moreno, 2020)

La mayor cantidad de palabras dentro de los tweets de Trump pertenecen a la categoría emocional de “confianza”. Aun así, la mayor cantidad de sentimientos se concentran entre ira, anticipación, miedo y tristeza.

La ilustración anterior puede resumirse de mejor manera si utilizamos el método **extract_emotion_terms()** ya que este clasifica a las palabras en 8 categorías.

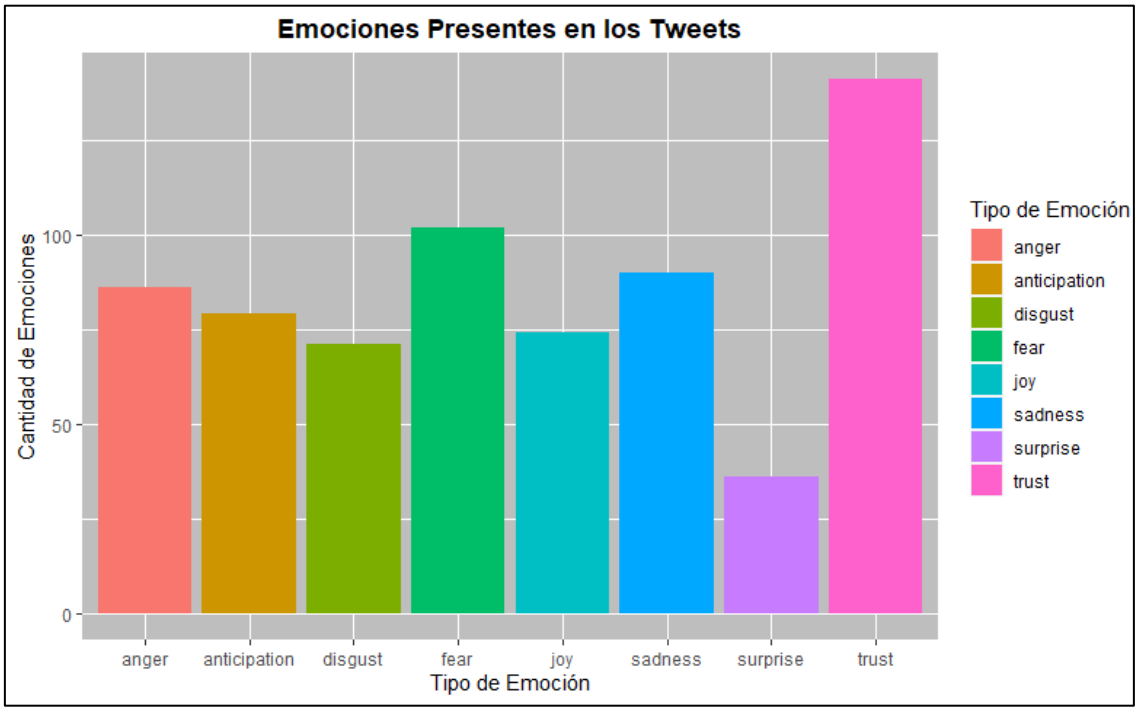


Ilustración 6.17 (Moreno, 2020)

La Ilustración 6.17 muestras las categorías emotivas en las que caen los términos presentes en los tweets de Donald Trump.

Aunque la mayor cantidad de palabras se encuentren en la categoría “confianza” la suma de las emociones negativas cómo ira, disgusto, miedo y tristeza son las que al final llegan a determinar la perspectiva global sobre Donald Trump. Esto junto con el análisis de sentimiento indican la poca simpatía con la que cuenta en la actualidad y cómo su rival Joe Biden es más aceptado. La gran presencia de palabras dentro de la categoría confianza tal cual como indica el gráfico de dispersión son poco emotivas (llegando a un valor cercano a 0.4 como máximo).

En conclusión, tanto la nube de palabras, análisis de sentimiento y emociones dan por sentado que Donald Trump es poco popular y en la mayor parte de casos hasta odiado por el público que lo sigue dentro de Twitter.

6.3.2 Resultado del Análisis de Emociones a Joe Biden

Dentro de esta sección se busca reforzar al análisis de sentimientos previo obteniendo conclusiones a través de las emociones presentes en los tweets que mencionan a Joe Biden.

La Ilustración 6.18 nos mostrará la dispersión de emociones que se encuentran en los tweets acerca de Joe Biden:

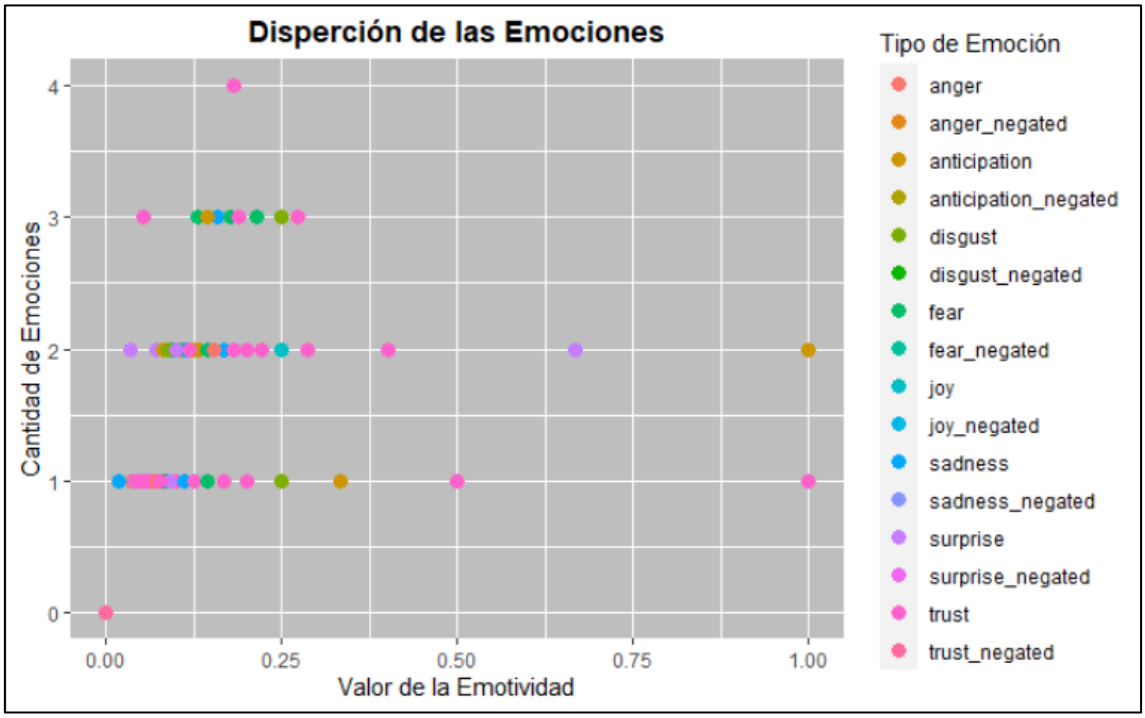


Ilustración 6.18 (Moreno, 2020)

El gráfico nos indica que la mayoría de los documentos son poco emotivos pero que la mayor parte de los términos evocan “confianza”. Además, podemos ver que la mayor emotividad se da en “confianza”, “anticipación” y sorpresa (todas mayores a 0.5). Las emociones negativas casi no aparecen y si lo hacen son “disgusto” y “tristeza”.

La Ilustración 6.19 nos mostrara que categoría emocional es la más presente dentro de los tweets de Joe Biden:

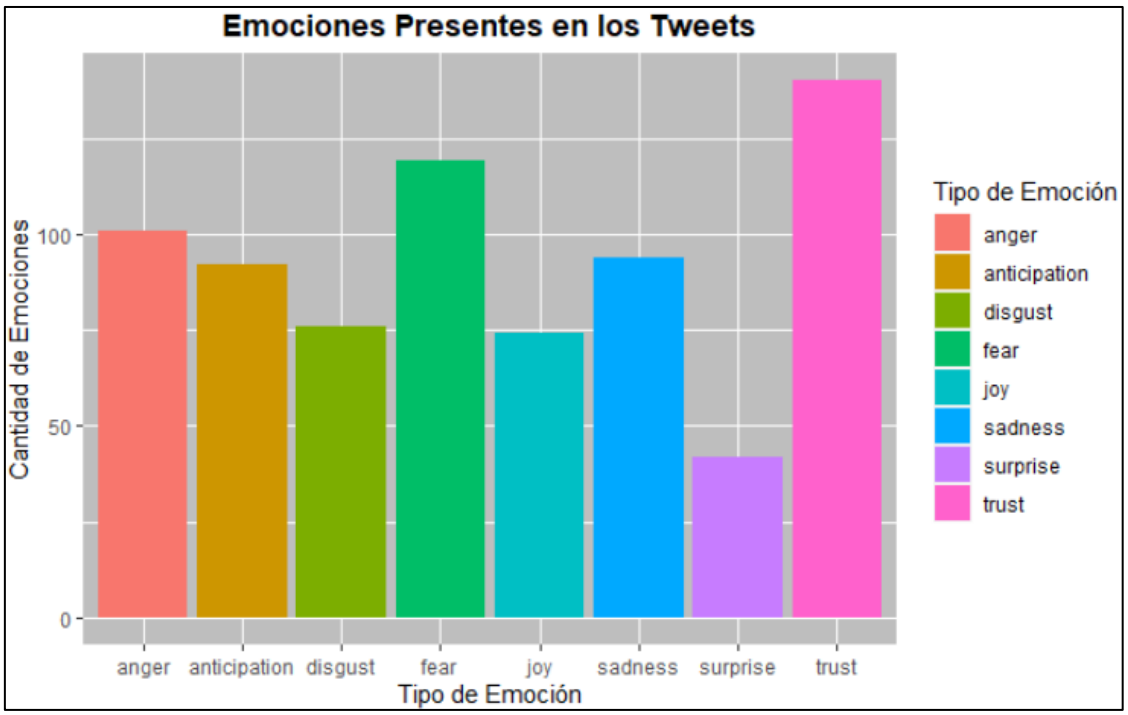


Ilustración 6.19 (Moreno, 2020)

Las emociones positivas más presentes son confianza, felicidad, sorpresa y anticipación. Aun así, las emociones negativas como ira y miedo son mayoría entre sorpresa y felicidad. Esto es el resultado de la presencia de "los modificadores de valencia" y la polivalencia emocional de una misma palabra según su contexto.

La presencia de emociones negativas es bastante grande pero no altamente emotiva, y si se revisa la Ilustración 6.18 podemos ver que el valor máximo que estos alcanzan es 0.25 en muchos casos. Esto quiere decir que son poco emotivas. A pesar de que las emociones positivas se presentan aparentemente como una minoría, es donde la emotividad llega a alcanzar valores de 1 en el caso de confianza y anticipación.

En conclusión, la perspectiva sobre Joe Biden es positiva dentro de los tweets recolectados sobre él, mientras que para Donald Trump los resultados son altamente negativos y prácticamente representan la otra cara de la moneda. Tanto el análisis de sentimiento cómo el de emociones han respondido a la pregunta planteada durante el capítulo 3 y han ayudado a la construcción del conocimiento sobre este tema.

6.4 Resultados de Combinar Porter con Lematización

El combinar el algoritmo de Porter con la lematización surgió como un experimento para conocer si existe algún tipo de resultado diferente a los obtenidos hasta el momento utilizando lo mejor de ambas aproximaciones.

Los resultados después de aplicar ambas técnicas se muestran en la Ilustración 6.20:

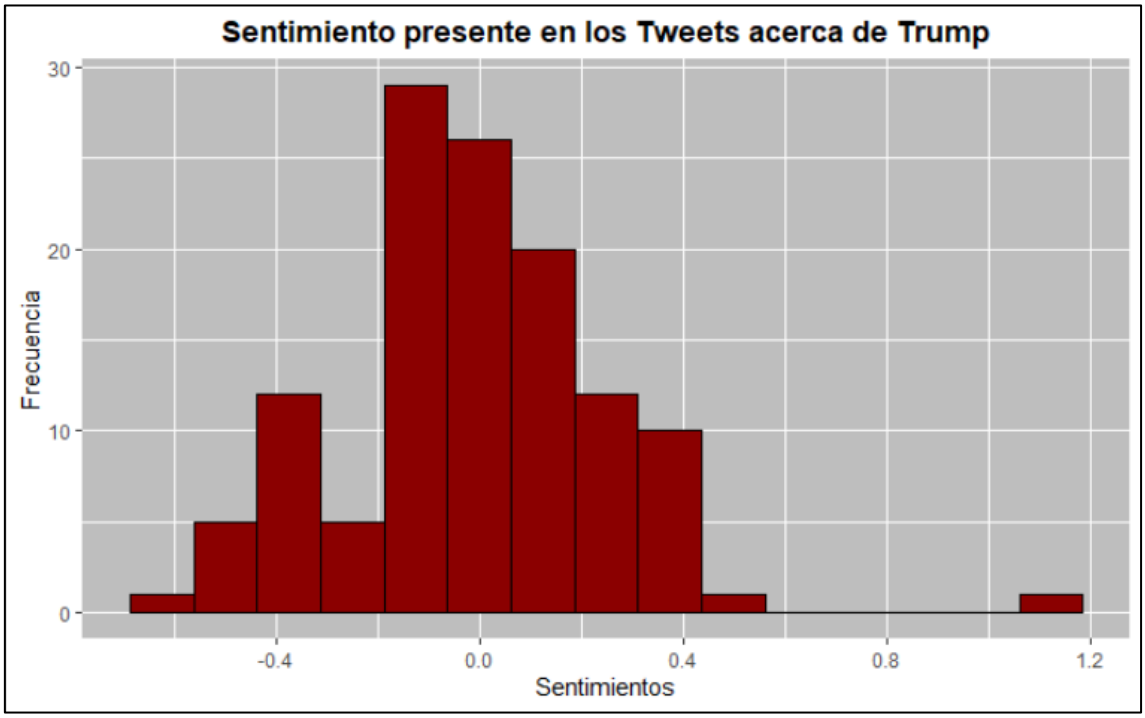


Ilustración 6.20 (Moreno, 2020)

Con un sesgo aparentemente positivo este gráfico se diferencia bastante del presentado en la ilustración 6.7 donde el sesgo es evidentemente negativo. Este gráfico se complementará con las ilustraciones 6.21, 6.22 y 6.23 para un mejor análisis.

El aparente sesgo positivo puede deberse al algoritmo de Porter ya que es probable que el algoritmo haya alterado términos negativos al no encontrarles “raíz”, de la misma manera este puede haber acabado con varios términos positivos.

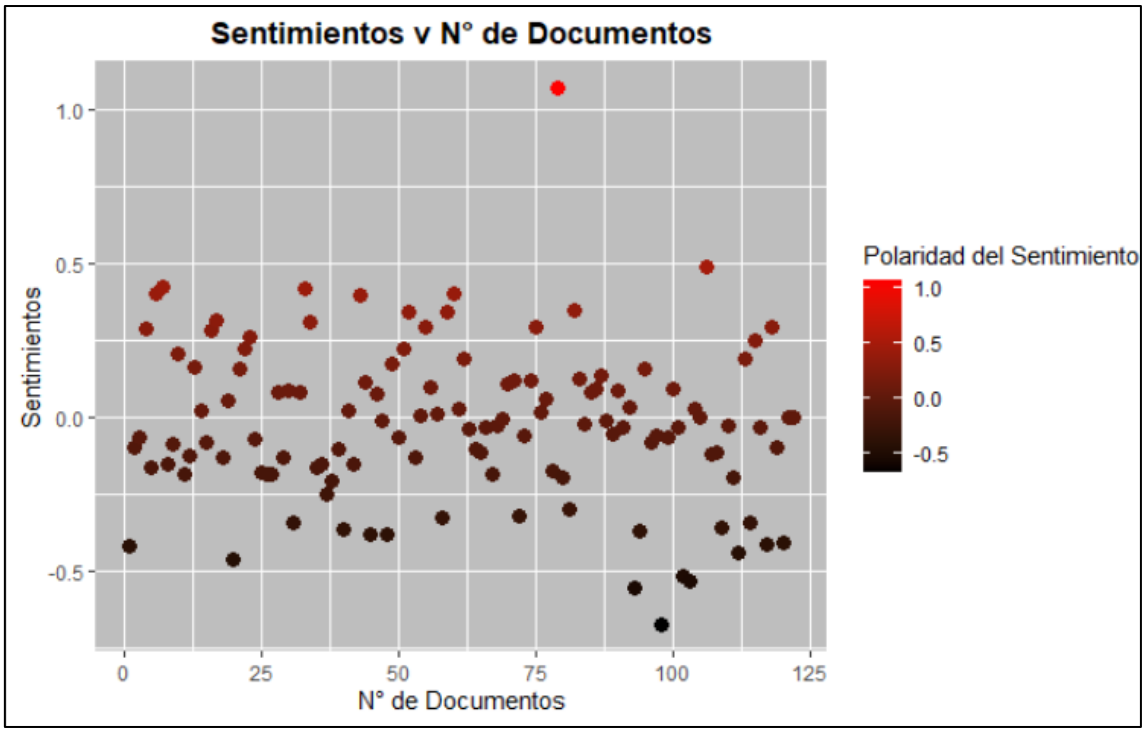


Ilustración 6.21 (Moreno, 2020)

La Ilustración 6.21 nos muestra que el haber aplicado el algoritmo de Porter en una etapa previa a la lematización ha reducido drásticamente la cantidad de términos neutros dentro los documentos. La tendencia sigue siendo la misma con Trump (existen más términos negativos que positivos en los documentos).

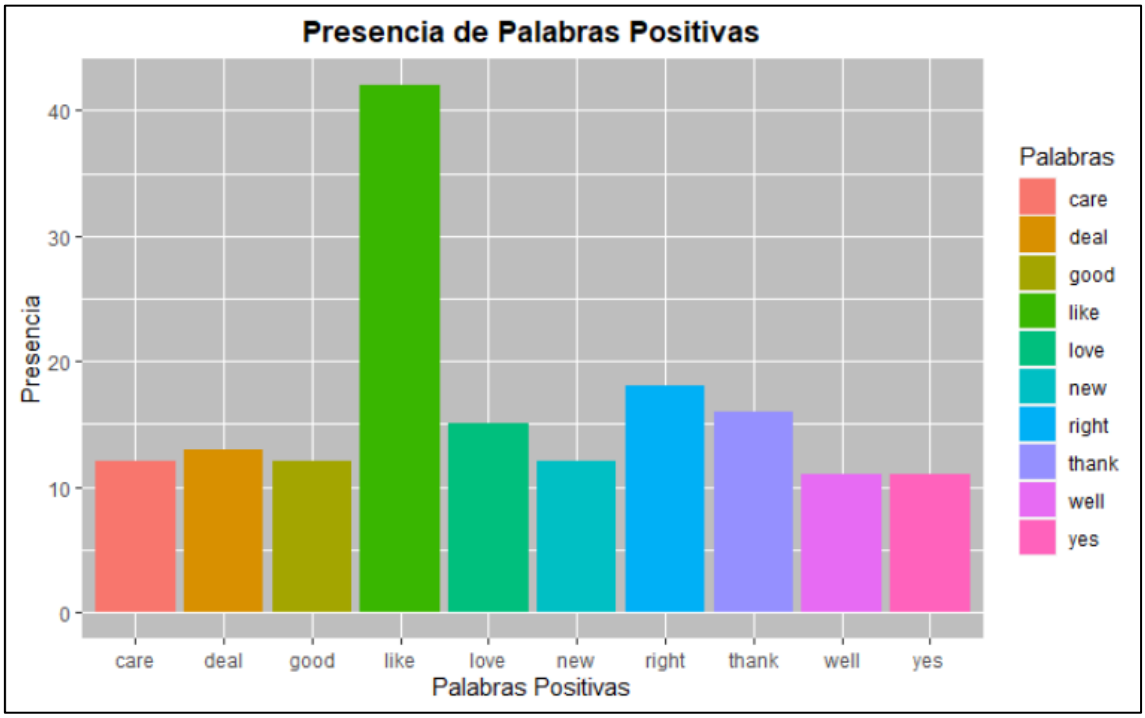


Ilustración 6.22 (Moreno, 2020)

La Ilustración 6.22 muestra las 10 palabras positivas más frecuentes en los tweets de Trump después de utilizar el algoritmo de Porter y la Lematización.

La palabra “like” al igual que en la aproximación convencional es el termino positivo más utilizado dentro de los tweets referentes a Donald Trump. Los términos “deal” y “thank” se hacen presentes dentro de este “top 10” y son los únicos términos que no aparecen en la ilustración 6.9 (donde solo se utilizó lematización).

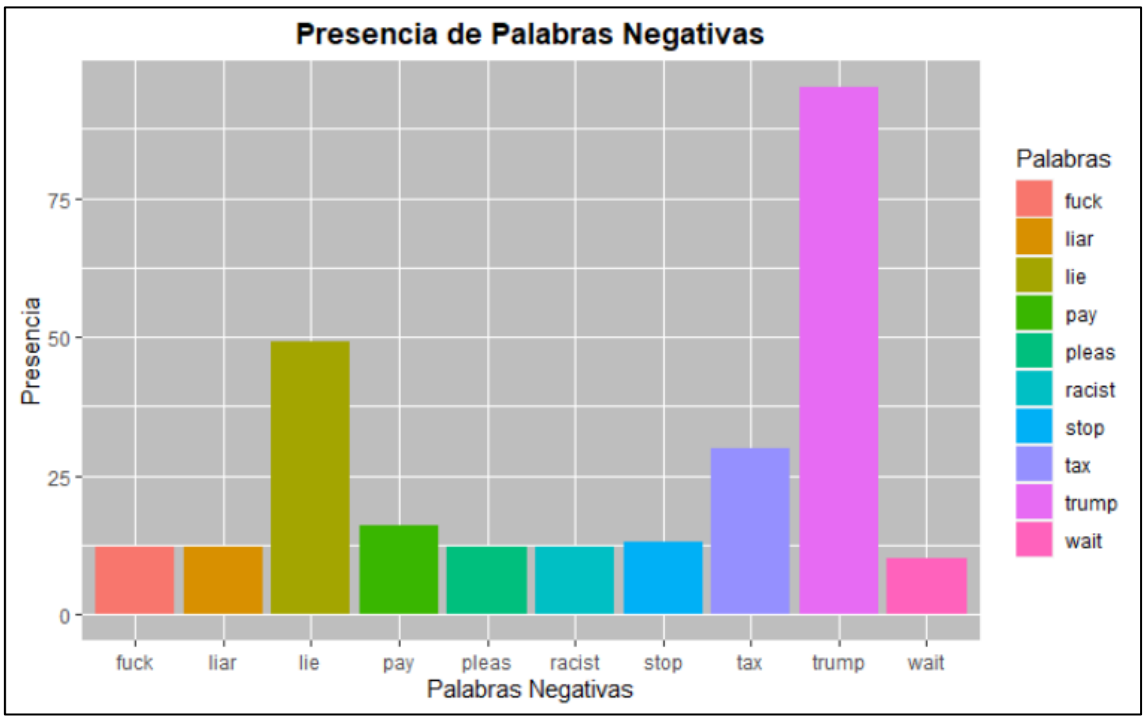


Ilustración 6.23 (Moreno, 2020)

La Ilustración 6.22 muestra las 10 palabras negativas más frecuentes en los tweets de Trump después de utilizar el algoritmo de Porter y la Lematización.

Al igual que con la aproximación convencional el termino negativo más repetido vuelve a ser Trump. Los nuevos términos que no aparecen en la ilustración 6.10 son “fuck”, “liar” y “wait”.

La combinación de lo mejor de ambas aproximaciones ha arrojado prácticamente los mismos resultados a excepción de la ilustración 6.14 en donde el resultado es bastante alejado al obtenido usando una sola aproximación. En conclusión, acelera y facilita el proceso de lematización, pero también se encarga de arrojar términos imprecisos los cuales pueden entorpecer o cambiar algún tipo de aproximación inicial. El uso de amabas técnicas al analizar algún caso en específico queda a discreción del investigador.

Capítulo 7 Conclusiones y Recomendaciones

7.1 Conclusiones

- El algoritmo de Porter es bastante ineficiente en lenguajes ajenos al inglés. Dicha conclusión se obtuvo probando el algoritmo con documentos en español, donde los resultados fueron prácticamente desastrosos debido a que aproximadamente un 70%-90% de los resultados arrojados eran palabras que no tenían ninguna asociación con el español.
- La aproximación por lematización al ser más moderna y contar con soporte activo de varias comunidades dedicadas al análisis de texto produce resultados superiores a los algoritmos de stemming (encontrar la raíz de una palabra en base a un algoritmo) debido a que los resultados arrojados son palabras que pertenecen a un lenguaje.
- La existencia de diccionarios de lematización para otros lenguajes cómo el español no existe, son escasos o son imprecisos. Esto obliga a que el análisis de sentimiento en este lenguaje se limite a las nubes de palabras o la creación de algoritmos propios bastante complejos.
- El candidato Joe Biden es más apreciado por el pueblo americano que el actual presidente Trump luego de haber realizado los respectivos análisis de sentimiento y emoción, en donde se encontró que la cantidad de términos y emociones negativas es mayor en el caso de Trump.
- La opinión que se tiene sobre los candidatos Yaku Pérez y Guillermo Lasso es negativa, debido a que la mayoría de los términos presentes dentro de sus nubes de palabras están llenas de sobrenombres, insultos y palabras que los asocian con acontecimientos polémicos cómo el feriado bancario y el paro nacional indígena provocado en el 2019 por el alza de los combustibles.
- El combinar al algoritmo de Porter con la lematización permite acelerar el proceso del análisis de sentimientos debido a que reduce la cantidad de palabras a indexar por los diccionarios y los términos neutros presentes en un corpus. Esto permite tomar casi directamente a todas las palabras que evoquen un sentimiento positivo o negativo.

7.2 Recomendaciones

- La elaboración de diccionarios en lenguajes ajenos al inglés para una correcta lematización en una mayor cantidad de idiomas, con el objetivo de hacer universal el análisis de texto.
- “Preparar” un diccionario propio cuando se trata con corpus en español para evitar la pérdida de información al momento de utilizar el algoritmo de Porter.
- Realizar cualquier tipo de análisis de sentimientos en lenguajes dedicados a la ciencia de datos como R y Python debido a que estos cuentan con librerías que simplifican y aceleran el proceso.
- Tener un amplio conocimiento en el idioma ingles debido a que la mayoría de las herramientas, algoritmos, fundamentos y recursos se encuentran escritos en este idioma en todo el tema referente al análisis de sentimiento, emoción y texto.

- Ampliar el conocimiento sobre los datos y la forma en la que estos se encuentran hoy en día, debido a que las aproximaciones convencionales no satisfacen la demanda de conocimiento actual.
- Tener conocimientos en estadística, gramática de gráficos y manejo de archivos para normalizar, resumir y procesar los textos que se lleguen a analizar.
- Las nuevas carreras deberían tener un camino “road map” enfocado a la ciencia de datos por si sola, esto quiere decir una fuerte fundación matemática en estadística, captura de datos, manipulación y tratamiento de datos. En caso de que aquello no sea posible una nueva carrera enfocada solo en el campo de los datos debería ser una consideración importante en un futuro próximo.
- La falta de desarrollo de algoritmos propios en español al igual que diccionarios y una buena versión del algoritmo de Porter limitan el análisis de sentimientos en épocas electorales o parecidas, a pesar de aquello una investigación colaborativa en los campos de la lingüística, la psicología y la computación pueden ser capaces de construir las herramientas necesarias para el español.

Capítulo 8 BIBLIOGRAFÍA

Bramer, M. (2016). *Principles of Data Mining.* London, United Kingdom: Springer-Verlag London Ltd.

Cebrían, M. (12 de Mayo de 2008). La Web 2.0 como red social de comunicación e información. Madrid, España: Universidad Complutense de Madrid.

Chaffey, D. (30 de Enero de 2020). *Smart Insights*. Obtenido de DIGITAL 2020: GLOBAL DIGITAL OVERVIEW: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science Big Data, Machine Learning, Using Python Tools.* Shelter Island: Manning Publications Co.

Clement, J. (21 de Noviembre de 2019). *Most popular social networks worldwide as of October 2019, ranked by number of active users*. Obtenido de Statista: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Clement, J. (17 de Octubre de 2019). *Number of internet users worldwide from 2009 to 2019, by region(in millions)*. Obtenido de Statista: https://www.statista.com/statistics/265147/number-of-worldwide-internet-users-by-region/

Collomb, A., Costea, C., Joyeux, D., Hasan, O., & Brunie, L. (2013). *A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation.* Obtenido de Semantic Scholar: https://www.semanticscholar.org/paper/A-Study-and-Comparison-of-Sentiment-Analysis-for-Collomb-Brunie/0e6e72c40f438c5c0e5c7ca47448d57e0a8c6e54

Comscore. (Octubre de 2019). *Comscore*. Obtenido de Top 50 Multi-Platform Properties (Desktop and Mobile) October 2019: https://www.comscore.com/Insights/Rankings

Economist, T. (6 de Mayo de 2017). The world's most valuable resource is no longer. *The Economist*, 1-4.

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource. Italia.

Fan, W., & Gordon, M. (Junio de 2014). *The Power of Social Media Analytics.* Obtenido de ResearchGate: https://s3.amazonaws.com/academia.edu.documents/43136252/Unveiling_the_Power_of_Social_Media_Analytics_CACM_final_2013.pdf?response-content-disposition=inline%3B%20filename%3DThe_Power_of_Social_Media_Analytics_Unve.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-

Hope, C. (6 de Junio de 2019). *Data*. Obtenido de Computer Hope: https://www.computerhope.com/jargon/d/data.htm

Hurtado, L., & Buscaldi, D. (Septiembre de 2015). ELiRF-UPV en TASS 2015: Análisis de Sentimientos en Twitter. Valencia, España.

IBM. (2013). *IBM Smarter Analytics Libe 2013.* Obtenido de IBM.

IBM. (2 de Diciembre de 2020). *IBM Watson Developer*. Obtenido de IBM Watson: https://www.ibm.com/watson/developer

Khan, G. F. (2015). SOCIAL MEDIA ANALYTICS: AN OVERVIEW. En G. F. Khan, *Seven Layers of Social Media Analytics: Mining Business Insights from Social Media* (pág. 21). CreateSpace Independent Publishing Platform.

Kharde, V. A., & Sonawane, S. S. (Abril de 2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications, 139*(11), 10.

Lacroix, H. (Mayo de 2020). *Data Engineering for Everyone.* Obtenido de DataCamp.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Claypool Publishers.

Moreno, I. (22 de Septiembre de 2020). *Developer Portal*. Obtenido de sentimientos_t_isra: https://developer.twitter.com/en/portal/apps/18854768/keys

NASDAQ. (2020 de Junio de 2020). *General Motors Company Common Stock*. Obtenido de Nasdaq: https://www.nasdaq.com/market-activity/stocks/gm

NASDAQ. (11 de Junio de 2020). *Tesla Inc. Common Stock*. Obtenido de Nasdaq: https://www.nasdaq.com/market-activity/stocks/tsla

Pérez, C. (2015). *R Lenguaje de programación y análsis estadístico de datos.* Madrid: Garceta.

Reinsel, D., Gantz, J., & Rydning, J. (Noviembre de 2018). *Data Age 2025 Sponsored by Seagate.* Obtenido de Seagate: https://www.seagate.com/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Statista, & TNW. (2019). *Number of people using social media platforms, 2009 to 2019*. Obtenido de Our World in Data: https://ourworldindata.org/grapher/users-by-social-media-platform?time=2009..&country=Facebook~Instagram~MySpace~Pinterest~Reddit~Snapchat~TikTok~Tumblr~Twitter~WeChat~Whatsapp~YouTube

Stieglitz, S., Dang-Xuan, L., Brans, A., & Neuberger, C. (Abril de 2014). *Social Media Analytics An Interdisciplinary Approach and Its Implications for Information Systems.* Obtenido de ResearchGate: https://www.researchgate.net/publication/271914787_Social_Media_Analytics

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). Data Ownership and Distribution. En *Introduction to Data Mining* (pág. 5). Harlow: Pearson Education Limited.

Vinodhini, G., & Chandrasekaran, R. (junio de 2012). Sentiment Analysis and Opinion Mining: A Survey. *International Journal of Advanced Research in Computer Science and Sofware Engineering, 2*(6), 283.

Wolfram. (2 de Diciembre de 2020). *WolframAlpha computacional intelligence*. Obtenido de https://www.wolframalpha.com/

Capítulo 9 Anexos

Script 1: Permite recuperar tweets utilizando los tokens entregados por el API del portal Twitter Developer. El límite máximo de tweets que se pueden recuperar es de 32000 actualmente.

```
library(tm) # para Text Mining
library(SnowballC) # para Stemming
library(twitteR) #descargar y trabajar con tweets
library(writexl) #transformar a xlsx

# Configurando el api de twitter
api_key<-'lwdtLoTQhehGEVcPSGZ97SZ1w'
api_secret_key<-'fLJsyRlt9woMLFo77ZnXkqvxpZSKuVYqQjIWovHhMbVWv3LVnQ'
access_token<-'1219922628-TVzuV1IWgLEFsRTHBUfBUVHsefRDIIilk5IHc42'
access_token_secret<-'AHqqKFyZ3GjE7bZJiqB3eGLRPci2CW1svGmSvT4Ks15vV'
setup_twitter_oauth(api_key,api_secret_key,access_token,access_token_s
ecret)

#Etapa de Recoleccion o Captura
# extrayendo tweets relacionados con la cuenta oficial de Donald Trump
tweets<-twitteR::searchTwitter('realDonaldTrump',n=2000)
#conservamos solo tweets no retweets
tweets<-twitteR::strip_retweets(tweets)
# cuenta el numero de tweets
n.tweets<-length(tweets)
n.tweets
#convierte la lista en un data frame
tweets.df<-twListToDF(tweets)
#permite visualizar el data frame
View(tweets.df)
#Guardamos el data frame como archivo en formato xls
# convierte en xls
#writexl::write_xlsx(tweets.df,"C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\archivos_ingles\\tweetstrump.xlsx")
```

Script 2: El siguiente script permite importar un archivo XLSX, transformar el archivo en un corpus, normalizar el corpus y crear una nube de palabras utilizando el algoritmo de Porter. Este script funciona solo para el idioma inglés.

```
library(tm)#permite mineria de texto
library(xlsx)#para leer archivos xlsx
library(SnowballC)#utilizar el algoritmo de Porter
library(wordcloud)#utilizar una nube de palabras

#leemos el archivo xlsx que contiene los tweets y lo transformamos en
un data frame
tweets.df<-read.xlsx("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R
Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
                     sheetIndex=1,
                     startRow=1,
                     colIndex=1)

#convirtiendo el data frame en un corpus para su normalizacion
myCorpus<-Corpus(VectorSource(tweets.df$text))
#inspeccion del corpus
inspect(myCorpus)
#limpieza y normalizacion del contenido del corpus
corpus_limpio<-tm_map(myCorpus,tolower)#convierte todo el texto en
minusculas
#removemos urls innecesarias dentrol del corpus
remove_url<- function(x) gsub("http[^[:space:]]*","",x)
corpus_limpio<- tm_map(corpus_limpio,content_transformer(remove_url))
#removiendo los numero dentro del corpus
corpus_limpio<-tm_map(corpus_limpio,removeNumbers)
#removiendo signos de puntuacion
corpus_limpio<-tm_map(corpus_limpio,removePunctuation)
#removiendo emojis
removeEmojo<-function(x)gsub("[^\x01-\x7F]","",x)
corpus_limpio<-tm_map(corpus_limpio,content_transformer(removeEmojo))
#removiendo palabras de parada
corpus_limpio<-tm_map(corpus_limpio,removeWords,stopwords(kind =
"en"))
#encontrnado las raices de las palabras
corpus_limpio<-tm_map(corpus_limpio,stemDocument)#se utiliza el
algoritmo de Porter
#arreglo de stopwords
corpus_limpio<-
tm_map(corpus_limpio,removeWords,c('realdonaldtrump','itsjefftiedrich'
,'gopchairwoman','trump','therightmelissa','rudygiuliani','breitbartne
w','page','that','minut','rudi',

'potus','joebiden','joe','biden','thehil','get','book','tri','let','go
p','your','even','just','noth','obama','amp','presssec',

'mani','anoth','one','look','hkrassenstein','barackobama','thing','new
','stevescalis','mikep','mean','still',

'now','tell','question','realli','hes','done','stahl','presid','lol',

'put','guy','wait','live','donald','said','ever','yes','isnt','wow',

'mayb','doesnt','ivankatrump','account','caslernoel','man','hand',

'tweet','ask','keitholbermann','post'))
#inspeccionamos una vez mas el corpus despues de su normalizacion
inspect(corpus_limpio)

#construccion de una Nube de Palabras
wordcloud(corpus_limpio,min.freq =10,random.order = F)
```

Script 3: Este script funciona de la misma manera que el anterior, la única diferencia es que aquí jamás se ejecuta el algoritmo de Porter. Esto debido a que el corpus se encuentra en español.

```
library(tm)#permite mineria de texto
library(xlsx)#para leer archivos xlsx
library(SnowballC)#utilizar el algoritmo de Porter
library(wordcloud)#utilizar una nube de palabras

#leemos el archivo xlsx que contiene los tweets y lo transformamos en
un data frame
tweets_lasso.df<-read.xlsx("C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\archivos_espanol\\tweetslasso.xlsx",
                     sheetIndex=1,
                     startRow=1,
                     colIndex=1)

#Convirtiendo el data frame en un corpus

#Normalizacion del corpus
myCorpus<-Corpus(VectorSource(tweets_lasso.df$text))
#inspeccion del corpus
inspect(myCorpus)
#limpieza y normalizacion del contenido del corpus
corpus_limpio<-tm_map(myCorpus,tolower)#convierte todo el texto en
minusculas
#removemos urls innecesarias dentrol del corpus
remove_url<- function(x) gsub("http[^[:space:]]*","",x)
corpus_limpio<- tm_map(corpus_limpio,content_transformer(remove_url))
#removiendo los numero dentro del corpus
corpus_limpio<-tm_map(corpus_limpio,removeNumbers)
#removiendo signos de puntuacion
corpus_limpio<-tm_map(corpus_limpio,removePunctuation)
#removiendo emojis
removeEmojo<-function(x)gsub("[^\x01-\x7F]","",x)
corpus_limpio<-tm_map(corpus_limpio,content_transformer(removeEmojo))
#removiendo palabras de parada
corpus_limpio<-tm_map(corpus_limpio,removeWords,stopwords(kind =
"es"))
#El algoritmo de porter no funciona bien en espanol
# documento<-stemDocument(as.character(corpus_limpio),
language="spanish")
# corpus_limpio<-Corpus(VectorSource(documento))
corpus_limpio<-
tm_map(corpus_limpio,removeWords,c('jajaja','parasos','inters','lassog
uillermo','aos','lfc','pas','est',

'sers','rutakritica','estn','cmo','aqu','dueo','tambin','van','jams',

'bosscec','fegasg','mashirafael','jaimenebotsaadi','mikeaulestia',

'ecuarauz','lasso','johncajasguijar','antenaunofm','caridadvela',

'comunicacionec','presidenciaec','teleamazonasec','reuterslatam',

'oeaoficial','seor','ttere','guillermo','carlosverareal',

'descacidh','rabascallcarlos','campaa','kchfmradio','guidochiriboga',

'isidroromeroc','leostagg','bancoguayaquil','creoecuador','panchoteran
',

'arauz','tinocotania','onuderechos','lolacienfuegos','mariapaularomo',
```

Script 4: Este script permite transformar archivos XLSX en corpus y a los corpus en archivos TXT. Esto permite luego usar los archivos TXT en una nube de palabras comparativa. Este script sirve tanto para inglés cómo para español.

```
#Transformacion de los corpus en textos
library(xlsx)#para leer archivos xlsx
library(tm)#herramientas para Text Mining
#tweets Trump
tweetsTrump.df<-read.xlsx("C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
                      sheetIndex=1,
                      startRow=1,
                      colIndex=1)
#tweets Biden
tweetsBiden.df<-read.xlsx("C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\archivos_ingles\\tweetsbiden.xlsx",
                      sheetIndex=1,
                      startRow=1,
                      colIndex=1)
#corpus Trump
CorpusTrump<-Corpus(VectorSource(tweetsTrump.df$text))

#corpus Biden
CorpusBiden<-Corpus(VectorSource(tweetsBiden.df$text))

#inspeccion de ambos Corpus
inspect(CorpusTrump)
inspect(CorpusBiden)

#transformacion los corpus a la extension txt
writeLines(as.character(CorpusTrump), con="C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\textos_ingles\\textoTrump.txt")

writeLines(as.character(CorpusBiden), con="C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\textos_ingles\\textoBiden.txt")
```

Script 5: Este script permite construir una nube de palabras comparativa a partir de dos documentos tipo TXT que se encuentren en el mismo directorio.

```
#Comparacion de nubes
library(tm)
library(SnowballC)
library(wordcloud)

#se crea un corpus que contiene los dos textos
corpus_TrumpyBiden<-Corpus(DirSource(directory = "C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R Sentimientos\\textos_ingles"))
#se comprueba que el corpus esta compuesto de dos documentos
summary(corpus_TrumpyBiden)

#Limpieza de los textos y el corpus
corpus_limpio<-tm_map(corpus_TrumpyBiden,tolower)#convierte todo el
texto en minusculas
#removemos urls innecesarias dentrol del corpus
remove_url<- function(x) gsub("http[^[:space:]]*","",x)
corpus_limpio<- tm_map(corpus_limpio,content_transformer(remove_url))
#removiendo los numeros dentro del corpus
corpus_limpio<-tm_map(corpus_limpio,removeNumbers)
#removiendo signos de puntuacion
corpus_limpio<-tm_map(corpus_limpio,removePunctuation)
#removiendo emojis
removeEmojo<-function(x)gsub("[^\x01-\x7F]","",x)
corpus_limpio<-tm_map(corpus_limpio,content_transformer(removeEmojo))
#removiendo espacios
corpus_limpio<-tm_map(corpus_limpio,stripWhitespace)
#removiendo palabras de parada
corpus_limpio<-tm_map(corpus_limpio,removeWords,stopwords(kind =
"en"))
#encontrnado las raices de las palabras
corpus_limpio<-tm_map(corpus_limpio,stemDocument)#se utiliza el
algoritmo de Porter
```

```
#eliminando terminos que no aportan valor
corpus_limpio<-
tm_map(corpus_limpio,removeWords,c('joebiden','realdonaldtrump','barac
kobama','steveschmidts','gopchairwoman',

'minut','therightmelissa','kamalaharri','trump','senronjohnson','rudyg
iuliani',

'joe','stevescalis','book','thehil','aoc','ananavarro','hkrassenstein'
,

'kayleighmcenani','gop','empti','itsjefftiedrich','page','biden','ivan
katrump',

'rvat','projectlincoln','breitbartnew','drbiden','toddfoot','hunter',

'mikep','devo','email','today','potus','that','gabbygifford','caslerno
el',

'sachabaroncohen','walshfreedom','amp','thedemcoalit','unleashthetea',

'youramerican','nataliegwint','dcexamin','shes','juli',

'scottpresl','week','markdic','ines','donald','proudsocialist','blank'
,

'mmpadellan','ilhanmnbeliev','ronna','keitholbermann','choic','call',

'anoth','ilhanmn','impeach','point','bidenrep','markmeadow','youv','st
ahl',

'lesley','shadi','paper','wolfblitz','ericbool','mean','ericbol',

'secpompeo','seen','cri','howardmortman','sarahdauterman','get','espn'
,

'bidenharri','nwhile','use','got','repdougcollin','checkmatest','ign',

'mani','presssec','son','cnnsitroom','asianamerican','ron','cent',

'republican','yes','parti','pleas','seem','question','look',

'whitehouse','even','whitehous','say','russianron','makeamericadecenta
gain',

'pennsylvania','traffick','florida','come','big','pictur','man',

'pay','new','jeffreyguterman','reason','benshapiro','jennpellegrino',

'glameleg','eugenegu','realdailywir','ear','bro','alli','twitter',
                                                    'will','kayleigh'))
wordcloud(corpus_limpio,min.freq = 10)
wordcloud(corpus_limpio,min.freq =10,colors =
brewer.pal(8,"Set2"),random.order = F,rot.per = .10)

#Creando la Nube de Comparacion
matrix_de_terminos<-TermDocumentMatrix(corpus_limpio) #crea una matriz
de terminos es un objeto propio de r
matrix_de_terminos<-as.matrix(matrix_de_terminos)# la transforma en
una matriz nxm
colnames(matrix_de_terminos)<-c('Biden','Trump')# nombramos a la nubes
```

Script 6: El siguiente script permite combinar el algoritmo de Porter con la Lematización para realizar un análisis de sentimientos.

```
library(xlsx)#para leer archivos xlsx
library(sentimentr)#Paquete completo para analisis de sentimientos
library(tidyverse) #Herramientas para Data Science
library(lexicon) #Lexicons disponibles
library(tm) #mineria de texto
library(SnowballC) #algoritmo de porter


tweets_trump.df<-read.xlsx("C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\archivos_ingles\\tweetstrump.xlsx",
                           sheetIndex=1,
                           startRow=1,
                           colIndex=1)

#Transformacion a Corpus Normalizacion
myCorpus<-Corpus(VectorSource(tweets_trump.df$text))

#Inspeccionando el corpus
inspect(myCorpus)

#Normalizando el Corpus
#limpieza y normalizacion del contenido del corpus
corpus_limpio<-tm_map(myCorpus,tolower)#convierte todo el texto en
minusculas
#removemos urls innecesarias dentrol del corpus
remove_url<- function(x) gsub("http[^[:space:]]*","",x)
corpus_limpio<- tm_map(corpus_limpio,content_transformer(remove_url))
#removiendo los numero dentro del corpus
corpus_limpio<-tm_map(corpus_limpio,removeNumbers)
#removiendo signos de puntuacion
corpus_limpio<-tm_map(corpus_limpio,removePunctuation)
#removiendo emojis
removeEmojo<-function(x)gsub("[^\x01-\x7F]","",x)
corpus_limpio<-tm_map(corpus_limpio,content_transformer(removeEmojo))
#removiendo palabras de parada
corpus_limpio<-tm_map(corpus_limpio,removeWords,stopwords(kind =
"en"))
#encontrnado las raices de las palabras
corpus_limpio<-tm_map(corpus_limpio,stemDocument)#se utiliza el
algoritmo de Porter
```

```
#Transformando el Corpus a texto
writeLines(as.character(corpus_limpio), con="C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\transformacion_corpus\\textoTrump.txt")

#Importando el corpus para que se lea como texto
texto<-readLines("C:\\Users\\Israel Moreno\\Desktop\\Israel\\R
Sentimientos\\transformacion_corpus\\textoTrump.txt")

#Obteniendo sentimientos y aplicando lematizacion
texto%>%
  get_sentences() %>%
  #sentiment(polarity_dt=lexicon::hash_sentiment_sentiword)-
>tweets_trump_sentimiento #obtiene el sentimiento de manera individual
  sentiment()->tweets_trump_sentimiento

#Obteniendo la polaridad de los sentimientos, positivo negativo,
neutral
texto%>%
  get_sentences() %>%
  extract_sentiment_terms()->tweets_trump_terminos_sentimiento

#Summary
terminos<-attributes(tweets_trump_terminos_sentimiento)$counts

summary(tweets_trump_sentimiento$sentiment)

#Obteniendo terminos positivos y negativos

#Obteniendo solo los terminos positivos
terminos_positivos<-terminos[polarity>0,]
#Obteniendo solo los terminos negativos
terminos_negativos<-terminos[polarity<0,]

#Revisando los data frames
head(terminos_positivos)
head(terminos_negativos)

#Obteniendo el top 10
#Top terminos Positivos
top_terminos_positivos<-terminos_positivos[n>=11,]

#Top terminos negativos
top_terminos_negativos<-terminos_negativos[n>=10,]

#Obteniendo solo los 10 primeros
top_terminos_positivos<-top_terminos_positivos[1:10,]

top_terminos_negativos<-top_terminos_negativos[1:10,]
```

```
#GRAFICOS
#Histograma de sentimientos utilizando GGPlot
ggplot(tweets_trump_sentimiento,aes(x=sentiment))+geom_histogram()
ggplot(tweets_trump_sentimiento,aes(x=sentiment))+geom_histogram(bins
= 15)
ggplot(tweets_trump_sentimiento,aes(x=sentiment))+geom_histogram(bins
= 15,fill="red4",col="black")->histograma_senti
#Etiquetas del Histograma y estilos
histograma_senti+labs(title="Sentimiento presente en los Tweets acerca
de Trump",x="Sentimientos",y="Frecuencia")->histograma_senti
histograma_senti+theme(plot.title = element_text(hjust =
0.5,face="bold"))->histograma_senti
histograma_senti+theme(panel.background = element_rect(fill =
"grey"))->histograma_senti
#Histograma final
histograma_senti

#Grafico de Dispercion de Sentimientos
ggplot(tweets_trump_sentimiento,aes(x=element_id,y=sentiment))+geom_po
int()
ggplot(tweets_trump_sentimiento,aes(x=element_id,y=sentiment,col=senti
ment))+geom_point(size=3)->disp_sentimiento
#Etiqeutas y estilos
disp_sentimiento+labs(title = "Sentimientos v N° de Documentos",x="N°
de Documentos",
                      y="Sentimientos",col="Polaridad del
Sentimiento")->disp_sentimiento
disp_sentimiento+theme(panel.background = element_rect(fill =
"grey"))->disp_sentimiento
disp_sentimiento+theme(plot.title = element_text(hjust = 0.5,face =
"bold"))->disp_sentimiento
disp_sentimiento+scale_color_gradient(low="#000000",high = "#ff0000")-
>disp_sentimiento
#Grafico de Dispercion final
disp_sentimiento
```

```
#GRAFICOS
ggplot(top_terminos_positivos,aes(x=words,y=n))+geom_col()
ggplot(top_terminos_positivos,aes(x=words,y=n,fill=words))+geom_col()-
>barras_positivo
#Etiquetas y estilos
barras_positivo+labs(title = "Presencia de Palabras
Positivas",x="Palabras Positivas",y="Presencia",fill="Palabras")-
>barras_positivo
barras_positivo+theme(panel.background = element_rect(fill = "grey"))-
>barras_positivo
barras_positivo+theme(plot.title = element_text(hjust = 0.5, face =
"bold"))->barras_positivo
#Grafico de Barras final
barras_positivo


ggplot(top_terminos_negativos,aes(x=words,y=n))+geom_col()
ggplot(top_terminos_negativos,aes(x=words,y=n,fill=words))+geom_col()-
>barras_negativo
#Etiquetas y estilo
barras_negativo+labs(title = "Presencia de Palabras
Negativas",x="Palabras Negativas",y="Presencia",fill="Palabras")-
>barras_negativo
barras_negativo+theme(panel.background = element_rect(fill="grey"))-
>barras_negativo
barras_negativo+theme(plot.title = element_text(hjust = 0.5,face =
"bold"))->barras_negativo
#Grafico de Barras final
barras_negativo
```

Script 7: El siguiente script permite realizar un análisis de sentimiento utilizando lematización, además también permite realizar un histograma, un gráfico de dispersión y dos gráficos de barras con los respectivos sentimientos encontrados. El script solo funciona para análisis en inglés.

```
library(xlsx)#para leer archivos xlsx
library(sentimentr)#Paquete completo para analisis de sentimientos
library(tidyverse) #Herramientas para Data Science
library(lexicon) #Lexicons disponibles

#Recuperando los tweets de Joe Biden
tweets_Biden.df<-read.xlsx("C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\archivos_ingles\\tweetsbiden.xlsx",
                     sheetIndex=1,
                     startRow=1,
                     colIndex=1)

#Obteniendo sentimientos
tweets_Biden.df %>%
  get_sentences() %>%
  sentiment()->tweets_Biden_sentimiento #obtiene el sentimiento de
manera individual

#Obteniendo sentimientos promedio
tweets_Biden.df %>%
  get_sentences() %>%
  sentiment_by()->tweets_Biden_sentimiento_promedio #obtiene el
sentimiento por grupos

#Obteniendo la polaridad de los sentimientos, positivo negativo,
neutral
tweets_Biden.df %>%
  get_sentences() %>%
  extract_sentiment_terms()->tweets_Biden_terminos_sentimiento

#Aplicando Summary
#Resumen estadistico de todas las columnas
summary(tweets_Biden_sentimiento)
#Resumen estadistico de la columan sentiment
summary(tweets_Biden_sentimiento$sentiment)

#Aplicando Summary al Sentimiento Promedio
summary(tweets_Biden_sentimiento_promedio)

summary(tweets_Biden_sentimiento_promedio$ave_sentiment)

#Obteninedo terminos postivos y negativos
terminos<-attributes(tweets_Biden_terminos_sentimiento)$counts

#Obteniendo solo los terminos positivos
terminos_positivos<-terminos[polarity>0,]
#Obteniendo solo los terminos negativos
terminos_negativos<-terminos[polarity<0,]
#Revisando los terminos positivos y negativos
head(terminos_positivos)
head(terminos_negativos)
```

```
#Graficos
#Histograma de sentimientos utilizando
ggplot(tweets_Biden_sentimiento,aes(x=sentiment))+geom_histogram()
ggplot(tweets_Biden_sentimiento,aes(x=sentiment))+geom_histogram(bins
= 20)
ggplot(tweets_Biden_sentimiento,aes(x=sentiment))+geom_histogram(bins
= 20,fill="blue4",col="black")->histograma_senti
#Etiquetas del Histograma y estilos
histograma_senti+labs(title="Sentimiento presente en los Tweets acerca
de Biden",x="Sentimientos",y="Frecuencia")->histograma_senti
histograma_senti+theme(plot.title = element_text(hjust =
0.5,face="bold"))->histograma_senti
histograma_senti+theme(panel.background = element_rect(fill =
"grey"))->histograma_senti
#Histograma final
histograma_senti

#Grafico de Dispercion de Sentimientos
ggplot(tweets_Biden_sentimiento,aes(x=element_id,y=sentiment))+geom_po
int()
ggplot(tweets_Biden_sentimiento,aes(x=element_id,y=sentiment,col=senti
ment))+geom_point(size=3)->disp_sentimiento
#Etiqeutas y estilos
disp_sentimiento+labs(title = "Sentimientos v N° de Documentos",x="N°
de Documentos",
                      y="Sentimientos",col="Polaridad del
Sentimiento")->disp_sentimiento
disp_sentimiento+theme(panel.background = element_rect(fill =
"grey"))->disp_sentimiento
disp_sentimiento+theme(plot.title = element_text(hjust = 0.5,face =
"bold"))->disp_sentimiento
disp_sentimiento+scale_color_gradient(low="#000000",high = "#ff0000")-
>disp_sentimiento
#Grafico de Dispercion final
disp_sentimiento

#Top terminos Positivos
top_terminos_positivos<-terminos_positivos[n>=10,]

#Top terminos negativos
top_terminos_negativos<-terminos_negativos[n>=5,]

#Obteniendo solo los 10 primeros
top_terminos_positivos<-top_terminos_positivos[1:10,]

top_terminos_negativos<-top_terminos_negativos[1:10,]
```

```
#GRAFICOS
ggplot(top_terminos_positivos,aes(x=words,y=n))+geom_col()
ggplot(top_terminos_positivos,aes(x=words,y=n,fill=words))+geom_col()-
>barras_positivo
#Etiquetas y estilos
barras_positivo+labs(title = "Presencia de Palabras
Positivas",x="Palabras Positivas",y="Presencia",fill="Palabras")-
>barras_positivo
barras_positivo+theme(panel.background = element_rect(fill = "grey"))-
>barras_positivo
barras_positivo+theme(plot.title = element_text(hjust = 0.5, face =
"bold"))->barras_positivo
#Grafico de Barras final
barras_positivo


ggplot(top_terminos_negativos,aes(x=words,y=n))+geom_col()
ggplot(top_terminos_negativos,aes(x=words,y=n,fill=words))+geom_col()-
>barras_negativo
#Etiquetas y estilo
barras_negativo+labs(title = "Presencia de Palabras
Negativas",x="Palabras Negativas",y="Presencia",fill="Palabras")-
>barras_negativo
barras_negativo+theme(panel.background = element_rect(fill="grey"))-
>barras_negativo
barras_negativo+theme(plot.title = element_text(hjust = 0.5,face =
"bold"))->barras_negativo
#Grafico de Barras final
barras_negativo
```

Script 8: El siguiente script permite realizar un análisis a nivel de emociones mediante lematización. El script permite visualizar a las emociones como un gráfico de dispersión y otro de barras. El script solo funciona para análisis en inglés.

```
library(xlsx)#para leer archivos xlsx
library(sentimentr)#Paquete completo para analisis de sentimientos
library(tidyverse) #Herramientas para Data Science
library(lexicon) #Lexicons disponibles

tweets_Biden.df<-read.xlsx("C:\\Users\\Israel
Moreno\\Desktop\\Israel\\R
Sentimientos\\archivos_ingles\\tweetsbiden.xlsx",
                           sheetIndex=1,
                           startRow=1,
                           colIndex=1)


#Obteniendo sentimientos
tweets_Biden.df %>%
  get_sentences() %>%
  emotion()->tweets_Biden_emociones #obtiene el sentimiento de manera
individual

#Obteniendo sentimientos promedio
tweets_Biden.df %>%
  get_sentences() %>%
  emotion_by()->tweets_Biden_emociones_promedio #obtiene el
sentimiento por grupos

#Obteniendo la polaridad de los sentimientos, positivo negativo,
neutral
tweets_Biden.df %>%
  get_sentences() %>%
  extract_emotion_terms()->tweets_Biden_terminos_emociones

#Emociones

terminos_valorados<-attributes(tweets_Biden_terminos_emociones)$counts
```

```
#Grafico de Dispercion de las emociones
ggplot(tweets_Biden_emociones)
ggplot(tweets_Biden_emociones,aes(y=emotion_count,x=emotion))+geom_poi
nt()
ggplot(tweets_Biden_emociones,aes(y=emotion_count,x=emotion,col=emotio
n_type))+geom_point(size=3)->dispercion_emociones
#Etiquetas y estilos
dispercion_emociones+labs(title="Disperci□n de las Emociones",x="Valor
de la Emotividad",y="Cantidad de Emociones",col="Tipo de Emoci□n")-
>dispercion_emociones
dispercion_emociones+theme(plot.title = element_text(hjust =
0.5,face="bold"))->dispercion_emociones
dispercion_emociones+theme(panel.background = element_rect(fill =
"grey"))->dispercion_emociones
#Grafico final
dispercion_emociones


#Grafico de Barras
terminos_valorados<-terminos_valorados[emotion>0,]
#El metodo
ggplot(terminos_valorados)
ggplot(terminos_valorados,aes(x=emotion_type,fill=emotion_type))+geom_
bar()->barras_terminos
#Etiquetas y estilos
barras_terminos+labs(title="Emociones Presentes en los Tweets",x="Tipo
de Emoci□n",
                     y="Cantidad de Emociones",fill="Tipo de
Emoci□n")->barras_terminos
barras_terminos+theme(plot.title = element_text(hjust =
0.5,face="bold"))->barras_terminos
barras_terminos+theme(panel.background = element_rect(fill = "grey"))-
>barras_terminos
#Grafico final
barras_terminos
```