CMA IADCOM - UBA CENTRO DE INVESTIGACIÓN EN METODOS CUANTITATIVOS APLICADOS A LA ECONOMÍA Y LA GESTIÓN

Universidad de Buenos Aires Facultad de Ciencias Económicas

http://www.economicas.uba.ar/institutos_y_centros/revista-modelos-matematicos/

ANÁLISIS DE CONTENIDO EN TWITTER Y EL AISLAMIENTO SOCIAL OBLIGATORIO

Salaberry, Natalia

Universidad de Buenos Aires, Facultad de Ciencias Económicas, Av. Córdoba 2122 – 1120AAQ. Ciudad Autónoma de Buenos Aires, República Argentina

Resumen

Recibido: 04-05-2020

Aceptado: 09-07-2020

Palabras clave

Cuarentena, Coronavirus, Tweets, Datos alternativos, Minería de texto

Este trabajo analiza el contenido en tweets que contienen una etiqueta asociada al aislamiento social obligatorio, medida adoptada por las Autoridades Nacionales de Argentina el 20 de marzo de 2020 frente a la expansión de la pandemia COVID-19. Mediante la utilización de técnicas de minería de texto, el objetivo de este trabajo es la clasificación de tweets a partir del sentimiento contenido en los mismos expresados por usuarios en la red social Twitter respecto del aislamiento social obligatorio. Con este propósito se procesan mediante técnicas de minería de datos, 2.965 tweets recolectados a través de la API (*Application Programming Interface*) de Twitter de uso público y gratuito. A partir de su análisis mediante técnicas de minería de texto, se concluye que el 49,8% de los tweets contiene un sentimiento negativo respecto del aislamiento social obligatorio. Si bien las técnicas utilizadas podrían ampliarse, la metodología presentada puede considerarse una buena primera aproximación al procesamiento y análisis de datos alternativos para la generación de valor agregado.

ISSN (En línea) 2362 3225

CONTENT ANALYSIS ON TWITTER AND MANDATORY SOCIAL ISOLATION

Abstract

KEYWORDS

Quarantine, Coronavirus, Tweets, Alternative Data, Text Mining.

This work analyzes the sentiment on tweets that contain a tag associated with mandatory social isolation, a decision adopted by the National Authorities of Argentina on March 20 of the present Year against the expansion of the COVID-19. Using text mining techniques, the objective of this work is the classification of tweets based on the sentiment expressed by users on the Twitter social network regarding mandatory social isolation. 2.965 tweets have been collected through the free public API (Application Programming Interface) of Twitter. Based on analysis performed using text mining techniques, it is concluded that 49.8% of tweets contain a negative feeling regarding compulsory social isolation. Although the techniques used could be expanded, the presented methodology can be considered a good first approach to the processing and analysis of alternative data.

ISSN (En línea) 2362 3225

INTRODUCCIÓN

La explosión de los medios digitales durante el presente siglo ha facilitado la posibilidad de contar con grandes volúmenes de datos. Esto dio lugar a la conformación de un ecosistema de datos, sobre el que las diferentes organizaciones se encuentran cada vez más interesadas, en la medida que les brinda la oportunidad de obtener más información sobre los individuos y mejorar su estrategia de toma de decisiones. En este sentido es posible hablar de un ecosistema de grandes volúmenes de datos (*Big Data*) (Kolanovic y Krishnamachari, 2017) que ha tenido un alto impacto generando desde nuevas formas comunicacionales hasta cambios de hábitos en los individuos, causando un cambio socioeconómico alrededor del mundo entero. Pero tales datos cuentan con la particularidad de poseer diversidad, de ser a gran escala y sin una estructura definida, lo que conlleva una dificultad para su tratamiento.

Una de las principales fuentes digitales de datos son las redes sociales. En gran medida esto se debe a que las organizaciones las utilizan como un canal de comunicación, para brindar información a sus clientes, o a la ciudadanía en caso de tratarse de organizaciones públicas. De esta manera, se convierten en un espacio generador de datos ya que son los propios individuos los que terminan exponiendo características sobre sí mismos. Entonces, su potencial radica en el hecho de ser una fuente de datos que permite obtener atributos sobre los diferentes individuos (Preot¸iuc-Pietro, Lampos y Apetras, 2015). Es así como surge un interés por parte de las organizaciones para poder extraer y explotar los datos que las diferentes redes sociales ponen a disposición, con el fin de poder sumar información nueva o completaría a sus procesos productivos, permitiéndoles generar mayor o nuevo valor agregado en el proceso de toma de decisiones.

En este sentido el presente trabajo persigue el objetivo de determinar el sentimiento de los individuos respecto de la cuarentena obligatoria establecida por las Autoridades Nacionales en Argentina. Para alcanzar el objetivo planteado se implementarán técnicas de minería de texto (*Text Mining*) sobre tweets que contienen una etiqueta asociada (*Hashtag*) a la temática "cuarentena obligatoria". De esta manera, se mostrará la potencialidad que posee la explotación de datos alternativos.

Para llevar adelante esta tarea, en primer lugar, se recolectarán tweets que contengan "cuarentenaobligatoria" como Hashtag, obteniéndose un conjunto de datos no estructurados. En segundo lugar, mediante el procesamiento de los datos obtenidos, se mostrará el tratamiento necesario que debe darse a los datos mediante técnicas de minería de datos (*Data Mining*) para que los mismos puedan estructurarse y de ese modo poder realizar un análisis de estos. Finalmente, a partir de la aplicación de algoritmos de minería de texto, se obtendrán resultados y se evaluará su potencial impacto para la toma de decisiones.

1. ACERCA DE GRANDES VOLÚMENES DE DATOS

Con la explosión de la disponibilidad de datos en grandes volúmenes, toma mayor fuerza la idea de que no es posible manejar aquello que no se puede medir (McAfee, Brynjolfsson, Davenport y Barton, 2012). La disponibilidad de datos en formato digital comienza a tomar mayor relevancia en el proceso de toma de decisiones, desde el momento que surge la posibilidad de obtener mejores resultados que llevan a un mayor conocimiento del público objetivo. En particular, la disponibilidad online de datos presenta una dinámica de volumen, variedad y velocidad de

disponibilidad de información que puede ser utilizada para adoptar cambios estratégicos con mayor inmediatez.

La definición de grandes volúmenes de datos (*Big Data*) no solo refiere a la idea de disponer de grandes volúmenes de datos, que pueden ser de tipo estructurados y no estructurados, sino también refiere a un conjunto de tecnologías que abarcan desde el almacenamiento, procesamiento y transformación de los datos. En primer lugar, se establecen tres características definitorias de grandes volúmenes de datos: volumen, velocidad y variedad. Por volumen se hace referencia a la tecnología necesaria para recolectar y almacenar grandes volúmenes de datos con el fin de procesarlos para transformarlos en información de utilidad. Dada la explosión de datos digitales en la última década, el volumen de datos comenzó a crecer de manera exponencial, con la particularidad de ser mayoritariamente datos de tipo no estructurado agregando complejidad para su almacenamiento y procesamiento. Este tipo de datos por ejemplo son los obtenidos de redes sociales, sensores e imágenes entre muchas otras fuentes. La magnitud del volumen de tal tipo de datos ha llevado al desarrollo de nuevas formas de negocio o de transformación de los existentes, en la medida que se ha encontrado una forma clara de explotarlos (Eberendu, 2016).

En cuanto a la velocidad, resulta ser en muchos casos más importante que el volumen (McAfee, Brynjolfsson, Davenport y Barton, 2012) en la medida que una organización cuente con la agilidad suficiente para captar los datos en tiempo real. En este sentido, los datos ya dejan ser un stock para ser un flujo constante (Eberendu, 2016), lo que lleva a la necesidad de procesamiento diario y hasta por hora en ocasiones. De aquí que, el procesamiento requiere de cierta inteligencia superior para lograr captar la mayor cantidad de diferentes tipos de datos provenientes de diversas fuentes y con la mayor velocidad posible, de forma tal de convertirlos en valor agregado para la toma de decisiones.

La variedad de datos resulta de la existencia de una amplia diversidad a partir de la disponibilidad digital de los mismos. Desde datos obtenidos a partir de emails, transacciones online hasta los derivados de videos, audios entre otros. Dada esta diversidad, el procesamiento de datos para la toma de decisiones requiere de procesos que sinteticen y simplifiquen la información que pueda obtenerse de ellos. A su vez, presentan la posibilidad de contar con información sobre cualquier tema de interés para la organización (McAfee, Brynjolfsson, Davenport y Barton, 2012), llegando a conformar entre el 70 y 80 por ciento de los datos utilizados por las organizaciones (Eberendu, 2016).

De esta manera, la conjunción de volumen, velocidad y variedad de datos que ofrece un sistema de Big Data, lo convierten en un ecosistema de datos que ha implicado una transformación disruptiva en las organizaciones a la hora de delinear estrategias para la toma de decisiones. Y es en este sentido que surge la necesidad de contar con un maestro de datos que sea transversal a la organización con el fin de brindar consistencia y calidad sobre estos para que realmente sean convertidos en valor agregado.

2. LOS DATOS ALTERNATIVOS EN EL PROCESO DE GESTIÓN DE DATOS

Los datos digitales resultan ser de una gran diversidad y con enorme escalabilidad, siendo en su mayoría datos de tipo no estructurado. La incorporación de estos requiere de una transformación

en el diseño del proceso de almacenamiento de datos. Tradicionalmente, dado un tipo de dato estructurado, el proceso de almacenamiento de los datos consistía en la extracción, transformación y carga (ETL sus siglas en inglés). Frente a la necesidad de incorporar un nuevo tipo de dato que esencialmente resultan ser de tipo no estructurado, se plantea pensar de manera diferente aquel proceso. En este sentido, se propone un enfoque de extracción, carga y transformación (Schmarzo, 2013).

Una característica particular de este tipo de datos es que son datos automáticos. Esto es, se obtienen de manera casi inmediata, a través de diferentes dispositivos como por ejemplo sensores y GPS entre muchos otros. Frente a ello, las organizaciones se encuentran interesadas en poder captarlos con mayor velocidad a diferencia de un procesamiento ETL. Tal interés surge con el objetivo de poder dar respuesta más inmediata de modo de seguir la dinámica veloz en que los datos son generados. Para este tipo de casos, existen diferentes alternativas de almacenamiento de datos que están orientadas a un procesamiento en paralelo de forma tal de garantizar un acceso performante a los datos y a alta velocidad. En este sentido, podría decirse que, desde una visión técnica, la tecnología necesaria para procesar tales datos se encuentra resuelta.

Al mismo tiempo, para lograr la calidad e integridad de los datos se requiere aplicar un proceso de datos maestros. Siendo los datos de diferente tipo originados en diferentes fuentes, se requiere de datos maestros (o definiciones transversales a una organización) que permitan su integración de una manera exitosa, garantizando su calidad. De esta manera, un proceso de datos maestros permitirá integrar datos de calidad con el fin de obtener información enriquecida generando valor agregado. Para ello la etapa de control se convierte en vital, dado que permite sincronizar, de ser necesario en tiempo real, los datos maestros. Esto es, durante la extracción y carga de datos podría detectarse nuevos valores que requieran una modificación de los datos maestros para que los nuevos datos puedan ser integrados. Tal detección puede obtenerse de manera automática o como resultado de un análisis realizado por alguna persona a cargo.

Bajo esta nueva metodología de integración de datos alternativos, podrían surgir metadatos que sean temporales y otros permanentes. Estos últimos requieren de almacenamiento en el maestro de datos mientras que los primeros pueden ser útiles durante el tiempo de ejecución de un proceso (Russom, 2015). A diferencia de los datos estructurados y su procesamiento en formato ETL, los datos no estructurados podrían presentar un ciclo de vida corto que no requerirán de un almacenamiento permanente. Es en este sentido que el tratamiento de datos alternativos marca una diferencia a ser tenida en cuenta.

Finalmente, a pesar de que la integración de datos alternativos requiere de adaptaciones o modificaciones de los procesos tradicionales de gestión de datos, representan un claro potencial para la generación de valor agregado. Bajo un esquema adecuado de gestión, puede lograrse la integridad de estos, y garantizar su calidad. De esta manera, se logrará un proceso de toma de decisiones más robusto en la medida que se sustenta en mayor conocimiento a partir de obtener más información.

3. PROCESAMIENTO Y ANÁLISIS DE DATOS ALTERNATIVOS

Los datos pueden ser principalmente de dos tipos: estructurados y no estructurados. Por datos estructurados se refiere a aquellos que se encuentran ordenados en formatos de columnas y filas

perfectamente legibles y listos para ser procesados. Por el contrario, los datos no estructurados son aquellos que no poseen una estructura interna, razón por la cual requieren de un tratamiento inicial para poder hacer uso y comprensión de estos.

En particular, los datos no estructurados, también llamados datos alternativos, pueden presentar alguna ventaja en la información que proporcionan. Tal ventaja puede consistir en descubrir nueva información no contenida en fuentes tradicionales, o descubrir una misma información, pero anticipadamente (Kolanovic y Krishnamachari, 2017). Entre este tipo de datos se encuentran los producidos y publicados por los individuos, como por ejemplo las publicaciones que realizan en redes sociales, los generados a partir de transacciones online como los provenientes de comercio electrónico, entre otros, o aquellos generados por sensores como por ejemplo las imágenes satelitales.

Dada las características de los datos no estructurados, las técnicas de minería de datos ofrecen la posibilidad de aplicar algoritmos que permiten procesar datos hasta en tiempo real y de gran volumen. Una metodología frecuente en la que se sustentan para el procesamiento de texto como es el caso del texto contenido en las publicaciones en redes sociales, es el procesamiento de lenguaje natural (*Natural Language Processing*). El procesamiento de lenguaje natural consiste en un análisis de datos de texto, utilizando métodos computacionales, cuyo objetivo es la construcción de una representación sobre el contenido del texto que agregue una estructura al lenguaje natural no estructurado contenido en el cuerpo de aquel (Verspoor y Khoen, 2013). Dicha estructura puede ser de naturaleza sintáctica, capturando las relaciones gramaticales entre los componentes del texto, o semántica, obteniendo el significado que está transmitiendo el contenido textual.

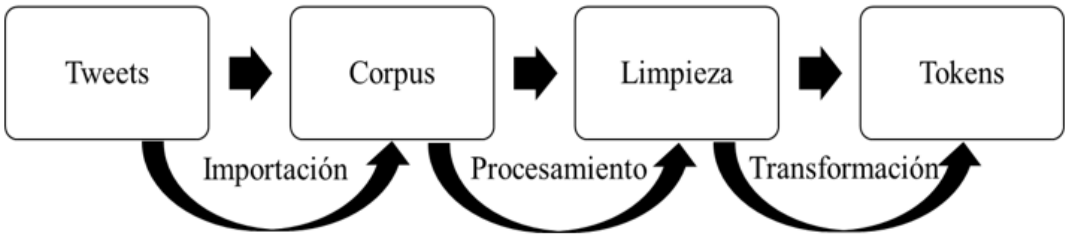
Tal proceso consiste en determinar las reglas en cada oración, eliminar aquellas palabras que no aportan significado al sentido del texto (*Stop Word*) y reducir las palabras a su nivel raíz removiendo los sufijos y la pluralidad (*Stemming*) con el fin de obtener un procesamiento más veloz (Fiore, Almodovar, Assoumou, Dutta, y Cotoranu, 2017). A partir de dicha estructura base de procesamiento es que surgen diferentes técnicas de análisis para este tipo de datos que se engloban en el conjunto de algoritmos de minería de texto.

4. OBTENCIÓN Y PROCESAMIENTO DE TWEETS

Frente al objetivo de recolectar datos no estructurados, en el presente apartado se describirá el procesamiento de datos realizado mediante técnicas de minería de datos. En cuanto a la recolección de tweets que poseen "cuarentenaobligatoria" como Hashtag, se realizó a través de una API de Twitter de uso público y gratuito, con conexión a través del software RStudio. Dado que existe una limitación en cuanto a la cantidad de tweets que es posible extraer para un limitado período de 7 días, se realizaron diferentes extracciones durante mayo y junio de 2020. Este período resulta relevante para el análisis llevado a cabo en este trabajo ya que coincide con un anuncio realizado por las Autoridades Nacionales de Argentina sobre la extensión del período de Aislamiento Social Obligatorio, el día 24 de mayo de 2020. De esta manera, se logró conformar un set de datos con 2.965 tweets a partir de las extracciones realizadas, el cual se almacenó en formato CSV (*Comma-separated values*). La cantidad de tweets obtenida no constituye un límite establecido por la API de Twitter, sino que, fue la cantidad posible de obtener en forma aleatoria en el período antes mencionado.

Conformado el conjunto de datos, se procedió a implementar técnicas de minería de datos para realizar la limpieza de estos y luego estructurarlos de forma tal que facilite su interpretación y posterior análisis. En la Figura 1 se especifican las etapas de procesamiento. En primer lugar, se importa el archivo que contiene los datos. Luego, se construyó un cuerpo de texto (*Corpus*) a partir del texto contenido en cada tweet, conformando una colección de documentos, donde cada documento es un texto. En tercer lugar, se lleva a cabo un proceso de limpieza del texto contenido en cada documento, mediante la función "tm_map" de la librería "tm"1 en RStudio. Esto con el fin de eliminar todo tipo de simbología y direcciones que no constituyen una palabra que pueda aportar significado. Adicionalmente, se eliminan aquellas palabras que no aportan significado al sentido del texto (*Stop Word*) en base a una lista dada2, como son las preposiciones entre otras. También se realiza reducción de palabras (*Stemming*) con el fin de reducir las palabras a su nivel raíz removiendo los sufijos y la pluralidad. Este procesamiento también se llevó a cabo mediante la función "tm_map" mencionada anteriormente. Finalmente, se obtienen unidades de texto (*Tokens*), conformadas por palabras simples.

Figura 1: Procesamiento de Tweets



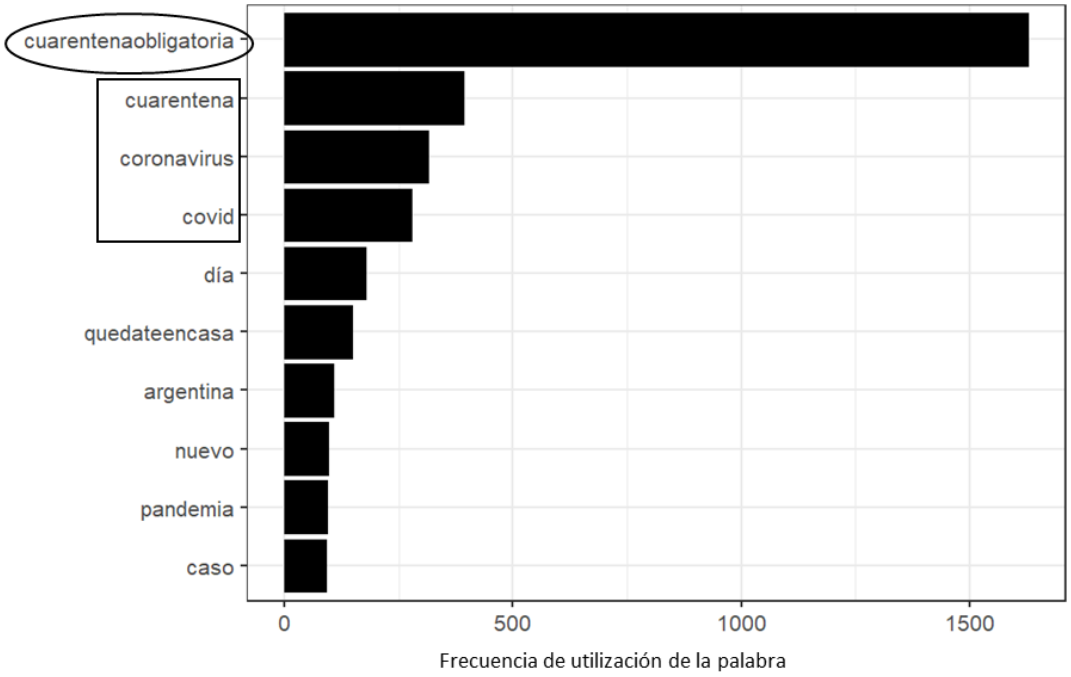
Fuente: Elaboración propia

De esta manera, se estructuraron los tweets iniciales en un cuerpo de texto que contiene 2.965 documentos, donde cada documento contiene unidades de texto. La totalidad de las unidades de texto alcanzan las 27.848. Realizando un top de palabras en función de su frecuencia de aparición a lo largo del cuerpo de texto, se puede observar en la Figura 2, que la palabra "cuarentenaobligatoria" se encuentra en el primer lugar. Este es un resultado esperado dado que se trata de detectar un sentimiento asociado al aislamiento social obligatorio. Resulta interesante que en segundo lugar aparece la palabra "cuarentena", siendo un indicativo de que los tweets bajo análisis refieren a una expresión de conformidad o no respecto del aislamiento social obligatorio. Luego, en tercer y cuarto lugar aparecen las palabras "coronavirus" y "covid", resultando esperable en la medida que es la causa por la cual se establece el aislamiento social obligatorio.

[1] Librería "tm" del Package "Text Mining Package" Ingo Feinerer [aut, cre], Kurt Hornik [aut], Artifex Software, Inc. [ctb, cph] http://tm.r-forge.r-project.org/

[2] Lista de Stop Word tomada de https://github.com/6/stopwords-json

Figura 2: Top de frecuencia de palabras.



Fuente: Elaboración propia con RStudio

Frente a la exploración inicial realizada, es posible notar que un procesamiento como el realizado permite comenzar a ver expresiones asociadas a una problemática o disconformidad o acuerdo que los usuarios expresan a través de la red social Twitter. En el siguiente aparatado se lleva a cabo un análisis de sentimientos que permitirá observar resultados más concretos acerca de la opinión de los usuarios sobre el aislamiento social obligatorio.

5. ANÁLISIS DE TWEETS MEDIANTE TÉCNICAS DE MINERÍA DE TEXTO

La minería de texto (*Text Mining*), puede definirse como el proceso de descubrimiento y extracción de conocimiento a partir de un texto no estructurado (Kao y Poteet, 2007). En este sentido constituye en un conjunto de herramientas que brinda la posibilidad de examinar grandes volúmenes de texto, con el objetivo de generar información para su posterior análisis. Existen diversos algoritmos de minería de texto, que permiten obtener nuevo conocimiento a partir de un texto. Entre los más frecuentemente utilizados se encuentra la Nube de Palabras (*Word Cloud*). Este suele utilizarse en una etapa inicial de análisis con el fin de detectar en un texto las palabras más frecuentemente utilizadas, mediante una visualización rápida y sencilla. Su potencial, para el caso de analizar tweets, radica en el hecho de que permite obtener una visión de lo que está pensando el usuario, basada en la frecuencia de utilización de determinadas palabras (Kabir, Karim, Newaz y Hossain, 2018).

Para poder construir una nube de palabras primero debe crearse un "*Document Term Matrix*" (DTM) que consiste en una matriz cuyas filas son los documentos creados y cuyas columnas son cada uno de los términos contenidos en los documentos. Luego, el contenido de cada celda será la frecuencia con que ocurrió cada término en cada documento. La ventaja que presenta esta forma

de representación consiste en que se pasa de un formato de texto a un formato numérico, permitiendo realizar operaciones en memoria de manera optimizada. Llevando a cabo esta implementación mediante la librería "wordcloud2"3 se obtiene la Figura 3.

Figura 3: Nube de palabras

Fuente: Elaboración propia con RStudio

En la Figura 3 se puede observar fácilmente que la palabra "cuarentena" resulta ser la más utilizada en los comentarios realizados por los usuarios de la red social, lo cual resulta esperable, como ya se mencionó anteriormente. Adicionalmente, es de destacar que surgen diferentes palabras con importancia alta a su alrededor. "covid, coronavirus, pandemia, salud, días, casa, argentina, noticias" son las que presentan mayor importancia alrededor de la palabra cuarentena. Llegado a este punto, puede establecerse que se cuenta con un indicio de posible disconformidad a partir de la identificación de palabras claves.

La disconformidad posiblemente surja como consecuencia de una medida adoptada por las Autoridades Nacionales que implican la imposibilidad de poder realizar las actividades de rutina, como por ejemplo ir a trabajar, hacer deporte al aire libre, entre muchas otras. Sin duda, existe un cambio de hábito en los individuos que los lleva a expresar un sentimiento respecto de la medida adoptada, conformando una huella acerca de su pensamiento. Es por ello por lo que resulta interesante poder captar tales sentimientos a la hora de analizar tweets. Una metodología frecuentemente utilizada en este sentido es el Análisis de Sentimiento (*Sentiment Analysis*). Este método, trata de buscar el sentimiento contenido en un texto, por ejemplo, en un tweet, en términos de actitudes, emociones y opiniones, con el fin de proporcionar una respuesta adecuada al usuario (Chen y Franks, 2016).

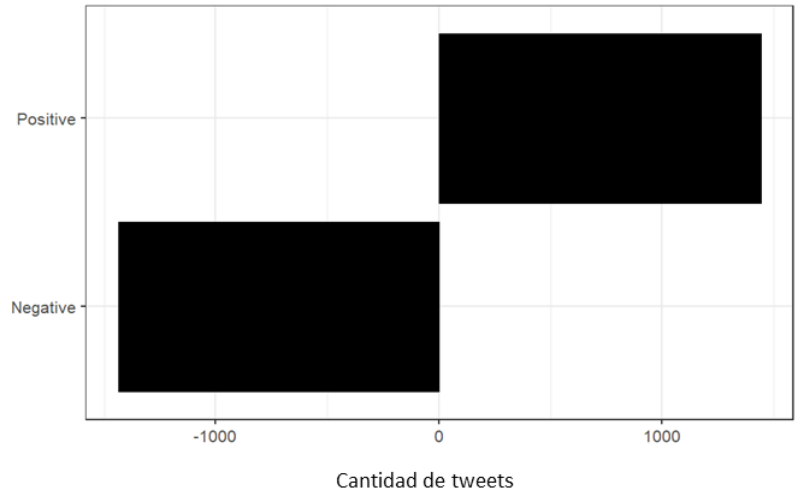
El método de análisis de sentimiento depende en gran medida de un léxico de sentimiento (u opinión) subyacente. Un léxico de sentimiento es una lista de características léxicas (por ejemplo,

[3] Liberia "wordcloud2" del Package "Word Cloud 2" https://cran.r-project.org/web/packages/wordcloud2/index.html

palabras) que generalmente se etiquetan según su orientación semántica como positiva o negativa (Hutto y Gilbert, 2014). Tal lista es validada previamente por el analista, de forma tal que pueda adecuarla al idioma bajo el cual esta expresado el texto bajo análisis. Luego, para aquellos sentimientos que no pueden ser definidos, se le asigna una semántica neutral. A su vez, dentro de la clasificación en negativo o positivo, se asigna un puntaje entre 1 y 6, pudiendo variar la escala según el algoritmo utilizado, para determinar la intensidad del sentimiento conformando un rango de valores positivos o negativos para cada caso. La determinación del puntaje sobre cada palabra proviene del resultado de diferentes análisis psicológicos y grafológicos sobre el impacto sentimental que le causa a un individuo determinada palabra, teniendo en cuenta la simbología utilizada, el uso o no de mayúsculas, entre otros[4].

Para llevar a cabo la implementación de un modelo de análisis de sentimiento, un método frecuentemente utilizado, y que se emplea en este trabajo, es reducir a unidades de texto los documentos contenidos en un cuerpo de texto, es decir, generar un conjunto de datos conformado por cada una de las palabras en formato de lista. Una vez llevada a cabo esta tarea mediante la función "tidy"[5], se realiza una unión con la lista de léxicos con el fin de clasificar a cada palabra en negativa, neutral o positiva, utilizando la librería "dplry"[6] en RStudio. Finalmente, se toma el puntaje asignado a cada palabra en cada documento y se sumariza el mismo. De esta manera, se obtiene una clasificación de tweets en positivo o negativo. De la implementación realizada, se obtuvo el siguiente resultado:

Figura 4: Sentimientos asociados a las palabras de cada tweet



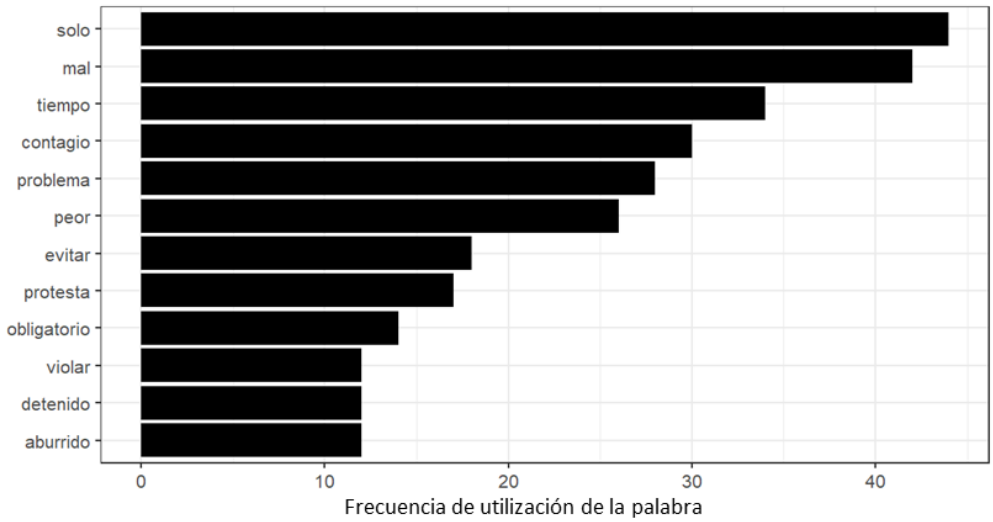
Fuente: Elaboración propia con RStudio.

En la Figura 4 se puede observar la distribución de los tweets en función de la clasificación del sentimiento expresado en cada palabra utilizada. De un total de 2.965 tweets analizados 1.437 contienen palabras asociadas a un sentimiento negativo y 1.444 contienen palabras asociadas con un sentimiento positivo. De esta manera, se obtuvo que el 49,9% del contenido de los tweets expresa un sentimiento negativo mientras que el 50,1% expresa un sentimiento positivo. A

[4] http://liwc.wpengine.com/
[5] Librería "tidyr" del Package "Tidy Messy Data" MIT + file LICENSE URL https://tidyr.tidyverse.org
[6] Librería "dplry" del Package "A Grammar of Data Manipulation", MIT + file LICENSE, https://cran.r-project.org/web/packages/dplyr/index.html

continuación, se realiza un top 10 de frecuencia de palabras asociadas a un sentimiento negativo que fueron utilizadas en los tweets.

Figura 5: Sentimientos negativos asociados a cada palabra



Fuente: Elaboración propia con RStudio

De la Figura 5 se puede observar que las principales palabras que surgen expresando un sentimiento negativo son palabras fácilmente identificables con este sentimiento. Así, la palabra "solo" se encuentra en el top de las palabras con sentimiento negativo, pudiéndose asociar a un sentimiento de soledad en un contexto de aislamiento social. Y en segundo lugar surge la palabra "mal". Luego las palabras "tiempo", "contagio" y "problema" también se encuentran como las de mayor frecuencia de utilización asociadas a un sentimiento negativo.

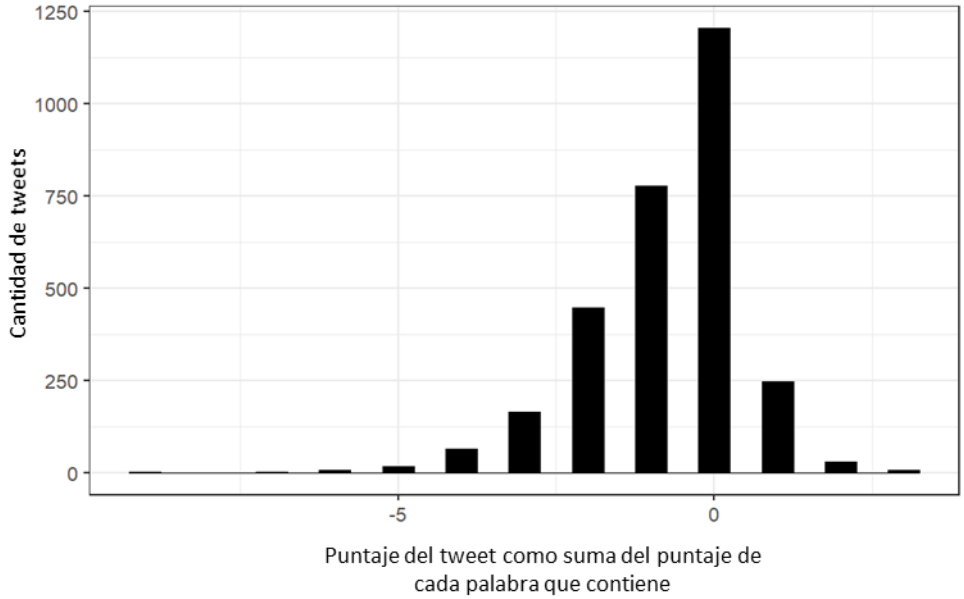
Por lo tanto, a través de la aplicación de la metodología de análisis de sentimiento se pudo determinar que prácticamente el 50% del contenido de los tweets que realizaron los usuarios expresan un sentimiento negativo respecto del aislamiento social obligatorio. Teniendo en cuenta que inicialmente se llevó a cabo una nube de palabras lo que permitió una visualización sencilla de frecuencia de términos donde también surgían las mismas palabras como las de mayor frecuencia de uso, es posible determinar que efectivamente se trata de tweets asociados a un sentimiento negativo.

6. RESULTADOS

Dado que el objetivo final de análisis consiste en poder hacer una clasificación sobre el sentimiento asociado a cada tweet, y teniendo en cuenta la metodología de análisis de sentimiento, se llevó a cabo la realización de un score de tweets que permitió ya no una clasificación individual de sentimientos por palabras sino del tweet completo. Para ello, se sumariza el puntaje obtenido por cada palabra en cada tweet. Si este resulta menor que cero, se considera que el tweet puede ser clasificado de manera negativa, mientras que si resulta mayor a cero se considera de manera

positiva. En caso de sumar cero, se considera que su contenido es neutral, es decir no pudo determinarse en forma polarizada cual es el sentimiento contenido en el tweet.

Figura 6: Score de Tweets en función del sentimiento asociado.



Fuente: Elaboración propia con RStudio

De la Figura 6 se puede observar que la cantidad de tweets clasificados como negativos es muy superior a los clasificados como positivos. A su vez, aproximadamente la mitad de la totalidad de los tweets no pudo ser clasificada de manera polarizada resultando en neutral la expresión del sentimiento de su contenido. Específicamente se pudo establecer que el 49,8% (1.476) de los 2.965 tweets analizados resultan ser negativos en cuanto a su contenido. Mientras que el 9,6% únicamente resulta con sentimiento positivo, y el 40,6% resulta con un sentimiento neutral.

De esta manera, en función de los resultados obtenidos a partir de la implementación de las diferentes técnicas de minería de texto, fue posible determinar una distribución sobre el sentimiento contenido en los tweets que poseen una etiqueta "cuarentenaobligatoria". De esta manera se logra mostrar que la utilización de técnicas de minería de texto permite la generación de valor agregado a partir de obtener información sobre datos alternativos, para la toma de decisiones.

CONCLUSIÓN

A lo largo del presente trabajo se ha demostrado cómo llevar a cabo el procesamiento de un volumen considerable de tweets, y la implementación de técnicas de minería de datos. Como resultado se pudo arribar a la detección de un sentimiento negativo respecto del aislamiento social obligatorio, en el 49,8% de los tweets analizados. En contraposición, solo el 9,6% contiene un sentimiento positivo. Para el 40,6% restante no logra identificarse un sentimiento polarizado. De esta manera se logró mostrar que la utilización de técnicas de minería de texto permite la generación de valor agregado a partir de obtener información sobre datos alternativos, para la toma de decisiones.

De esta manera, se considera que el presente trabajo logra sentar las bases de un aporte para las distintas organizaciones que se encuentren interesadas en llevar adelante la incorporación de datos alternativos con el objetivo de mejorar su proceso de toma de decisiones. En este sentido, los desafíos pueden resultar diversos, implicando desde un cambio cultural en la organización hasta un mayor requerimiento de especialistas en los diferentes eslabones de la cadena de implementación. No obstante, si el interés es poder generar un mayor valor agregado para optimizar las decisiones a tomar, se considera que puede ser un buen comienzo en la búsqueda de tal objetivo.

Por otra parte, resulta de interés llevar adelante una ampliación de este trabajo, que permita realizar una implementación más profunda de otros métodos de minería de texto que sean complementarios o innovadores para ampliar la obtención de resultados, así como del conocimiento. En este sentido, la implementación de la metodología de *Latent Dirichlet Allocation* puede resultar adecuada. También la búsqueda de predicción a partir de llevar a cabo una implementación de aprendizaje automático, podría ser un objetivo interesante, en la medida que permitiría establecer un modelo automático de detección de sentimientos.

REFERENCIAS BIBLIOGRAFICAS

Chen, H. M., & Franks, P. C. (2016). Exploring Government Uses of Social Media through Twitter Sentiment Analysis. Journal of Digital Information Management, 14(5).

Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Big Data. International Journal of Computer Trends and Technology, 38(1), 46-50.

Fiore, V., Almodovar, K., Assoumou, A., Dutta, D., & Cotoranu, A. (2017). The Correlation between the Topic and Emotion of Tweets through Machine Learning. Seidenberg School of CSIS. Pace University, Pleasantville, New York

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

Kabir, A. I., Karim, R., Newaz, S., & Hossain, M. I. (2018). The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R. Informatica Economica, 22(1).

Kao, A., & Poteet, S. R. (Eds.). (2007). Natural language processing and text mining. Springer Science & Business Media.

Kolanovic, M., & Krishnamachari, R. T. (2017). Big data and AI strategies: Machine learning and alternative data approach to investing. JP Morgan Global Quantitative & Derivatives Strategy Report.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Preoţiuc-Pietro, D., Lampos, V., & Aletras, N. (2015, July). An analysis of the user occupational class through Twitter content. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 1754-1764).

Russom, P. (2015). Modernización de la integración de datos para dar cabida a los nuevos requisitos de negocio y Big Data. TDWI, tdwi.org.

Schmarzo, B. (2013). Big Data: Understanding how data powers big business. John Wiley & Sons.

Verspoor, Karin & Cohen, Kevin. (2013). Natural Language Processing. 10.1007/978-1-4419-9863-7_158.

SOFTWARE: RSTUDIO

R Score Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/,

RStudio (Febrero 2011), AGPL v3, Northern Ave, Boston, https://www.rstudio.com/

PAQUETES UTILIZADOS:

Package 'broom' (Abril 2019) "Convert Statistical Analysis Objects into Tidy Tibbles" David Robinson [aut] y otros. https://cran.r-project.org/web/packages/broom/index.html

Package 'dplyr' (Mayo 2008), "A Grammar of Data Manipulation", MIT + file LICENSE, https://cran.r-project.org/web/packages/dplyr/index.html

Package 'factoextra' (Agosto 2017) "Extract and Visualize the Results of Multivariate Data Analyses" Alboukadel Kassambara [aut, cre], Fabian Mundt [aut] https://cran.r-project.org/web/packages/factoextra/index.html

Package 'ggplot2' (Agosto 2019) "Create Elegant Data Visualisations Using the Grammar of Graphics" r Hadley Wickham [aut, cre], Winston Chang [aut] y otros. https://cran.r-project.org/web/packages/ggplot2/index.html

Package 'gridGraphics' (Mayo 2019) "Redraw Base Graphics Using 'grid' Graphics" Paul Murrell [cre, aut], Zhijian Wen [aut] https://cran.r-project.org/web/packages/gridGraphics/index.html

Package 'gridExtra' (Septiembre 2017) "Miscellaneous Functions for ``Grid'' Graphics", Baptiste Auguie [aut, cre], Anton Antonov [ctb] https://cran.r-project.org/web/packages/gridExtra/index.html

Package 'lubridate' (Abril 2018) "Make Dealing with Dates a Little Easier" Vitalie Spinu [aut, cre], Garrett Grolemund [aut], Hadley Wickham [aut], Ian Lyttle [ctb], Imanuel Constigan [ctb], Jason Law [ctb], Doug Mitarotonda [ctb], Joseph Larmarange [ctb], Jonathan Boiser [ctb], Chel Hee Lee [ctb] http://lubridate.tidyverse.org

Package 'RColorBrewer'(Febrero 2015) "ColorBrewer Palettes" Erich Neuwirth [aut, cre] https://cran.r-project.org/web/packages/RColorBrewer/index.html

Package 'scales' (Agosto 2018) "Scale Functions for Visualization" Hadley Wickham [aut, cre], RStudio [cph] https://cran.r-project.org/web/packages/scales/index.html

Package 'stringr' (Febrero, 2019), "Simple, Consistent Wrappers for Common String Operations", Hadley Wickham [aut, cre, cph], RStudio [cph, fnd], https://cran.r-project.org/web/packages/stringr/index.html

Package 'tidyr' (Septiembre 2019) "Tidy Messy Data" MIT + file LICENSE URL https://cran.r-project.org/web/packages/tidyr/index.html

Package 'tidytext' (Octubre 2018), "Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools", MIT + file LICENSE, https://cran.r-project.org/web/packages/tidytext/index.html

Package 'tidyverse' (Noviembre 2017) " Easily Install and Load the Tidyverse" Hadley Wickham [aut, cre], RStudio [cph, fnd] https://cran.r-project.org/web/packages/tidyverse/index.html

Package 'tm' (Diciembre 2018) "Text Mining Package" Ingo Feinerer [aut, cre], Kurt Hornik [aut], Artifex Software, Inc. [ctb, cph] https://cran.r-project.org/web/packages/tm/index.html

Package 'tokenizers'(Marzo 2018) "Fast, Consistent Tokenization of Natural Language Text" MIT + file LICENSE, https://cran.r-project.org/web/packages/tokenizers/index.html

Package 'twitteR' (Agosto 2016), "Provides an interface to the Twitter web API", Jeff Gentry, https://cran.r-project.org/web/packages/twitteR/index.html

Package 'topicmodels' (Diciembre 2018) "Topic Models" Bettina Grün [aut, cre], Kurt Hornik [aut] https://cran.r-project.org/web/packages/topicmodels/index.html

Package 'widyr' (Septiembre 2019) "Widen, Process, then Re-Tidy Data" David Robinson [aut, cre], Kanishka Misra [ctb] https://cran.r-project.org/web/packages/widyr/index.html

Package 'wordcloud2' (Enero 2018), "Word Clouds 2", Dawei Lang and Guan-tin Chien, https://cran.r-project.org/web/packages/wordcloud2/index.html