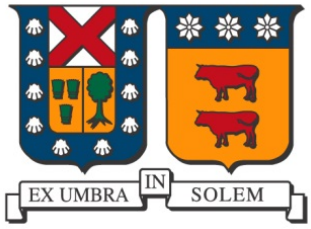
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
CIUDAD - CHILE



“DISEÑO DE UNA GUÍA DE CONDUCTAS PARA LA VIRALIZACIÓN DE ALTO IMPACTO DE CONTENIDOS EN TWITTER”

CARLOS ALBERTO ANDRADE CABELLO

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: José Luis Martí Lara
Profesor Correferente: ???

Septiembre - 2019

DEDICATORIA

A San Marcelino, a quien inicialmente de broma me encomendé en los peores momentos de mi proceso universitario, para luego ser una de mis cábalas más importantes en la vida.

AGRADECIMIENTOS

Quiero agradecer a mi familia por todo el apoyo que han significado en mi vida, mi madre que siempre tiene palabras de ánimo, mi padre por inspirarme a hacer preguntas y mi hermana por siempre tener un buen concejo.

También quiero agradecer a mis amigos que siempre me han ayudado en los momentos de mayor dificultad, en especial a Alfredo, quien ha sido una de mis más grandes fuentes de conocimientos a lo largo de toda la carrera.

Quiero destacar también todo el apoyo recibido por el profesor José Luís Martí, quien durante estos 7 años me ayudó a mantener la calma y ver el final de este camino con optimismo.

Finalmente quiero agradecer la fuerza emocional sobrehumana que he logrado a lo largo de este proceso universitario, puesto que sin ella, no habría llegado hasta aquí.

RESUMEN

Resumen—

En este trabajo se realizaron diversos análisis sobre publicaciones en Twitter, para los cuales se definieron dimensiones y métricas de estudio, se ejecutaron diversas técnicas de análisis y se asociaron estos resultados en búsqueda de patrones que mejoren el alcance de estas publicaciones, los cuales finalmente permitieron la construcción de una guía de conductas para la viralización de alto impacto de contenidos en Twitter.

Palabras Clave— Análisis de texto; Minería de datos; Redes sociales; Twitter; Viralización de contenidos.

ABSTRACT

Abstract—

In this work, several analyzes were carried out on Twitter publications, for which dimensions and study attributes were defined, various analysis techniques were executed and these results were associated in search of the best behaviors in the publication of contents, which finally allowed the construction of a behavior guide for the viralization of content on Twitter.

Keywords— Text analysis, data mining, social networks, Twitter, Content viralization

GLOSARIO

Contenido Viral: unidad de información que se ha propagado de forma rápida a una cantidad muy grande de usuarios.

Cronología: cuando un usuario sigue a otro, este se suscribe al contenido generado de la persona a seguir. Los tweets se presentan en la página inicial de twitter de cada usuario, a esta colección ordenada de forma cronológica se llama Cronología o timeline (TL) en inglés.

Entidad: a veces las cuentas de Twitter no pertenecen a una persona, sino que a una empresa, organización u otro tipo, por lo que se define como entidad a todo tipo de cuenta que no sea personal.

Follower o seguidor: cuenta suscrita a los contenidos de otra, también es llamado Follower, por su nombre en inglés.

Friend o Seguido: cuenta a la cual se encuentra suscrita otra cuenta.

Influencer: cuenta de interés para una gran cantidad de usuarios, estos pueden influenciar a través de lo que comunica o realiza.

***Like* o me gusta**: anteriormente llamado *Favorite* o favorito es una funcionalidad presente en cada tweet, en el cual el usuario puede demostrar afinidad con el contenido presentado.

Red Social: sitio Web, aplicación o cualquier plataforma que utilice internet para conectar a personas a través de diversas dinámicas, como lo son el compartir y consumir contenido generado por otros usuarios, mensajería, etc.

Respuesta: es un tweet escrito a modo de respuesta a otro tweet, generando dialogo entre los participantes.

Retweet: abreviado como RT, es la forma de compartir contenido generado por otro usuario de la plataforma, al hacer RT, los seguidores de la persona que realiza el retweet verán en su cronología el tweet original.

***Thread* o hilo**: funcionalidad de la plataforma Twitter, la cual permite escribir varios Tweets que se referenciarán de forma secuencial, permitiendo visualizar contenido extenso.

Tweet: contenido generado por lo usuarios de la plataforma Twitter, se basa en un texto limitado a 280 caracteres. Estos pueden poseer imágenes, videos, enlaces, entre otros.

Twittear: acción de generar un Tweet.

Twitter: red social de Microblogging en la cual los usuarios registrados comparten contenidos llamados Tweets.

Viralización: fenómeno en el cual los contenidos se propagan de forma rápida e independiente, sin publicidad ni marketing, llegando de forma exponencial a nuevas personas.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

Plantearse el desafío de ingresar a un mundo nuevo siempre es abrumador, la existencia de conocimientos que inicialmente no se poseen puede llegar a preocupar a quien desee participar de éste. Ya sea ingresar a una saga de juegos o libros que posea un trasfondo creado a lo largo de años de entregas distintas, irse a vivir a un país nuevo que posee costumbres y tradiciones que pueden llegar a ser completamente diferentes a las que se poseen, o incluso registrarse en una red social y no saber nada de ella, solo que es popular entre las personas que uno considera importantes. En todos estos casos no saber qué conductas a seguir puede llegar a ser un desafío muy interesante, pero a la vez agobiante.

Actualmente la sociedad está cada día más interconectada, la participación en las redes sociales aumenta de forma sostenida, siendo una de estas Twitter, la cual permite compartir contenidos de texto, imágenes y otros tipos entre las personas que decidan seguir al usuario que los publica. Este fenómeno no ha pasado inadvertido en distintas empresas que buscan aumentar el alcance de sus productos pero que no poseen conocimientos de cómo proceder en este mundo, por lo que una guía de conductas para la viralización de éstos llega a ser necesaria.

Para realizar un proyecto como es el de una guía de conductas para la viralización es necesario conocer qué comportamientos son los que generan un impacto positivo en el alcance de los contenidos, en conjunto con analizar cómo se relacionan éstos en el mundo en el que están insertos.

Para estudiar estas publicaciones y reconocer ciertos patrones se deben poseer conocimientos de análisis de texto, puesto que la mayoría de los contenidos son en este formato. También son necesarias técnicas y herramientas que permitan la asociación de los diversos atributos que se vayan descubriendo y que influencian el alcance de estos contenidos.

Este trabajo esta compuesto por la definición del problema, presentación de los objetivos y alcances, los 3 presentes en el capítulo 1; En el capítulo 2 se encuentra el marco conceptual donde se presentará un breve análisis de la plataforma, se definirán las dimensiones del contenido a estudiar, se analizarán metodologías existentes para trabajos similares a éste y se presentarán posibles métricas y técnicas de análisis; Luego en el capítulo 3 se presentará la propuesta de solución para el trabajo a realizar, aquí se presentará la metodología propuesta, la cual define la recolección, almacenamiento y manipulación de los datos, la elaboración de métricas, el trabajo a realizar con las dimensiones y técnicas de estudio; luego se presentará el desarrollo y la validación de la solución, en donde se presentarán los resultados del trabajo, el análisis para cada métrica y dimensión, en conjunto con el diseño de la guía de conductas, todo esto conforma el capítulo 4; Finalmente el trabajo será cerrado a través de una conclusión que recalcará el trabajo realizado, los conocimientos que fueron necesarios para la realización de este y algunas recomendaciones e ideas para trabajos futuros.

CAPÍTULO 1
DEFINICIÓN DEL PROBLEMA

En este capítulo se define el contexto y el problema a resolver, además de presentar el objetivo general y los objetivos específicos; el capitulo finaliza definiendo el alcance del trabajo.

1.1. Contexto del problema

En la actualidad la sociedad chilena se encuentra hiperconectada, con una penetración del 71,7 % de personas con conexión a internet, de las cuales, el 94 % se conecta principalmente a través de un dispositivo móvil [1]. Estas personas utilizan en promedio 17 aplicaciones, siendo en su mayoría redes sociales. Lo anterior convierte a las redes sociales en uno de los mercados más interesantes para difundir contenidos, debido al alcance y penetración que presentan en Chile.

Una de las redes sociales más utilizadas es la plataforma de microblogging Twitter, en la cual se pueden compartir diversos contenidos en publicaciones llamadas Tweet, la cual posee un límite de 280 caracteres (anteriormente 140).

El 2017 fue estrenada una funcionalidad en Twitter llamada "*Threads*" o hilos, con la cual es posible escribir diversos Tweets seguidos, los cuales se verán agrupados. Esta funcionalidad es muy útil cuando la información que se desea compartir es mucho más extensa como para ser presentada en un único Tweet.

Gracias a la nueva funcionalidad presentada, es posible difundir contenidos que anteriormente eran más difíciles de viralizar debido a su extensión, lo que ha generado un gran interés para diversas entidades que buscan expandir sus receptores de información (ya sea clientes, adherentes políticos u otros, según sea quien difunda el contenido).

Considerando lo interesante del mercado presentado, en conjunto con las nuevas funcionalidades introducidas en las diversas plataformas sociales, es que empresas, organizaciones, entidades de gobierno e incluso personalidades públicas están invirtiendo en personal (Community Manager) para posicionar entre los usuarios su marca o contenido asociado, quienes actualmente puede que posean o no los conocimientos necesarios para la difusión de contenidos.

Al existir dineros invertidos, es de gran interés poseer conocimientos específicos para generar un mayor impacto y viralización de los contenidos desarrollados, por lo que se reconoce

[1] Chile lidera la penetración de internet en la región y el smartphone continúa siendo el favorito Fuente: Emol.com - https://www.emol.com/noticias/Tecnologia/2017/05/04/856853/Chile-lidera-la-penetracion-de-internet-en-la-region-y-el-smartphone-continua-siendo-el-favorito.html

como problema el no existir una guía o manual de buenas prácticas para difundir contenido a través de Twitter.

Finalmente, se reconoce que una guía de buenas prácticas para la viralización de contenidos aporta un gran valor para las diversas entidades que desean posicionar productos, marcas o campañas en la plataforma social, puesto que optimizaría los recursos invertidos para este objetivo.

1.2. Identificación del problema

Se identifica como problema la dificultad de popularizar contenidos en una plataforma tan utilizada y saturada de información como lo es Twitter; debido a esto, se busca una forma de maximizar su ***impacto***, por lo que se desea analizar una gran cantidad de Tweets en búsqueda de características que generen interés y viralización de éstos.

Puesto que se busca estudiar el texto compartido según diversas técnicas de análisis de texto plano, se reconoce que la corta extensión de éstos pueda llegar a afectar de gran manera la calidad de los resultados, por lo que se decide estudiar hilos de Tweets.

Para el estudio es necesario un dataset, por lo que se deben recolectar una vasta cantidad de hilos, estos deben ser almacenados y procesados de tal manera que permitan realizar diversos tipos de análisis sobre sus atributos obtenidos y generados.

Puesto que se busca popularizar contenidos, es de vital importancia hacer estudios sobre las características de la presentación de la información en sí, por lo que se deben realizar diversos análisis de texto, como lo son la detección de tópicos y de emociones.

En conjunto con todo lo anterior, es necesario destacar la necesidad de almacenar toda la información obtenida para que sea posible manipular los datos de forma rápida a través de código y herramientas de análisis, por lo que se debe definir una estructura y una técnica de almacenamiento para los datos óptima para el estudio.

1.3. Objetivos

El objetivo general de esta memoria consiste en: Diseñar y validar una guía de conductas para la viralización de alto impacto de contenidos en la plataforma Twitter a través de la funcionalidad de *Threads*, mediante el análisis de múltiples dimensiones de éstos.

Objetivos específicos

- Estudiar el impacto generado según el largo, la emoción y el tópico de los *Threads*, para identificar factores en común en los mensajes populares.
- Estudiar la relación entre el autor y el impacto de los *Threads*, para reconocer la importancia del autor en relación al impacto del contenido.
- Analizar la relación de Favoritos, RTs y respuestas del Tweet principal y los siguientes en un Thread, para estudiar el comportamiento viral de estos.
- Examinar y asociar relaciones descubiertas entre variables de estudio, para proponer buenas prácticas de difusión de contenido.

1.4. Alcances

Para el trabajo a realizar se analizarán los atributos de alrededor de 6.000 Tweets, agrupados en aproximadamente 400 *Threads*. Los documentos de estudio se encuentran escritos en inglés, algunos sólo poseen texto, mientras que otros poseen también elementos multimedia, los cuales no serán considerados.

De manera inicial se genera sólo un archivo en el cual se registra el ”id” del último Tweet de un hilo para luego rescatar el resto. Puesto que se busca realizar un análisis tanto del impacto del contenido, como la influencia en sí del usuario en la viralización del *Thread*, se debe también rescatar la información del emisor.

En el trabajo no se realizarán estudios del grafo de Followers de la red, ni tampoco análisis sobre Tweets que no sean parte del *dataset* inicial.

CAPÍTULO 2
MARCO CONCEPTUAL

En este capitulo se presenta el marco conceptual que sirve como base para el trabajo a desarrollar. Se reconoce la plataforma fuente de los documentos a estudiar, se definen las dimensiones a analizar de los documentos y, finalmente, se exploran las metodologías de trabajo existentes para el análisis de Twitter.

2.1. Plataforma Twitter

Twitter es una plataforma de *microblogging* desarrollada por Jack Dorsey el 2006, en la cual es posible compartir contenido, ya sea texto, videos e imágenes, entre otros. Los usuarios utilizan la plataforma para conversar sobre diversos temas, como por ejemplo sus pensamientos a lo largo del día, noticias que encuentra interesantes o comentarios sobre política o programas de televisión y cualquier otro tema que consideren contingentes.

Figura 1: Ejemplo de un *timeline* a la izquierda y ejemplo de un perfil público a la derecha. Fuente: Aplicación de Twitter para sistema operativo Android.

Al Twittear (acción de compartir contenidos), cada usuario genera contenido en su propia cuenta, por lo que para consumir el contenido de otra cuenta, los usuarios deben seguir a aquella de interés, lo que conlleva a que el contenido de ésta aparezca en el inicio de Twitter del usuario que realizó el seguimiento. Esta acción es dirigida, por lo que se genera la distinción entre seguidores y seguidos.

Los contenidos suscritos se presentan en el inicio de la plataforma; a esto se le llama cronología o *Timeline (TL)*. Un ejemplo de esta aparece en la imagen izquierda de la figura 1. En dicha vista, los usuarios pueden ver los contenidos generados por las cuentas que sigue el usuario.

Una cuenta puede seguir a muchos usuarios y tener muy pocos seguidores, lo que significa que ésta consuma mucho contenido pero su contenido no sea consumido. El caso contrario es cuando una cuenta posee muchos seguidores pero muy pocos seguidos; a esto se le llama una cuenta popular entre usuarios o *influencers*. En la imagen de la derecha de la figura 1 se puede observar un ejemplo de una cuenta popular.

Cuando una cuenta desea compartir contenidos, es posible que un solo Tweet no sea suficiente, por lo que deberá crear varios de forma seguida. Teniendo esto en cuenta, Twitter posee una funcionalidad llamada hilos o *Threads*, con la cual es posible redactar varios Tweets de forma continua, los cuales luego serán asociados unos con otros para así poder ser mostrados de forma consecutiva. Un ejemplo de esto se puede observar en la figura 2.

En diversos trabajos basados en información obtenida de Twitter son considerados varios valores presentes en la plataforma, los que pueden provenir de la cuenta o del Tweet a estudiar. Los valores más típicos a considerar de una cuenta son las listas y cantidad de seguidores y seguidos, la cantidad de Tweets y la fecha de registro. En relación a un Tweet, los valores más comunes a la hora de realizar un análisis son la lista y cantidad de *hashtags*, respuestas, Retweets, Likes y fecha de publicación, entre otros.

Figura 2: Ejemplo de *Thread*
Fuente: Twitter para Android.

2.2. Dimensiones de los documentos procedentes de Twitter

Según el problema inicial, se pueden definir 3 ejes claramente distinguibles en relación a los contenidos en la red social Twitter: el primero son los tópicos o temas tratados en uno o más Tweets, el segundo son las emociones con las que están redactadas los Tweets, y el tercero es la influencia de la cuenta generadora de contenidos.

2.2.1. Tópicos

La detección de tópicos en el análisis de lenguaje natural (NLP) hace referencia al estudio de colecciones de documentos para reconocer ciertos ”tópicos” o temas principales que se puedan estar tratando, pudiendo organizar colecciones según los temas descubiertos [Blei, 2012].

Una de las técnicas utilizadas en la detección de tópicos es *Latent Dirichlet Allocation* (LDA), el cual es un modelo estadístico generativo que aprende los tópicos latentes presentes en una colección de documentos [Petrović *et al.*, 2010], es decir, modela temas en común en diversos documentos a través de análisis estadísticos que estudian la ocurrencia de palabras.

Para el caso de estudio, es posible considerar la colección de documentos como el conjunto de todos los Tweets pertenecientes a cada *Thread*. Es factible también considerar cada hilo estudiado como un solo documento, lo que hace posible realizar una detección de tópicos más precisa, esto debido a que es poco probable que un Tweet sea lo suficientemente complejo como para detectar tópicos coherentes debido a que las limitaciones presentadas por la plataforma condicionan la calidad del lenguaje utilizado [Roberts *et al.*, 2012], por lo que al aglomerar varios Tweets en un documento más grande, el documento resultante posee más palabras asociadas al tema tratado.

Existen muchas variantes de LDA para realizar detección de tópicos. En [Petrović *et al.*, 2010] se presenta una adaptación para detección de nuevos eventos, también conocido en inglés como *First Story Detection*; el algoritmo, a través de un *Stream* de Tweets, busca detectar la primera historia relacionada a un evento particular el cual, a través de *hashing*, analiza la similitud entre documentos, para así generar un conjunto de cada tema detectado. Cada tema tiene un límite de documentos asociados, por lo que si se debe agregar uno nuevo, se elimina el más antiguo.

Para complementar el trabajo, se genera una red de Tweets por cada tema, para luego analizar la velocidad de crecimiento y estudiar el interés e impacto generado por el tema, y así diferenciar los temas más relevantes.

Otra variante sobre LDA se encuentra en [Lau *et al.*, 2012], en el cual se busca algo similar al mencionado anteriormente, diferenciándose en el método de obtención y análisis de documentos. En esta variante se utiliza un vocabulario dinámico, enfocado en la clusterización

y análisis de co-ocurrencias más que sólo la frecuencia de términos. Este vocabulario se va actualizando en porciones de tiempo definidas, lo que permite mantener actualizados los tópicos. La forma de manejar el texto en relación al *Stream* de datos en [Lau *et al.*, 2012] se basa en ventanas de tiempo, lo que hace que los documentos analizados se dividan en cada ventana; esto se hace para mantener un tamaño constante de documentos analizados.

2.2.2. Emociones

La detección de emociones es parte de un área mayor llamada *Affective Computing*, la cual busca que los computadores sean capaces de detectar y expresar emociones humanas [Canales y Martínez-Barco, 2014].

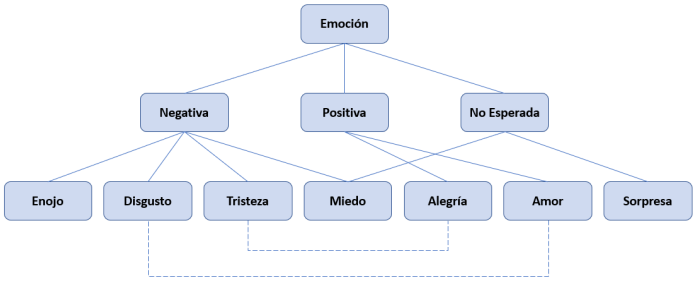


Figura 3: Relación entre las 6 emociones básicas además de amor; las lineas sólidas representan herencia, mientras que las punteadas sentimientos contrarios.
Fuente: [Roberts *et al.*, 2012], traducción propia.

Es posible clasificar las emociones a través de diversos modelos, poseyendo éstos mayor o menor nivel de especificación en sus categorías. Dependiendo del estudio realizado podrán ser clasificadas en distintas clases, como se observa en la figura 3. Por ejemplo, con *Sentiment Analysis* se obtiene un rango para los sentimientos entre Positivo, Negativo o Neutro, mientras que otras categorías se basan en estudios psicológicos, los cuales reconocen varias emociones básicas: Enojo, Disgusto, Miedo, Felicidad y Sorpresa [Canales y Martínez-Barco, 2014]. Vale decir que éstas emociones también pueden ser categorizadas en los sentimientos descritos, usando un modelo como el presentado en la figura 3.

Para la detección de emociones se reconocen 3 posibles tipos de métodos para realizarlo: basados en léxicos, basados en máquinas de aprendizaje supervisadas y aquellos basados en máquinas de aprendizajes no supervisadas, cada uno de los cuales con ventajas y desventajas [Canales y Martínez-Barco, 2014].

Es posible construir un *corpus* para la detección de emociones presentes en Tweets a través de diversas técnicas, tales como extracción y etiquetado de *hashtags*, detección de tópicos previamente etiquetados según emociones y máquinas de aprendizajes guiadas

con Tweets previamente etiquetados, entre otros [Hasan *et al.*, 2014], [Mohammad, 2012], [Roberts *et al.*, 2012].

Para cada tipo de modelo de detección se presentan diversos problemas [Hasan *et al.*, 2014], ya sea por el lenguaje utilizado, que genera problemas a las técnicas basadas por léxicos; la falta de etiquetas por parte del texto a analizar, que no permite entrenar máquinas de aprendizaje; e incluso la gran cantidad de temas tratados en Twitter, que genera un gran número de potenciales tópicos, los que presentados en forma de vectores generaría una gran cantidad de valores cero para un Tweet específico en técnicas de análisis basados en estos. Ejemplos de dos modelos de reconocimiento son presentados en la figuras 4 y 5.

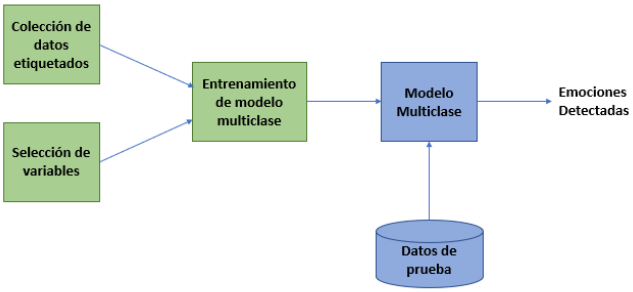


Figura 4: Ejemplo de modelo de aprendizaje supervisado para la detección de emociones en mensajes. Fuente: EmoTex [Mohammad, 2012], traducción propia.

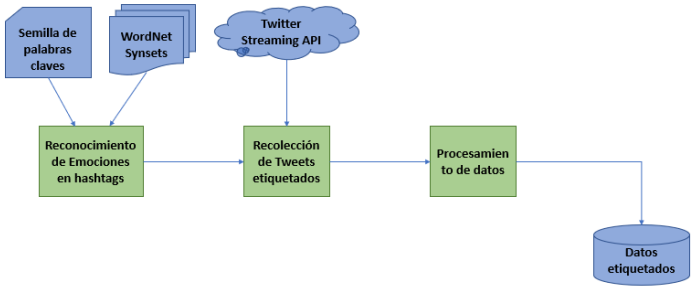


Figura 5: Ejemplo de modelo para la generación de una colección de datos etiquetados. Fuente: EmoTex [Mohammad, 2012], traducción propia.

2.2.3. Influencia de la cuenta generadora de contenido

Además de analizar el contenido de los Tweets en sí, también es necesario estudiar el impacto que genera. Actualmente existen varias formas de interactuar con el contenido en Twitter: es posible responder, compartir en forma de *Retweet* (RT) y poner *Like* a los Tweets (anteriormente llamado Favorito), siendo cada una de estas acciones muy distinta en la intención con la que se hacen.

Una dimensión que hay que tomar en cuenta al momento de analizar el impacto que tiene un Tweet es la calidad e interés que genera la misma cuenta que crea el contenido, puesto que dependiendo del interés que éste posea, distinto será el comportamiento de las reacciones de los Tweets [Cha *et al.*, 2010].

En la literatura se reconocen principalmente 3 tipos de influencias que pueda generar un usuario [Cha *et al.*, 2010]: popularidad (*Indegree Influence* según [Garcia *et al.*, 2017]), la cual se refiere a la cantidad de seguidores que posea el usuario; *Retweet Influence*, la que medida según la cantidad de RTs que reciben sus publicaciones, indica la habilidad del usuario de generar contenidos con valor de ser compartidos; y *Mention Influence*, la cual indica el interés generado por el usuario para que otros deseen 'conversar' con él. Además de los tipos anteriores, otros autores definen una cuarta influencia, *Active Account*, que considera el nivel de actividad que presenta una cuenta en la plataforma [Drakopoulos *et al.*, 2017].

Para cada tipo de influencia se reconocen ciertos perfiles que generan mayor tipo de interés [Cha *et al.*, 2010]. Para *Indegree Influence* existe una gran variedad de tipos de usuarios, siendo la mayoría canales de noticias, políticos, famosos y celebridades, mientras que para *Retweet Influence* existe predominancia de cuentas de noticias y otros contenidos, como por ejemplo memes y videos. Finalmente se reconoce que las cuentas que generan mayor interés en el ámbito de *Mention Influence* son, en su mayoría, cuentas de celebridades.

Es posible también reconocer qué tan transversal es la influencia de las cuentas populares en cada una de las 3 categorías de interés. Se reconoce en la literatura que las cuentas populares en RT y menciones poseen una correlación no despreciable, tal como se puede observar en la figura 6, además de que las cuentas más importantes poseen un impacto elevado en diversos tópicos, mientras que cuentas menos populares poseen popularidad en temas más acotados [Cha *et al.*, 2010].

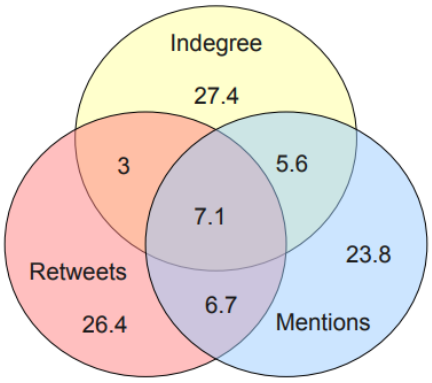


Figura 6: Diagrama de Venn de los top 100 de cada categoría de influencia, los valores se encuentran normalizados para que sumen 100 %.
Fuente: Measuring user influence in Twitter. [Cha *et al.*, 2010].

2.3. Metodologías de trabajo para el análisis de Twitter.

Es necesario considerar una metodología para el trabajo a realizar, por lo que en este apartado se analizan diversos documentos que realizan estudios de contenidos obtenidos desde Twitter. Se presentan también las métricas más relevantes encontradas, además de diversas alternativas para el almacenamiento de los datos a estudiar. El capítulo finaliza con la presentación de algunas técnicas de minería de datos más utilizadas en la actualidad para el análisis inteligente de información.

2.3.1. Metodologías existentes

Debido a que busca estudiar diversas metodologías, se abordarán 3 textos que realizan análisis sobre datos extraídos desde Twitter; en cada caso, se presenta brevemente el trabajo, los pasos que realizan y se exponen pros y contra de la metodología usada.

El primer texto a analizar es [Eysenbach, 2011], el cual busca estudiar la factibilidad de predecir la cantidad de citaciones de *papers* según las menciones en Twitter. El estudio realizado se basa en el número de veces que son compartidos las publicaciones, además considera las citaciones en otros trabajos. El primer valor proviene de una base de datos de la *Journal of Medical Internet Research*, mientras que la cantidad de citaciones son obtenidas desde Google Scholar y Scopus.

El trabajo realizado considera la extracción de la cantidad de citaciones que posee una publicación, la elaboración de diversas medidas en base a la cantidad de menciones en diversos periodos de tiempo, el análisis de distribución de las medidas presentadas, principalmente la cantidad de citaciones por *paper* y la cantidad de citaciones por periodos de tiempo, entre otras. Además del análisis de distribución, realiza análisis de correlación entre la cantidad de menciones y las citaciones en cada plataforma (Google y Scopus).

Después del estudio de los datos obtenidos, realiza diversas regresiones lineales con el fin de encontrar la relación entre citaciones y menciones; además de esto, define diversas métricas para evaluar la calidad de las regresiones, como lo son *tweetations* (Cantidad de menciones), *twimpact* (Cantidad de menciones por cierto periodo de tiempo) y *twindex* (Posición entre trabajos según cantidad de menciones).

Finalmente, realiza un análisis sobre las limitaciones de los datos, las mediciones y el estudio en sí.

Uno de los puntos más importantes a rescatar de este trabajo son las técnicas utilizadas para estudiar dos dimensiones que de manera inicial no se encontraban relacionadas, además de las métricas definidas, las cuales pueden ser extrapoladas para otro tipo de estudios.

El segundo trabajo de estudio es [Garcia *et al.*, 2017], el cual busca analizar el efecto que genera la popularidad y reputación en Twitter sobre la permanencia del usuario a través del tiempo. Los datos utilizados abarcan todos los usuarios existentes en la plataforma registrados hasta el 2009. En este *paper* se define la diferencia entre popularidad de los usuarios y la reputación de éstos, características que pueden ser medidas de diversas formas.

Este trabajo se basa tanto en el análisis del grafo de seguimiento de cuentas, como también en la actividad de cada usuario para definir la popularidad y reputación.

El estudio comienza definiendo el *dataset*, para luego presentar cada dimensión (Popularidad, reputación, influencia e inactividad); a continuación define las diversas métricas a utilizar. Después de esto, se realizan análisis de cada dimensión, para después estudiar la interacción entre la inactividad y las otras dos dimensiones por separado. Finalmente, se presenta una discusión sobre el trabajo realizado y destaca los resultados más concluyentes.

Es muy importante destacar este trabajo debido al método de análisis realizado sobre las dimensiones, por separado y en conjunto.

El tercer y último texto de estudio es [Drakopoulos *et al.*, 2017], en el cual se definen y evalúan métricas de influencia en Twitter. Este trabajo es mucho más teórico, por lo que los datos de estudio son presentados luego de las definiciones de medidas y métricas con las que se trabajan y no son presentados con mayor precisión que ”Tweets recolectados entre noviembre y diciembre de 2016 a través de la búsqueda de diversos *hashtags* universitarios”.

El trabajo comienza definiendo las medidas que serán consideradas; luego procede a definir las métricas, clasificándolas entre métricas de primer orden (directas de una o más medidas) y de segundo orden (indirectas, obtenidas a través del grafo de la red).

Luego de la presentación de las medidas y métricas, hace una breve exposición del *dataset*, para así proceder a definir y analizar diversas relaciones de distribución de influencia. A continuación, se evalúan la correlación y divergencia entre las definiciones anteriores, y luego la eficacia de cada distribución. Finalmente se realiza un estudio de *performance* de cada mecanismo de *ranking* presentado.

En resumen, este trabajo define diversas métricas de estudio, junto con un estudio exhaustivo de varias técnicas de análisis de influencia, de las cuales se presentan su precisión y *performance* en cálculo. Si bien el trabajo no busca estudiar un conjunto de datos, presenta una gran cantidad de técnicas de gran interés.

2.3.2. Etapas más comunes

De los documentos estudiados, se reconocen algunas etapas en común, aunque no todos poseen el mismo orden, ni tampoco la misma prioridad para cada una. Se agruparon las más similares entre sí para facilitar el orden de la presentación:

- **Recolección y Almacenamiento de datos:** es aquí donde los autores pueden presentar los datos en los cuales basan su trabajo y como serán almacenados. Dependiendo el objetivo del *paper* este fase será ubicada al principio del documento o luego de la teoría a aplicar.

  Algunos autores dividen esta en dos o más, mientras que otros las presentan inserta en otro punto de mayor relevancia para ellos.

- **Manipulación de datos y obtención de medidas y métricas:** es posible observar que los autores no siempre obtienen los datos de la forma en que ellos lo necesitan, por lo que en algunos trabajos se definen los pasos que deben realizar para poder estandarizar sus datos de la forma que les sean útiles.

  Los autores también pueden definir las medidas y métricas en las cuales se basarán sus estudios. Dependiendo la importancia que posea esta fase en el trabajo, la definición y cálculo de estos valores puede extenderse de forma considerable, llegando a ser 3 etapas por separado, mientras que en otros trabajos ocurren ambos casos de forma más acotada sin definir mayor diferencia.

- **Presentación y elaboración de técnicas de estudio:** esta etapa es la más diferente entre los documentos, puesto que es donde se presenta el trabajo principal del estudio. Es aquí donde se encuentran las definiciones de las técnicas de análisis de los datos y se obtienen modelos del fenómeno de estudio, entre otros pasos a desarrollar.

  En esta etapa los autores presentan el trabajo con el cual pretenden cumplir con los objetivos planteados, por lo que dependiendo de éstos es que ocurren diversas opciones para el desarrollos esta fase. En algunos casos los autores presentan primero todos los modelos, para luego así estudiar como se comportan las métricas. También ocurre que el autor genere las métricas y modelos en paralelo.

- **Análisis y comparación de resultados:** se observa que esta etapa es la más importante, puesto que aquí se evalúa la calidad de los resultados obtenidos. Dependiendo de lo realizado en la etapa anterior, se estudia la calidad de los modelos y la relación entre estos, además se realizan análisis sobre el comportamiento del fenómeno estudiado a través de los resultados obtenidos. También ocurre que dependiendo de los resultados del análisis realizado en esta etapa se proceda por realizar nuevamente alguna de las fases anteriores para mejorar la calidad de éstos.

- **Discusión y conclusiones:** se reconoce como etapa final la discusión y conclusiones del trabajo realizado; en ésta se busca destacar los resultados obtenidos, evaluar el por qué de la calidad de estos, en conjunto con postular ideas de trabajos a futuros que se basen en los resultados o cómo se debe proceder para realizar mejoras a éstas, entre otros.

  Dependiendo de diversos motivos, como por ejemplo los objetivos del trabajo, se posee un desarrollo más teórico en esta etapa, con una discusión enfocada principalmente a las condiciones óptimas para cada técnica desarrollada, en conjunto con estudios

de rendimiento en la obtención de resultados según variables internas y externas al trabajo realizado.

En conjunto con las ideas de trabajos futuros, en esta etapa también se pueden acotar los alcances de los resultados obtenidos, declarando cuales son las condiciones y variables que permitieron la calidad de los resultados, además de definir cuales son las etapas más importantes que se deban repetir si se desean realizar estudios similares.

2.3.3. Métricas de análisis

Es posible reconocer múltiples métricas de análisis, en relación a los Tweets y a los usuarios, en conjunto o por separado. En [Eysenbach, 2011], [Garcia *et al.*, 2017] y [Drakopoulos *et al.*, 2017] son definidas múltiples métricas de gran interés para el trabajo a realizar.

Según [Drakopoulos *et al.*, 2017] es posible definir métricas directas e indirectas. Las métricas directas se refieren a las que son calculables realizando un análisis de valores y medidas del usuario de forma aislada, esto quiere decir, considerar sólo cuales provienen de su perfil, ignorando el grafo de usuarios de la red social; mientras que las métricas indirectas estudian diversos valores en conjunto con la posición que representan los usuarios en el grafo de estudio.

La mayoría de las métricas de estudio en [Drakopoulos *et al.*, 2017] consideran valores directos como la cantidad de Tweets (TW), Retweets (RT), hashtags usados (HT), seguidores (FWR), seguidos (FWE), Likes (FAV), respuestas (RES), menciones (MEN) y frecuencia (FR) de actividad. Mientras que algunas métricas indirectas consideran valores de centralidad de nodos calculados por diversas metodologías.

Tres de las métricas más interesantes presentadas en [Drakopoulos *et al.*, 2017] son la **influencia conversacional**, la cual considera la mayor cantidad de valores directos, (cuya formula se puede observar en la ecuación 1), la actividad promedio del usuario basada en FR, y el puntaje ponderado del usuario (presentado en la ecuación 2).

$$Cv = TW + RT + MEN + FAV + RES \tag{1}$$

$$Pp = TW \cdot RT \cdot HT \cdot log(1 + FWR)^{\frac{1}{4}} \tag{2}$$

Otras métricas interesantes de destacar son la popularidad y la reputación, presentadas en [Garcia *et al.*, 2017]. La popularidad puede ser definida como la cantidad total de FWR, mientras que la reputación se define basándose en la distribución de un grupo de usuarios según la proporción entre FWR y FWE, presentando dos métodos de obtención: el primero a través del cálculo de *incoreness* (medida recursiva de centralidad del nodo sobre el

grafo), mientras que el segundo mediante el análisis de la estructura *bowtie* que presenta la red (distribución del grafo donde la mayoría de los nodos se presenta al centro de éste, mientras que el resto se encuentra en uno de 2 extremos opuestos, tal como es posible ver en la figura 7).

En la estructura presentada en la figura 7 se estudia la distribución de todos los usuarios de Twitter registrados hasta el 2009, En el sector azul llamado *SCC* se presenta el grupo más fuertemente conectado, en donde la mayoría de las cuentas se siguen entre ellas. En el sector rojo se encuentra el grupo llamado *Out*, en donde las cuentas presentan una gran cantidad de conexiones hacia ellos (FWR), pero una baja conexión al sector azul (FWE). En el sector verde se presenta el grupo *In* el cual posee una gran cantidad de conexiones hacia el sector azul (FWE) y una baja conexión desde él (FWR). Un cálculo aproximado de esta distribución se puede realizar a través de una función sigmoídea (cuya fórmula es presentada en la ecuación 3)

$$P_s = \frac{log(1 + FWR)}{1 + log(1 + FWE)} \tag{3}$$

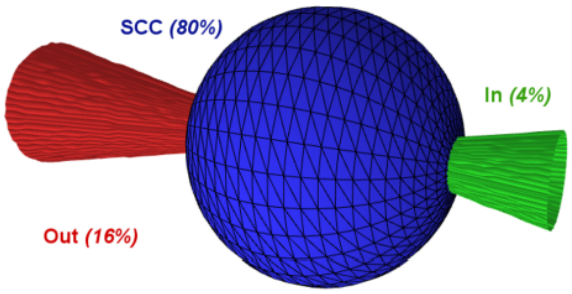


Figura 7: Representación de la distribución de la relación entre FWR y FWE de todos los usuarios en Twitter hasta el 2009.
Fuente: Understanding Popularity, Reputation and Social Influence in the Twitter Society. [Garcia *et al.*, 2017].

Un tercer conjunto de métricas destacables son las presentadas en [Eysenbach, 2011], siendo las más relevantes para este estudio Twimpact y Twindex. Los autores definen como Twimpact la cantidad de MEN recibidas por un tema por día, siendo posible definir diversos periodos de tiempo en el cual se estudiará este valor. Twindex, por su parte, es definido como la posición relativa del tema entre los demás según alguna métrica de estudio, cuyo valor se mueve entre 0 y 100.

2.4. Almacenamiento de datos

Los datos sobre los cuales se realizarán los análisis provienen de Twitter, la cual presenta su propia API para la extracción de información. Esta API posee varias limitaciones de cantidad y tamaño de consultas, haciendo necesario acotar estas consultas al mínimo, por lo que se busca no solicitar la misma información reiteradas veces, para lo cual se debe almacenar los datos que ya se hayan solicitado.

Existen diversas alternativas para almacenar información, por lo que se analizan dos; archivos y bases de datos.

Almacenamiento basado en archivos

El almacenamiento basado en archivos es una de las soluciones más simples para el almacenamiento de datos, puesto que no se debe montar ningún tipo de sistema o programa de administración. Debido a esto, es común ver problemas de pequeñas y medianas complejidades depender de un sistema de archivos para almacenar su información.

Algunas de las características del almacenamiento basado en archivos es el enfoque menos estructurado de datos, existiendo necesariamente una interrelación entre archivos. El almacenamiento en archivos es para problemas más generales que no necesiten una mayor optimización en las consultas y escrituras. No está diseñado para un alto consumo de los datos.

Existen diversos formatos populares para el almacenamiento de información basada en archivos, de los cuales 3 de los más comunes e importantes:

- **Comma Separated Values (CSV):** tal como se menciona en su nombre, son archivos que separan los valores o registros por coma, lo cual lo hace eficiente en espacio. Estos registros no poseen una mayor estructuración, por lo que los datos no tienen una relación entre ellos a no ser que se defina en el primer registro.

  Los registros realizados en este formato son de difícil lectura para los usuarios, además de ser poco óptimo para realizar operaciones más complejas que requieran cierta lógica de negocios.

- **Extensible Markup Language (XML):** los archivos XML están diseñados para almacenar información que posea relaciones jerárquicas, esto quiere decir, que los datos posean cierta dependencia entre los valores.

  Este tipo de archivo posee una estructura basada en etiquetas de la siguiente forma:

```
<etiqueta-1>
    <etiqueta-2>
        valor
    </etiqueta-2>
</etiqueta-1>
```

Debido a que cada valor posee una etiqueta de apertura y cierre, los archivos son de mayor tamaño, pero permiten una lectura simple para el usuario y una manipulación mucho más óptima para el procesamiento de datos.

- **Javascript Object Notation (JSON):** este tipo de archivo está basado en la estructura que Javascript maneja los objetos, lo que convierte a este tipo de archivo como uno de los más óptimos para los sistemas que dependan de este lenguaje.

```
{
    clave: valor,
    clave2: {
        clave3: valor,
        clave4: [1, a, valor, undefined],
        clave5: {
            clave6: valor
        }
    }
}
```

Además de los beneficios que trae tener un enfoque basado a Javascript, los datos también tienen la opción de tener estructuras jerárquicas como ocurre en XML, con el beneficio de ser mucho más liviano.

Almacenamiento en bases de datos

Al contrario del almacenamiento de datos basados en archivos, las bases de datos almacenan de una forma mucho más estructurada, los datos pertenecen al mismo contexto y son almacenados de forma sistemática.

Otra diferencia con el almacenamiento en archivos, es que las bases de datos poseen sistemas administradores de bases de datos, los cuales realizan las consultas y manipulación de todos los datos.

Dependiendo de la estructura y el manejo que se dé a los datos, pueden existir diversos tipos de bases de datos, por lo que a continuación se mencionan los 3 tipos más importantes para este trabajo.

- **Bases de datos relacionales:** estas bases de datos son las más populares; están basadas en registros con campos claramente definidos y rígidos, cada campo es una columna y cada registro es una fila que posee valores en cada una (dependiendo de las reglas de negocio, es posible la existencia de algunos valores vacíos o nulos). A nivel de definición de las bases de datos relacionales, no existe una relación directa en el orden que poseen los registros. La gran mayoría de los sistemas de bases de datos relacionales basan sus consultas en lenguaje SQL. Dos de las bases de datos más populares para esta categoría son ORACLE y MySQL.
- **Bases de datos orientada a documentos:** están diseñadas para trabajar con archivos JSON y XML. Estos sistemas de bases de datos son relativamente nuevos y buscan entregar soluciones flexibles en las cuales las bases de datos relacionales son demasiado estructuradas. Las bases de datos orientadas a documentos poseen ciertas dificultades cuando se deben realizar operaciones en grandes conjuntos de datos. Dos de las bases de datos más populares para esta categoría son MongoDB y CouchDB.
- **Bases de datos orientada a grafos:** enfocadas principalmente en la manipulación de datos que posean grandes cantidades de relaciones entre ellos. En estas bases de datos, los datos son representados como nodo, relación y propiedades, los que las hace muy útiles cuando las relaciones entre datos poseen tanta importancia como los datos en sí. Dos de las bases de datos más populares para esta categoría son Neo4j y OrientBD.

2.5. Minería de datos

La minería de datos es un campo de las ciencias de la computación y estadística, la cual a través de conocimientos de inteligencia artificial, máquinas de aprendizaje, estadística y sistemas de bases de datos, se pueden descubrir patrones en grandes conjuntos de datos [Tan, 2018].

A través de la minería de datos, es posible detectar, por ejemplo, conductas como grupos entre los datos, los cuales representan ciertas tendencias; también se pueden detectar anomalías y dependencias entre los datos.

Es posible realizar dos tipos de análisis, descriptivo y predictivo. El primero permite estudiar los datos en sí, evaluar su distribución y posibles tendencias que presenten, mientras que el segundo es utilizado para modelar fenómenos y predecir el comportamiento que tendrán datos no estudiados. El análisis predictivo es utilizado principalmente para hacer prediciones de comportamientos futuros o no conocidos, mientas que el análisis descriptivo es utilizado para evaluar el estado actual y tomar desiciones en respecto a este.

Existen diversas técnicas de minería de datos, las cuales pueden realizar análisis descriptivos o predictivos, tales como redes neuronales artificiales, clustering, arboles de decisión, reglas de asociación y regresión lineal [Tan, 2018].

Redes neuronales artificiales

La redes neuronales artificiales son modelos computacionales basados en el mismo funcionamiento del cerebro, poseyendo nodos o neuronas que reciben *inputs* y entregan uno o más *outputs*, que dependen de la configuración de estas. Se debe destacar que las redes neuronales artificiales en sí no son un algoritmo, sino más bien un *framework* para diversos algoritmos que utilizan esta lógica [Tan, 2018].

La estructura básica de las redes neuronales artificiales está constituida por grupos de nodos, los cuales son llamados capas de nodos, las cuales van procesando como información de entrada el resultado entregado en la capa anterior.

Las capas de nodos se pueden categorizar en 3: entrada, la cual recibe la información ”externa” del sistema; capas ocultas o *hidden*, que solo interactúan con las capas adyacentes del sistema; y salida, la cual entrega el resultado del aprendizaje realizado dentro del modelo [Tan, 2018]. Un ejemplo de la estructura mencionada se puede observar en la figura 8.

En estos modelos el sistema va aprendiendo y generalizando a través de ejemplos reales. La forma en lo que esto ocurre es debido a funciones de activación y de salida, en conjunto con un factor de procesamiento llamado peso.

Existen 3 tipos básicos de aprendizaje: supervisado, en el cual se le indica a la red cual deben ser los valores de salida; no supervisado, en el cual la máquina no recibe indicaciones de la salida, por lo cual se va adaptando; y el híbrido, en el cual se le indica solo si el resultado es de buena o mala calidad.

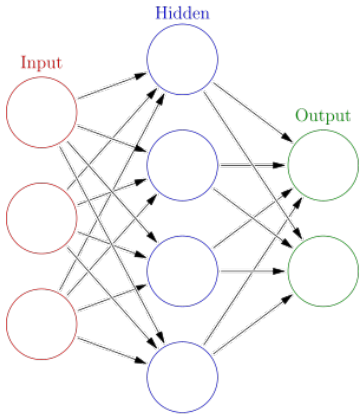


Figura 8: Ejemplo de red neuronal artificial
Fuente: Wikipedia.

Árboles de decisión

Los árboles de decisión son representaciones gráficas de funciones multivariadas; son considerados modelos predictivos, puesto que construyen diagramas lógicos que permiten la toma de decisiones en base a sucesos ya medidos.

Los árboles de decisión se componen de nodos y arcos. Cada nodo significa un estado de las variables y cada arco es la definición del estado de una variable. A cada nodo se puede llegar por sólo un camino y pueden salir tantos caminos como estados tenga la variable que está siendo decidida [Tan, 2018].

El objetivo de los árboles de decisión es poder graficar el resultado de la función analizada en relación a las conjugaciones de las variables, esto para entregar información de manera sencilla de comprender. Es posible analizar tanto variables categóricas como numéricas. Un ejemplo de un árbol de decisión se puede observar en la figura 9.

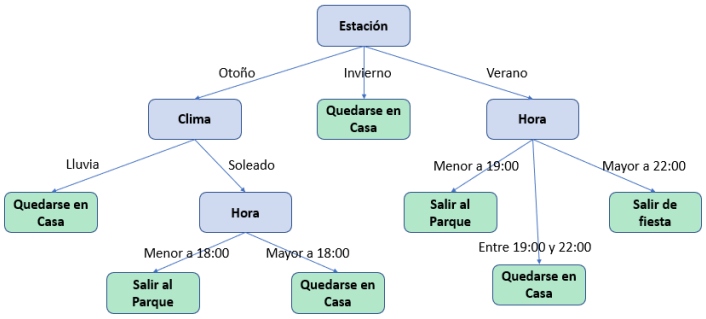


Figura 9: Ejemplo de un árbol de decisión
Fuente: Elaboración propia

Clustering

Los algoritmos de *clustering* son técnicas para separar los datos en grupos (*cluster*), los cuales son significativos, utiles, o ambas. Este análisis de agrupamiento de datos se basa sólo en la información del dato y sus relaciones [Tan, 2018]. El valor que se utiliza para determinar la asociación puede ser calculado a través de diversas definiciones dependiendo del caso de estudio; un ejemplo es la distancia euclídea, la cual se basa en el teorema de Pitágoras.

Las 2 clasificaciones de *clustering* más comunes son jerárquicas y de particionamiento por centroides. El *clustering* jerárquico crea dendrogramas según la cercanía entre nodos. Los algoritmos jerárquicos aglomerativos van generando uniones entre los dos *clusters* más cercanos, mientras que los algoritmos jerárquicos divisivos inician con un único *cluster*, el cual van dividiendolo según qué tan lejanos son.

El *clustering* por particionamiento divide los grupos de vectores en un número definido de grupos. Existen diversos tipos de algoritmos para esto, los cuales van variando en como se calcula la distancia, se posiciona el centroide inicial y como se van agrupando los vectores, entre otras posibilidades.

Reglas de asociación

La detección de reglas de asociación es una técnica que analiza un conjunto grande de datos en búsqueda de relaciones interesantes ocultas [Tan, 2018].

Si se tiene un conjunto de todos los ítemes de la forma $I = \{i_1, i_2, i_3...i_n\}$, una transacción como $t_x = \{i_{x1}, i_{x2}...i_{xm}\}$ y una base de datos como $D = \{t_1, t_2, t_3...i_k\}$, se puede definir como una regla de asociación una combinación de dos o más items que apunta a otra combinación de uno o más ítemes, de la forma $\{i_{xa}, i_{xb}\} \Rightarrow \{i_{xc}\}$. Lo anterior significa que en varias transacciones t_x de D existen diversas combinaciones de $\{i_{xa}, i_{xb}, i_{xc}\}$, de forma que las primeras dos condicionen la aparición de la tercera.

Debido a que es factible generar gran cantidad de asociaciones, es posible aplicar diversas métricas para encontrar sólo reglas interesantes, siendo las dos más utilizadas son la confianza y el soporte.

El soporte de un conjunto X se define como la proporción en la que aparece este dentro de toda la la base de datos, tal como se presenta en la ecuación 4.

$$Sop(X) = \frac{|X|}{|D|} \tag{4}$$

La confianza de la regla $X \Rightarrow Y$ se define como la proporción entre el soporte del conjunto $X \cup Y$ sobre el soporte del conjunto X, tal como se presenta en la ecuación 5.

$$Conf(X) = \frac{|X \cup Y|}{|X|} \tag{5}$$

Regresión lineal

La regresión es una técnica de modelamiento predictivo donde se estima que una variable objetivo es continua [Tan, 2018]. La regresión lineal es un modelo matemático que busca relacionar de forma lineal una variable dependiente Y con múltiples variables independientes X_i, siendo Y y X_i numéricos, de la forma $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2... + \epsilon$, donde β_i es un factor que indica la influencia de la variable independiente X_i y ϵ es un valor aleatorio.

CAPÍTULO 3
PROPUESTA DE SOLUCIÓN

Para el correcto desarrollo del análisis a realizar, se necesita definir el modo de trabajo y los elementos a utilizar, por lo que en este capitulo se presenta la metodología del trabajo a realizar, en conjunto con las técnicas y herramientas necesarias.

3.1. Metodología de trabajo

La estructura de trabajo utilizada posee las siguientes etapas:

- Recolección y Almacenamiento de datos
- Manipulación de datos y obtención de medidas y métricas.
- Presentación y elaboración de técnicas de estudio
- Análisis y comparación de resultados.
- Discusión y conclusiones.

De estas etapas, las primeras 3 son presentadas en este capítulo, mientras que las últimas 2 son desarrolladas en el capítulo 4.

3.1.1. Recolección y almacenamiento de datos

Debido a que Twitter no tiene ningún tipo de herramienta o funcionalidad para detectar hilos en la plataforma, es necesario realizar una búsqueda manual a través del buscador de ésta, por lo que de manera inicial no existe una técnica automatizada para conseguir un dataset de estudio; eso sí, existen diversas cuentas cuyo objetivo es presentar en páginas externas hilos en un formato parecido a una publicación de *blog* o noticia, donde cada Tweet es considerado como un párrafo de una entrada más grande. Para que un hilo sea procesado por alguna de estas cuentas es necesario que otro usuario las etiquete en alguno de los Tweets del hilo con algún comando específico (Comúnmente la palabra *unroll*). Gracias a la existencia de estas cuentas es posible obtener hilos con sólo una consulta en el buscador.

Para este trabajo se decidió buscar los hilos en los cuales la cuenta @threadreaderapp haya sido etiquetada, debido a la popularidad que ésta posee, por lo que la *query* realizada es ”@threadreaderapp unroll”, eso sí, encontrar hilos es sólo el primer paso manual para la

construcción del *dataset*, esto debido a que la mención puede ser en cualquier Tweet del hilo y Twitter no mantiene registros sobre si pertenecen a una conversación más grande o no, sólo mantiene un atributo llamado ”in_response_to”, por lo que luego de localizar un hilo, es necesario encontrar el último Tweet. La siguiente problemática que representa este tipo de búsqueda es definir cual será considerado el último Tweet, ya que es común que los autores inicialmente escriban el hilo completo, pero en el transcurso del tiempo agreguen información complementaria, anexos y nueva información que encuentren necesaria, por lo que se decide que un Tweet es el último cuando: éste posee palabras o *#hashtags* que indiquen fin, como por ejemplo ”*End*”, ”*#EndThread*”, etc. Cuando el Tweet siguiente posea palabras o *#hashtags* que indiquen que la información fue agregada después de la publicación inicial, como por ejemplo ”*postdata*”, ”*#PS:*”, etc. (debido a que este proceso es manual), se aprovecha esta instancia para filtrar hilos que posean sólo imágenes o hipervinculos, puesto que éstos no podrían ser estudiados a través de las técnicas a utilizar.

Luego de encontrar el último Tweet del hilo, se procede a registrar su ”id” en un achivo hilo.txt, el cual luego será leído a través de un código Python, mientras que las consultas a la red social se realizarán a través de *Twython*, una biblioteca que enmascara la API oficial de Twitter.

La lógica detrás del código para obtener todos los Tweets se basa en poseer una lista con todos los ”id” que se deben consultar, los cuales debido a restricciones de las funciones dispuestas por Twitter se pueden realizar con un máximo de 100 ”id” por consulta. En cada iteración se obtiene la información completa de todos los Tweets solicitados que se encuentren disponibles (es posible que algunos hayan sido borrados o que la cuenta que haya publicado el contenido sea privada y no se posean los permisos para ver sus Tweets). Luego de tener la información de la consulta, se procede a agregar los ”id” de los Tweets predecesores al final de la lista sobre la que se itera, los que se consiguen a través de la variable ”in_response_to”. A través de este método es posible que un Tweet intermedio del hilo no esté disponible, lo cual generará hilos incompletos. Este caso se manejará más adelante.

Luego de obtener obtener los datos y limpiarlos de variables que no entregan valor al trabajo, se tiene una lista de Tweets en formato diccionario, estos con los siguientes atributos:

- ”created_at”: Fecha en la que se publicó el Tweet
- ”id”: *id* del Tweet.
- ”id_str”: mismo valor a ”id” pero el tipo de la variable es *String*.
- ”in_response_to”: *id* del Tweet predecesor en el hilo, si el Tweet es el primero, este toma valor ”None”
- ”text”: contenido del Tweet, es decir, la cadena de carácteres publicada.
- ”user”: diccionario con los siguientes atributos de la cuenta que publicó el contenido:

    - ”id”: *id* de la cuenta.
    - ’id_str”: mismo valor a ”id” pero el tipo de la variable es *String*.
    - ”name”: el nombre de pantalla del usuario, no confundir con el *username* (se decidió no almacenar este último).
    - ”followers_count”: cantidad de seguidores a la fecha de consulta.
    - ”friends_count”: cantidad de cuentas que sigue la cuenta a la fecha de consulta.
    - ”created_at”: Fecha en la que se registró la cuenta.
    - ”verified”: valor *boolean* que indica si la identidad de la cuenta fue verificada por Twitter.
    - ”statuses_count”: cantidad de Tweets publicados por la cuenta.
- ”retweet_count”: veces que se le ha hecho Retweet al Tweet.
- ”favorite_count”: veces que se le ha marcado Like al tweet.
- ”hashtags”: Lista de todos los *#hashtags* utilizados en el Tweet.

Debido a la estructura que poseen los datos, además de la constante manipulación de éstos como diccionarios en Python, se utiliza como motor de base de datos *MongoDB Community Edition*, almacenando cada Tweet como un documento distinto.

3.1.2. Manipulación de datos y obtención de métricas

Luego de recolectar los datos, es necesario hilar los Tweets; esto se hace a través de la variable ”in_response_to”, existiendo 2 posibles opciones: la primera es buscar todos los Tweets que su ”id” no sea referenciada por ningún otro, los cuales se considerarán el último de cada hilo; luego de forma recursiva se consultará por el Tweet anterior con ”in_response_to” hasta que esta variable indique ”None”, lo cual indicaría que éste es el primer Tweet del hilo. La segunda opción es la versión inversa a la primera, partiendo con los Tweet que posean ”in_response_to” = ”None”, para luego encontrar el siguiente Tweet del hilo buscando el ”id” del Tweet actual en la variable ”in_response_to” hasta que no se obtengan resultados, lo cual indicará fin del hilo.

Como fue mencionado anteriormente, es posible que un hilo no haya sido rescatado íntegramente debido a la imposibilidad de obtener un Tweet intermedio, lo que genera una pérdida de los primeros Tweets. Este escenario genera un problema en la primera opción de construcción de hilos, puesto que si se recorre desde el final hacia el inicio, el código evaluará Hilos que finalmente no podrán ser construidos debido a que nunca se obtendrá un ”in_response_to” = ”None”. Debido a esto se decide proceder con la segunda opción, la cual asegura que sólo se evaluarán hilos completos.

Luego de reconocer todos los hilos, se agregan las siguientes variables a los documentos:

- ”Hilo”: indica a que hilo pertenece el Tweet.
- ”Pos_hilo”: posición del Tweet en el hilo.
- ”Progreso_hilo”: posición relativa al largo del hilo, siendo 0 el primer Tweet y 100 el último.
- ”Aporte_retweet”: relación entre el total de Retweets del Tweet sobre el total de Retweets del hilo.
- ”Aporte_favoritos”: relación entre el total de Likes del Tweet sobre el total de Likes del hilo.

Además de lo anterior, y debido a que se busca estudiar el comportamiento de hilos y no de Tweets por separado, se procede a crear una segunda colección de documentos, los cuales aglomeran las características de los Tweets pertenecientes a cada Hilo. La estructura de estos documentos es la siguiente:

- ”hilo”: *id* del hilo.
- ”Texto”: unión ordenada del texto de todos los Tweets que componen el hilo.
- ”Usuario”: Diccionario de todos los atributos anteriormente mencionados del usuario.
- ”fecha”: Fecha en la que se publicó el primer Tweet.
- ”total_tweets”: Cantidad de Tweets que componen un hilo, también llamado ”Largo” del hilo.
- ”total_retweets”: suma de todos los RT que recibieron todos los Tweets que componen el hilo.
- ”total_favorites”: suma de todos los Likes que recibieron todos los Tweets que componen el hilo.
- ”total_Hashtags”: Total de Hashtags distintos utilizados a lo largo del hilo.
- ”promedio_retweet”: Total de Retweets dividido por el largo del hilo
- ”promedio_favorites”: Total de Likes dividido por el largo del hilo
- ”all_hashtags”: Lista de todos los *hashtags* utilizados en el hilo.

Luego de definir la recolección y prepararación de los datos se procede con el estudio de tópicos, emociones, influencia del usuario e impacto.

Tópicos:

Tal como fue mencionado en el marco conceptual, es posible categorizar conjuntos de documentos modelando tópicos en común entre ellos a través de diferentes técnicas. Para esto se utilizó *Latent Dirichlet Allocation (LDA)* debido a la popularidad que presenta en los documentos estudiados.

La implementación de *LDA* se hizo a través de la biblioteca *Gensim* en conjunto con *Mallet*, un paquete de procesamiento estadístico de lenguaje natural basado en *Java*, puesto que una de sus implementaciones de *LDA* permite probar un rango de cantidad de tópicos a construir y evaluar la calidad de estos[2]. Este paquete se encuentra ya englobado para ser utilizada como función de *Gensim*.

Para la evaluación de la calidad de los tópicos se utilizó la métrica *Coherence Score:* C_v, esta indica la similitud de información mutua entre las palabras de los tópicos, con un valor dentro del rango 0 y 1.

El algoritmo desarrollado posee la siguiente estructura:

1. Extracción de datos de la base de datos.
2. Unión de todos los Tweets de un mismo hilo en un sólo *String*.
3. Limpieza de caracteres especiales de formato de texto, menciones (palabras iniciadas con), hipervínculos y *hashtags* (palabras iniciadas con #).
4. Limpieza de signos de puntuación, creación de bigramas (grupo de palabras que indiquen una sola idea, por ejemplo ”Donald Trump”), remoción de *Stopwords* (palabras sin mucho valor semántico, tales como ”the”, ”and” entre otros conectores) y lematización (convertir las palabras conjugadas a una raíz común llamada lema).
5. Construcción del Diccionario y Corpus con el cual se correrá el *LDA Mallet*.
6. Definición del rango de un valor alfa a evaluar y rango de cantidad de tópicos a construir.
7. Modelado de tópicos y cálculo de *Coherence Value*.
8. Selección de mejor combinación de alfa y cantidad de tópicos según *Coherence Value*.

Emociones:

Para la detección de emociones se decide utilizar 1 de los 3 modelos entrenados a través de *deep learning* presentados en [Colnerič y Demsar, 2018], debido a que estos están es-

[2]Mallet Homepage: `http://mallet.cs.umass.edu/` - Cunsultado el 06/09/2019

pecíficamente diseñados para el lenguaje utilizado en Twitter. El código disponible posee 3 modelos distintos, uno para cada categorización de emociones; Ekman, Plutchik y POMS. Se decide utilizar el modelo basado en las 6 emociones presentadas por Ekman por sobre Plutchik y POMS, puesto que por definición son las emociones más básicas que presentan los humanos, lo que se condice con el enfoque que se le desea dar a esta dimensión.

Al utilizar la biblioteca dispuesta por [Colneriĉ y Demsar, 2018] es necesario primero importar el modelo a utilizar; luego de esto se carga el conjunto completo de Tweets a analizar. Esta biblioteca posee 2 formatos para entregar resultados: el primero es una matriz en la cual cada Tweet posee por cada emoción un puntaje de ser categorizado en esta, mientras que el segundo formato es una tabla en la que cada Tweet es clasificado en una sola emoción, siendo ésta la con mayor puntaje de la matriz anteriormente mencionada.

Debido a que el modelo utilizado está diseñado para reconocer emociones sobre un solo Tweet y no un conjunto de Tweets como se desea, es necesario primero definir una técnica para evaluar el hilo.

Se definen 2 variables para esta dimensión: emoción principal y tupla de emociones principales. La primera selecciona la emoción principal según la técnica de estudio, mientras que la segunda selecciona las 2 primeras emociones. Si bien la primera variable es suficiente para categorizar los hilos, se define también la segunda variable puesto que se considera que debido a la extensión del texto, es probable que no sólo una emoción sea la determinante.

Se definen 4 posibles técnicas de detección de emociones, en todos los casos se elige la primera emoción para la variable Emoción principal y las primeras dos emociones para la tupla de emociones principales:

- Emoción media: a partir de la matriz de todos los Tweets del hilo se calcula el promedio de todos los puntajes para cada emoción y se seleccionan las emociones con los valores superiores.
- Emoción más intensa: a través de la matriz de todos los Tweets del hilo se selecciona el puntaje mayor de cada emoción y se seleccionan las emociones con los valores superiores.
- Emoción moda: a través de la tabla de todos los Tweets del hilo se selecciona la primera emoción moda (Primeras dos emociones moda para tupla de emociones principales).
- Emoción hilo: se unen de forma ordenada los Tweets en un solo texto y se detecta la emoción de cada hilo a través de la tabla.

Influencia del usuario:

A través de diversos estudios, como por ejemplo [Cha *et al.*, 2010], se sabe que un Tweet recibirá más atención si es que proviene de una cuenta perteneciente a un pequeño grupo llamados *Influencers*, los cuales son considerados de cierta forma ”lideres de opinión”, debido a que sus actos y decisiones llegan a influenciar a grandes grupos de personas. Considerando esto, si se desea estudiar el impacto de un hilo en la red social, es necesario definir métricas que tomen en consideración la cuenta generadora del contenido.

A través de la información recolectada es posible definir 3 métricas en relación a la influencia del usuario:

- Popularidad: se define como popularidad la cantidad de Followers que posee la cuenta; este valor indica el tamaño inicial de público que poseen las publicaciones realizadas. Eso sí, tal como fue presentado en el marco conceptual, este valor por sí solo no es completamente indicativo sobre la calidad de la audiencia que la cuenta posee [Cha *et al.*, 2010].
- Reputación: la reputación se define como la calidad del usuario a la vista de los demás, es decir, qué tanto interés genera y que tan importante es en la red a la que pertenece. Se puede medir a través de qué tan activa es la cuenta en relación entre sus Followers y Friends. Según [Drakopoulos *et al.*, 2017], es posible relacionar ambos valores a través una función sigmoídea (cuya fórmula se puede observar en la ecuación 3).
- Verificado: Twitter posee una variable llamada ”Verified´´ la cual indica si la cuenta fue reconocida como la identidad que dice ser. Las cuentas que poseen calidad de Verificada comúnmente son políticos, marcas y personalidades públicas, por lo que se puede asumir que este valor entrega nociones de importancia por sobre las demás cuentas no validadas.

Debido a que las 3 métricas son complementarias, se decide utilizar todas en el estudio a realizar.

Impacto:

Como se desea realizar una viralización de contenidos a través de hilos, es necesario medir el impacto que genera el contenido en la red social, esto es posible gracias a la existencia de 2 acciones realizables por los usuarios sobre los Tweets, marcar una publicación con un Like, lo que se considera que la cuenta está de acuerdo, simpatiza o encuentra interesante el Tweet, es decir, genera una reacción positiva sobre el usuario y este desea indicarlo. La otra acción disponible se llama Retweet, cuando uno hace Retweet a un Tweet, este se comparte en el *Timeline* de la cuenta que realizó la acción, permitiendo que sus Followers vean el contenido

como si fuese otra publicación del usuario, pero manteniendo los atributos originales de la cuenta publicadora.

Debido a que se comprende la viralización de un contenido como el suceso de que una publicación se vea por la mayor cantidad de usuarios posibles, se considera como una métrica de esto la cantidad de Retweets que recibe un hilo, mientras que los Likes se consideran como otra variable de estudio.

Puesto que se desea estudiar el impacto de los hilos por sobre el impacto de los Tweets, es necesario definir métricas de estudio para el hilo, tanto para los Retweets como para los Likes:

- Total RT (o FV): se llama Total RT a la suma de Retweets de todos los Tweets pertenecientes a cada hilo.
- Promedio RT (o FV): Se llama Promedio RT al ”Total RT” dividido por ”Total Tweet”.
- Log Total RT (o FV): de define como el logaritmo en base 10 del total de RT por cada hilo; debido a la naturaleza de los logaritmos, al total de RT se le suma 1.
- Log Promedio RT (o FV): se define como el logaritmo en base 10 del promedio de RT por cada hilo; debido a la naturaleza de los logaritmos, al total de RT se le suma 1.

Las ultimas 2 métricas se definieron debido a la distribución que presentan los Retweets, donde una muy pequeña cantidad de hilos suman una gran cantidad, mientras que la mayoría recibe valores mucho menores, por lo que para estandarizar el crecimiento, se aplica la función logaritmo.

3.1.3. Presentación y elaboración de técnicas de estudio

Luego del procesamiento de los datos y la obtención y selección de métricas a utilizar, se debe continuar con el estudio de éstas. El objetivo principal del trabajo a realizar es descubrir qué factores influyen sobre el impacto de cada hilo, por lo que se debe estudiar tanto el comportamiento de las medidas en conjunto y por si solas, además del efecto que generan éstas sobre la métrica de impacto seleccionada.

Para el estudio en sí de las medidas y métricas, se decide partir con un análisis gráfico a través del *software Microstrategy Desktop*. Este software está diseñado para el análisis de datos, los cuales pueden ser cargados a través de archivos o conectándose a uno de los múltiples motores de bases de datos disponibles[3]. Se decide trabajar sobre esta herramienta debido a

[3]Microstrategy Desktop, Get started: `https://www.microstrategy.com/us/get-started/desktop` - Consultado el 06/09/2019

la capacidad de conexión con MongoDB, cosa que actualmente no todas las herramientas similares lo pueden hacer de forma nativa. También entra en consideración los conocimientos previos que se poseen sobre esta herramienta.

Para el análisis y validación de las variables seleccionadas, se decide construir múltiples tipos de modelos de minería de datos y evaluar la calidad de estos. Para la construcción se elige trabajar sobre Orange 3, una herramienta *open source* para realizar minería de datos y análisis predictivo. Se decide utilizar este *software* debido a su facilidad de uso a través de sus *workflows Drag and Drop*[4]. Se debe recalcar que el objetivo del uso de esta herramienta es más que nada evaluar el efecto que generan las variables sobre el impacto, más que conseguir un modelo que prediga el comportamiento de éste, por lo que se estudia la variación de la calidad más que la predicción en sí.

[4]Orange - Datamining Fruitful and fun: `https://orange.biolab.si/` - Consultado el 06/09/2019

CAPÍTULO 4
DESARROLLO Y VALIDACIÓN DE LA SOLUCIÓN

En este capitulo se presentan los resultados del modelado de los tópicos, la detección de emociones, el cálculo de las métricas de impacto, el análisis de los datos, la construcción de recomendaciones y la validación de la solución.

4.1. Modelado de tópicos

Para la selección del mejor conjunto de Tópicos primero se modelan 5 a 21 tópicos variando el valor del parametro *alpha* [0.01, 0.1, 1, 10, 30, 100, 600] , con un incremento de 1 tópico para cada iteración. Los resultados de coherencia para cada conjunto se pueden ver en la figura 10, mientras que el comportamiento general del *Coherence Score* para cada *alpha* se puede observar en la figura 11.

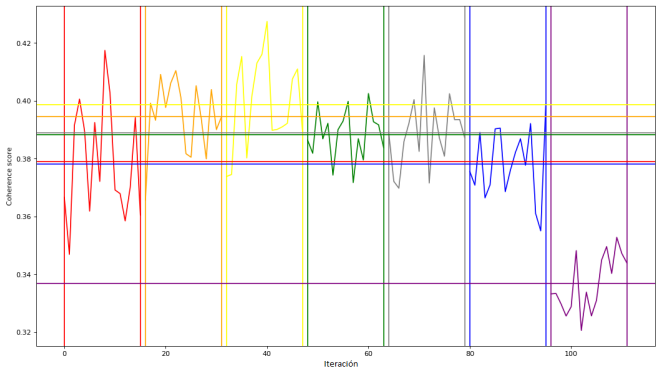


Figura 10: Gráfico de *Coherence Score* para cada iteración. Cada color indica un valor *alpha* distinto [0.01, 0.1, 1, 10, 30, 100, 600], los que están ordenados de manera ascendente; las barras horizontales indican el promedio de cada alfa, mientras que las verticales separan los resultados de cada intervalo *alpha*.

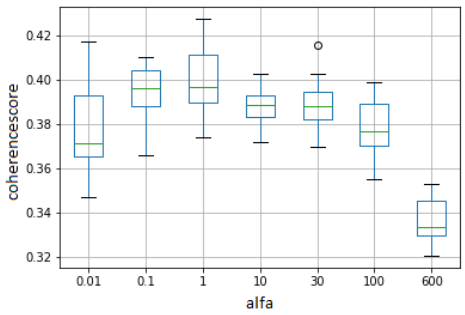


Figura 11: *Boxplot* de *Coherence Score* para cada valor alfa [0.01, 0.1, 1, 10, 30, 100, 600].

De las figuras 10 y 11 se puede observar que los mejores valores *alpha* son 0.1 y 1, por lo que se procede a realizar una segunda búsqueda para valores *alpha* entre 0.1 y 1, más específicamente [0.5 , 0.7, 0.9], y entre 10 y 31 tópicos. Los resultados se pueden observar en las figuras 12 y 13. En éstas se observa que un valor *alpha* 0.9 entrega mejores valores para *Coherence Score*, siendo la mejor combinación de valores alfa = 0.9 con 14 tópicos generados y un $C_v = 0{,}4372$.

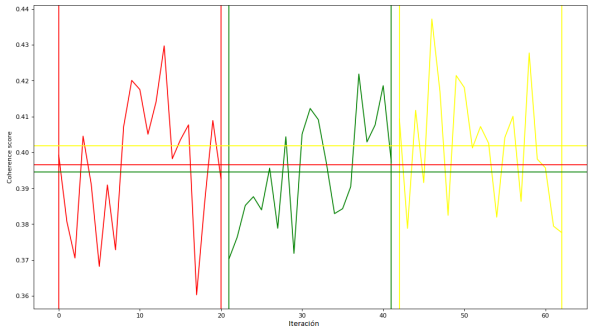


Figura 12: Gráfico de *Coherence Score* para cada iteración. Cada color indica un valor alfa distinto [0.5 , 0.7, 0.9], los que están ordenados de manera ascendente.

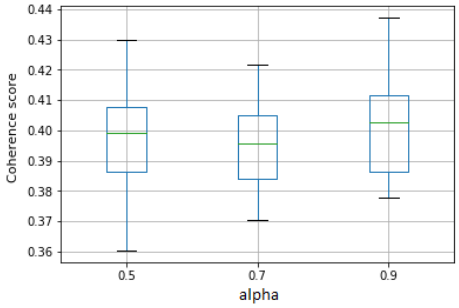


Figura 13: *Boxplot* de *Coherence Score* para distintos valores del parametro *alpha* [0.5 , 0.7, 0.9] .

Topico 0	Topico 1	Topico 2	Topico 3	Topico 4	Topico 5	Topico 6
23	15	27	17	127	19	11
Topico 7	Topico 8	Topico 9	Topico 1o	Topico 11	Topico 12	Topico 13
35	16	31	22	13	17	19

Tabla 1: Distribución de hilos según tópicos

La distribución de los hilos para cada tópico se presenta en la tabla 1. La detección de tópicos entrega como resultados vectores de palabras con sus respectivos pesos, en la tabla T1 del anexo se presenta el posible tema de cada tópico, en conjunto con las 6 primeras palabras más importantes. Una de las observaciones más recurrentes que se realiza en la literatura, es que no siempre es posible reconocer el tema que tratan los tópicos, puesto que si bien se reconocen las palabras de mayor importancia, estas no siempre poseen una relación muy

clara entre ellas, por ejemplo, es posible asumir que el Tópico 0 trata sobre política de Estados Unidos, caso contrario es el Tópico 4, el cual presenta una diversidad semántica mucho mayor.

Después de modelar los tópicos se procede a reconocer las posibles temáticas de cada uno. Para esto, se observa que tienen en común las 40 palabras más importantes de cada tópico, se escribe una idea inicial, luego se leen algunos hilos al azar a modo de comparación con la idea escrita y finalmente se decide que tema está tocando cada tópico. Los resultados de este análisis se presentan en el anexo, más precisamente en la tabla T1, se presenta un extracto de esta seleccionando los 5 tópicos más populares en la tabla 2.

Luego de describir los tópicos en la tabla T1 del anexo, se observa que existen 3 temáticas principales en los hilos del dataset. Estos son:

- Estados Unidos: 0, 5, 9, 10 y 11.
- Política y economía (excluyendo EEUU): 6, 7, 12 y 13.
- Sociedad: 1, 2, 3 y 8.

Se observa, también, que existen tópicos específicos que hablan de un hecho o un grupo de hechos o problematicas específicas, como lo son 0, 1, 5, 6, 8, 9, 10, 11 y 13. mientras que existen otros que son de carácter general, como los tópicos 2, 3, 7 y 12.

Tópico	Hilos	Posible tema según palabras	6 primeras palabras
Tópico 0	23	Política actual estadounidense, discusión sobre partidos políticos y reformas migratorias.	American, Conservative, People, Democrat, Republican, Vote.
Tópico 2	27	Familia, hijos y salud.	Child, Kid, Day, Work, Remember, Story.
Tópico 4	127	No es posible reconocer un tema en específico.	People, Make, Work, Time, Good, Thing.
Tópico 7	35	Política internacional y gobierno.	Law, State, President, Order, Official, Office.
Tópico 9	31	Noticias sobre el reporte Mueller sobre actos ilícitos de Donald Trump.	Investigation, Report, Muell, Flynn, Crime, Mueller .

Tabla 2: Resumen de la clasificación de los tópicos modelados.

4.2. Detección de emociones

Se presentan los resultados obtenidos a través de las técnicas de detección de emociones sobre hilos en la sección 3.1.2. En la tabla 3 se presenta la cantidad de hilos para cada emo-

ción según la técnica utilizada, mientras que en la tabla 4 se muestra el porcentaje de hilos por emoción para cada técnica.

Técnica	Anger	Disgust	Fear	Joy	Sadness	Surprise
Media	0	0	57	324	0	11
Intensa	4	6	80	279	2	21
Moda	2	0	62	317	2	9
Hilo	19	8	109	220	16	20

Tabla 3: Distribución de los hilos por emociones según cada técnica de etiquetado.

Técnica	Anger	Disgust	Fear	Joy	Sadness	Surprise
Media	0	0	14,5	82,7	0	2,8
Intensa	1,0	1,5	20,4	71,2	0,5	5,4
Moda	0,5	0	15,8	81,1	0,5	2,3
Hilo	4,8	2,0	27,8	56,1	4,1	5,1

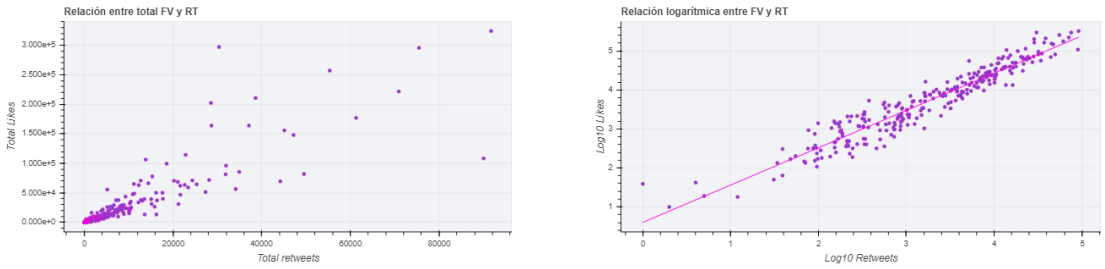
Tabla 4: Porcentaje de los hilos por emociones según cada técnica de etiquetado.

En la tabla 4 se observa que las técnicas emoción media y emoción moda categorizan más del 80 % del total de los hilos con la etiqueta *Joy*, mientras que técnica de emoción intensa presenta un comportamiento similar, etiquetando sobre el 70 % de los hilo con la misma etiqueta ya mencionada. Se observa también que en las 2 primeras técnicas mencionadas existen emociones no detectadas.

Finalmente se decide trabajar con emoción hilo, puesto que posee una mejor distribución de las emociones en comparación con las demás.

4.3. Métrica de impacto

Puesto a que se definió Retweet por sobre Like como la variable de impacto, se desea estudiar si la segunda tiene alguna relación directa o si se debe considerar como una de las medidas de estudio para el trabajo posterior. Para realizar esto se decide elaborar una regresión lineal entre ambas variables a través de las 4 métricas definidas: Total RT vs Total FV, Promedio RT vs Promedio FV, Log Total RT vs Log Total FV y Log Promedio RT vs Log Promedio FV. En la figura 14a se puede observar del comportamiento entre Total RT vs Total FV que existe una relación entre ambas, además, Promedio RT vs Promedio FV tiene el mismo comportamiento que este, por lo que no se muestran sus gráficos. En la figura 14b se observa del comportamiento entre Log Total RT vs Log Total FV una clara relación lineal, además, Log Promedio RT vs Log Promedio FV posee el mismo comportamiento.



(a) Relación entre los totales de RT y FV

(b) Relación logarítmica entre RT y FV

Figura 14: Gráficos de la relación entre Retweets (RT) y Likes (FV)

Luego de analizar el comportamiento en los gráficos, se realiza el cálculo de la regresión lineal para los 4 casos mencionados. Los resultados de esto se puede observar en la tabla 5. A través del coeficiente de correlación R^2 se puede afirmar que ambas variables son dependientes, puesto que sin importar cuál métrica se utilice, todos poseen un valor R^2 cercano a $0{,}9$. También se observa que al trabajar con el promedio de las variables se obtiene un mejor resultado por sobre los totales de éstas; se considera que debido al existir una mayor cantidad de Tweets en el hilo, mayor es la probabilidad de pequeñas diferencias, por lo que promediando éstas, es posible normalizar el comportamiento. Se observa también que el mejor resultado es la regresión lineal de la relación logarítmica del promedio de RT y FV, con un valor $R^2 = 0{,}901$.

Finalmente, se concluye que existe una clara relación de dependencia entre Retweets y Likes, por lo que se decide que no es necesario trabajar con la segunda variable.

	Relación	Intercepto	Coeficiente	R^2
1	RT vs FV	-12832.81232	5.26569	0.887
2	log(RT) vs log(FV)	0.59636	0.96618	0.869
3	prom(RT) vs prom(FV)	-603.53960	4.40572	0.897
4	log(prom(RT)) vs log(prom(FV))	0.50719	0.98419	0.901

Tabla 5: Resultados de regresión lineal sobre Retweets (RT) y Likes (FV). Se considera Log como logaritmo en base 10, mientras que prom es el promedio de la variable.

4.4. Análisis de datos

La siguiente tarea consiste en realizar un estudio inicial de los datos, analizar el comportamiento de diversas variables y reconocer patrones que puedan ser de utilidad para los análisis posteriores.

Para el análisis de datos se genera una muestra aleatoria de aproximadamente el 70 % del total del dataset a través de una función de selección *random*; el tamaño del conjunto es

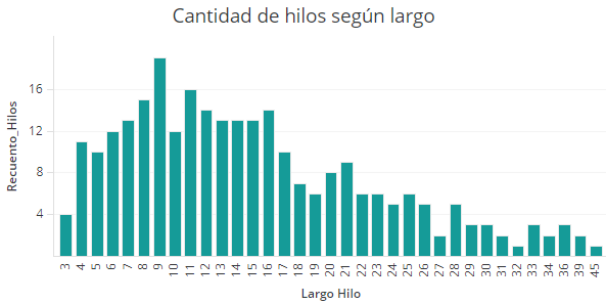
de 274 hilos. La función para la creación de la muestra fue realizada a través de Python y almacenada como un conjunto de documentos distinto al principal.

El resto de los hilos se utilizará como dataset de evaluación para los modelos generados.

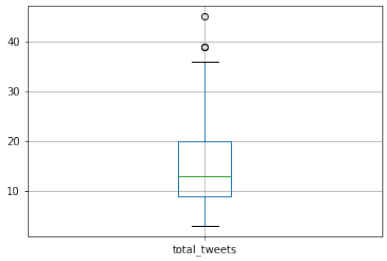
4.4.1. Análisis de hilos

Se estudia, primero que todo, el largo de los hilos (total Tweets) y la cantidad de *hashtags* (HT) para tener una base comparativa de ambas variables para análisis posteriores.

Se tiene como el largo de los hilos a la cantidad de Tweets que lo componen. En la figura 15a se observa que los hilos tienen un largo entre 3 y 45 Tweets, mientras que en 15b se ve que la primera mitad de los hilos tiene un largo menor o igual a 13 Tweets, mientras que el 75 % se encuentra bajo los 20 Tweets de largo. Se tiene que el promedio de Tweets por hilo es de aproximadamente 15.



(a) Gráfico de distribución de hilos según largo.



(b) Boxplot de distribución de hilos según largo.

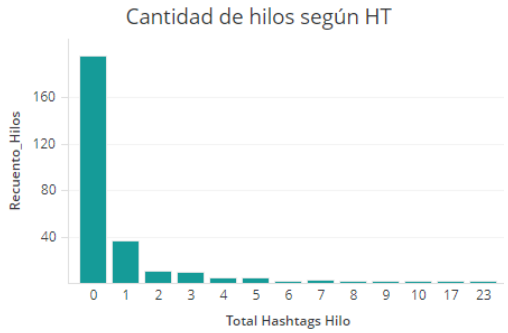
Figura 15: Distribución de los hilos según el largo.

En relación a los *hashtags*, se puede observar en la figura 16a que la gran mayoría no los utiliza (196 hilos), mientras que la segunda mayoría utiliza sólo 1 (37 hilos). De la figura 16b se observa que la cantidad de hilos va aumentando según más reciente es la fecha de publicación.

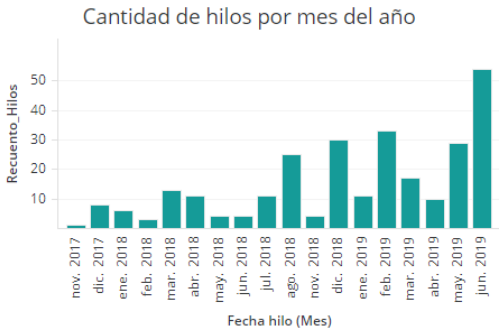
4.4.2. Análisis de influencia

En relación a la influencia, se estudian diversas variables según el puntaje sigmoídeo (P_s), el estado de verificación de la cuenta (*verified*) y la cantidad de Followers (FWR). Se debe dejar claro que para los gráficos se decidió aproximar P_s al primer decimal, meramente con fines de visualización.

Se observa en la figura 17a que la gran mayoría de los hilos son publicados por cuentas que poseen una reputación entre $[0{,}8, 1{,}4]$, los que comparados con el rango total de la muestra,



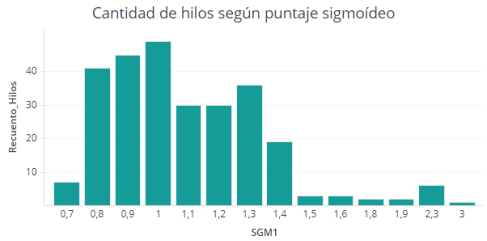
(a) Distribución de hilos según la cantidad de *hashtags* utilizados.



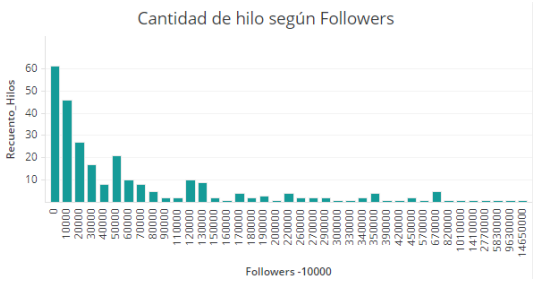
(b) Distribución de hilos según el mes en el que fueron publicados.

Figura 16: Distribución de los hilos según cantidad de *hashtags* y mes de publicación

son valores relativamente bajos. También se sabe de los datos que la mayoría de los usuarios escribió un hilo de la muestra, por lo que se puede asumir que la distribución presentada en este gráfico también representa la distribución de las cuentas.



(a) Distribución de hilos según puntaje sigmoídeo.

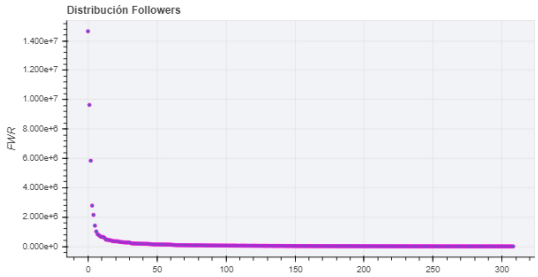


(b) Distribución de hilos según cantidad de Followers.

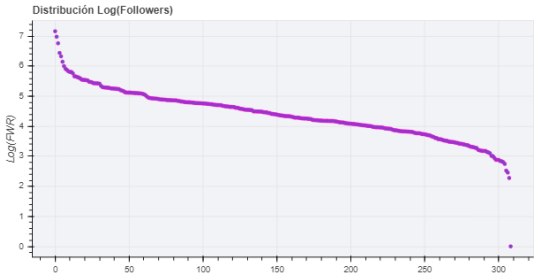
Figura 17: Análisis de hilos según influencia

Si se asume que a través del tiempo el comportamiento de la influencia de los usuarios presentada en [Garcia *et al.*, 2017] sigue siendo válida, es posible utilizar la figura 7 de la sección 2.3.3 como referencia para la reputación. Esta figura presenta el conjunto *In* como los usuarios con menor reputación, los cuales significan aproximadamente un 4 % de la muestra, el conjunto *SCC* un 80 % y el conjunto *Out* un 16 % de la población, siendo estos los con mejor reputación. En la figura 17a es posible observar un comportamiento similar al explicado.

Al comparar los grupos presentados en la figura 7 de la sección 2.3.3 con los valores obtenidos de P_s se tiene que el 4 % de los datos representa aproximadamente 11 hilos, por lo que si se ordenan de manera ordenada ascendentemente, se tiene que el conjunto *In* se define por los valores P_s = $[0{,}68, 0{,}78]$. El 16 % de la muestra tiene aproximadamente 44 hilos, los cuales de forma ordenada descendentemente se definen por valores P_s = $[1{,}27, 3{,}00]$. Finalmente se tiene que el conjunto *SCC* se define por P_s = $[0{,}79, 1{,}26]$. Lo importante de este análisis es que permite definir qué valores se consideran como puntaje sigmoídeo bajo, normal y alto.



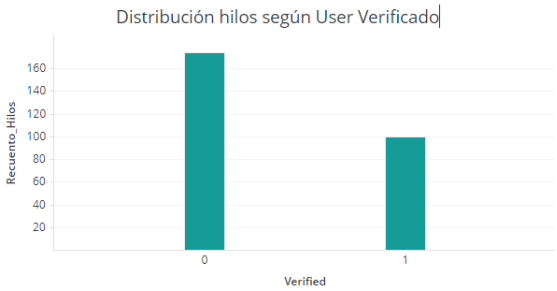
(a) Distribución de usuarios según cantidad de FWR.



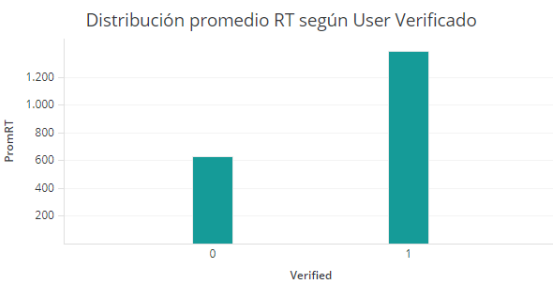
(b) Distribución de usuarios según logaritmo de FWR.

Figura 18: Análisis de influencia según FWR

En respecto a los Followers, se puede observar en la figura 17b que la gran mayoría de los hilos poseen valores relativamente bajos en comparación al rango total de seguidores. Se ve este mismo comportamiento en el gráfico 18a, el cual compara los FWR de cada usuario, lo que demuestra una relación directa entre los usuarios y los hilos. En la figura 18b se presenta la relación entre el logaritmo de FWR de cada usuario, esta gráfica muestra un comportamiento cercano a una recta en su centro. El gráfico de la figura 18b muestra que existe una pequeña cantidad de elementos con una gran cantidad de FWR, una gran mayoría con valores cercanos y una minoría con valores muy bajos.



(a) Distribución hilos según usuario verificado.



(b) Promedio de RT según usuario verificado.

Figura 19: Análisis de influencia según usuario verificado. La barra izquierda indica usuarios no verificados.

Para estudiar la importancia del atributo *verified* del usuario, se decide comparar la cantidad de hilos según el estado (figura 19a) vs la cantidad de Retweets en promedio que reciben las cuentas verificadas (figura 19b). A primera vista es posible darse cuenta que los usuarios verificados (valor 1 en ambos gráficos) reciben en promedio una mayor cantidad de RT; también se observa que la mayoría de los hilos provienen de cuentas que no están verificadas.

Finalmente, se recuerda que según [Cha *et al.*, 2010] no solo las personas naturales son usuarios de Twitter, sino que también existen cuentas de entidades (cuentas que no representan una persona, sino que a una organización, u otras que tienen como objetivo compartir contenidos específicos), las que poseen una reputación mayor, además de muchas veces poseer el

estado de verificado. Esto explica en parte la clara tendencia de estas cuentas, de presentar un impacto promedio mayor por sobre las no verificadas.

4.4.3. Análisis de Emociones

Como se poseen dos variables categóricas para el estudio de emociones en los hilos, primero se procede a analizar la Emoción principal, luego se continua con un análisis sobre la tupla de emociones principales, finalmente se comparan ambas para definir los casos en los cuales se utiliza cada una.

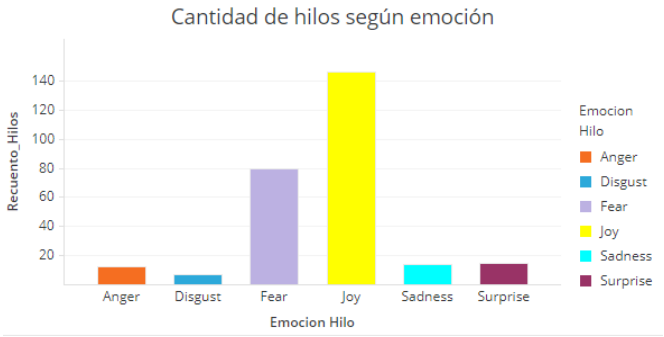


Figura 20: Distribución de hilos según emoción principal.

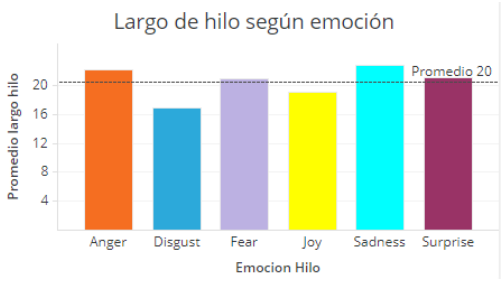
En el gráfico 20, se puede observar que tal como fue mencionado en la sección 4.2, las emociones más populares son *Joy* y *Fear*, sumando entre ambas 226 hilos, lo que representa un $80\,\%$ de la submuestra de estudio.

A través del gráfico 21a se puede observar que el largo promedio de los hilos para cada emoción es de 20 Tweets por hilo. Se reconoce también que no existe una gran diferencia de largo según las emociones, teniendo un rango de 17 a 23 Tweets en promedio. Se ve que al estudiar los hilos según emoción el largo promedio es mayor a los 14 Tweets promedio presentada en la sección 4.4.1, la cual estudia la distribución de los hilos en sí.

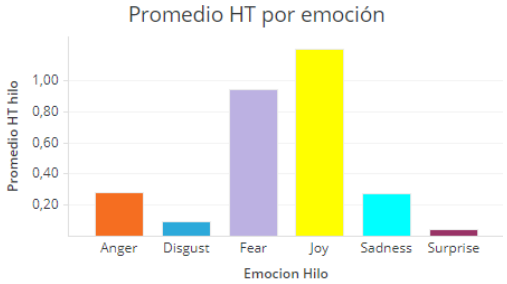
De la figura 21b se puede observar que sólo los hilos escritos con una emoción cercana a *Joy* y *Fear* tienden a usar 1 *hashtag*, con $1{,}2$ HT y $0{,}8$ HT respectivamente, mientras que las demás emociones no los utilizan.

En relación a tupla de emociones principales, se puede ver en la figura 22, que la gran mayoría de los hilos son [*Joy*, *Fear*] o [*Fear*, *Joy*], sumando ambas combinaciones un $55\,\%$ del total de los hilos. Considerando los resultados obtenidos en la figura 20, se comprueba que las dos emociones más predominantes en Twitter son *Joy* y *Fear*. Se observa también que al comparar las figuras 22 y 20, ignorando las 2 emociones más populares, el resto de las combinaciones posee una distribución similar de aproximadamente 5 hilos por tupla.

En relación al largo, se observa de la figura 23 que las tuplas de emociones poseen un largo



(a) Promedio de largo de hilos según emoción principal.



(b) Promedio de *hashtags* según emoción principal

Figura 21: Análisis de emociones según largo y *hashtags*.

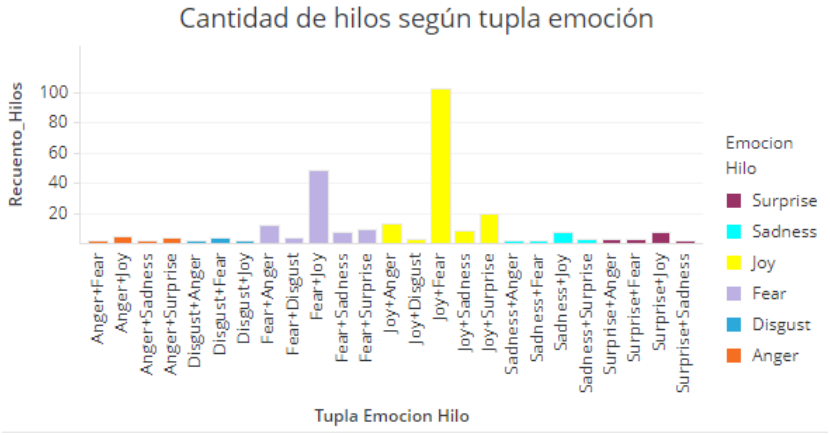


Figura 22: Distribución de hilos según tupla de emociones principales.

mucho más variado que en la figura 21a. Al comparar con la distribución de hilos de la figura 22 se puede asumir que este fenómeno se debe a la baja cantidad de hilos por tupla. Se observa también que el promedio de largo promedio por tupla es de 18 Tweets.

De la figura 24 se observa que todas las combinaciones de *Fear* y *Joy* presentes poseen hilos con *hashtags*, algo esperable según los resultados obtenidos en 21b. Un comportamiento interesante se observa en la tupla [*Sadness*, *Fear*], la cual tiene en promedio $HT = 1{,}8$, eso sí, se observa en la figura 22 que esta combinación sólo tiene 2 hilos, por lo que no se puede considerar como un resultado concluyente. Ocurre un fenómeno similar en las otras 2 combinaciones que resaltan.

Luego de los análisis de ambas variables, se observó que el comportamiento de tupla de emociones principales es una subdivisión de emoción principal, por lo que para los análisis gráficos posteriores sólo se presentarán figuras que estudien la emoción principal, mientras que para los estudios realizados a través de código se utilizará la tupla de emociones principales, a no ser que se indique lo contrario.

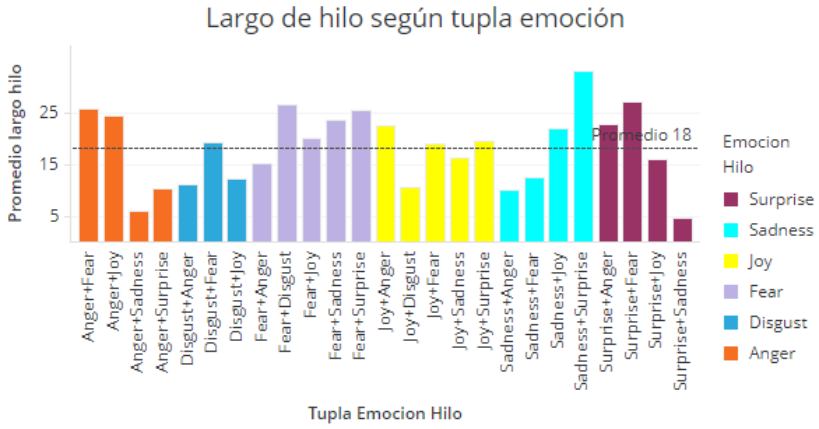


Figura 23: Promedio de largo de hilos según tupla de emociones principales.

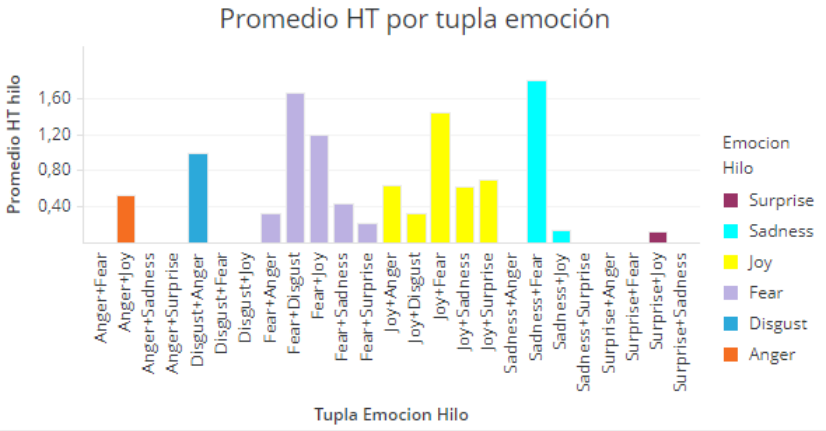


Figura 24: Promedio de *hashtags* según tupla de emociones principales

4.4.4. Análisis de tópicos

En la figura 25 se observa que la cantidad de hilos por tópico es relativamente similar, por lo que considerando el tópico 4 como outlier, se calcula el promedio de hilos del resto obteniendo aproximadamente 14 hilos por tópico.

En respecto a la cantidad de *hashtags* utilizados por cada tópico, se puede observar en la figura 26a que el tópico con mayor uso de estos es el número 4, con aproximadamente uno por cada hilo; otras categorías con valores superiores a $0{,}8$ son los tópicos 0 (política estadounidense), 2 (Familia, hijos y salud) y 10 (Guerra e intervencionismo estadounidense en medio oriente).

En relación al largo promedio según el tópico se puede observar en el gráfico 26b que el promedio es de 15 Tweets por hilo, el mismo promedio obtenido en el análisis de la figura 15. Es digno de rescatar también que el largo promedio se mueve entre 12 y 20 Tweets, un

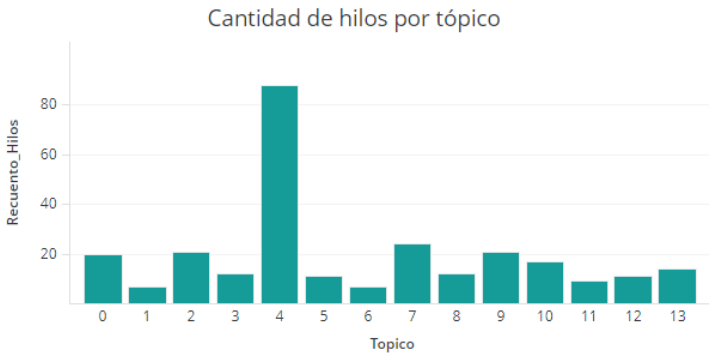
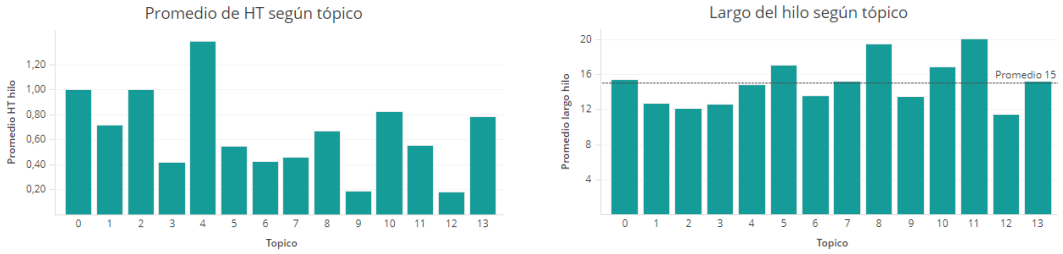


Figura 25: Distribución de hilos según tópico.

rango mucho menor comparado con el rango de todos los hilos (3 a 45 Tweets).



(a) Cantidad de *Hashtags* según tópico.

(b) Largo del hilo según tópico.

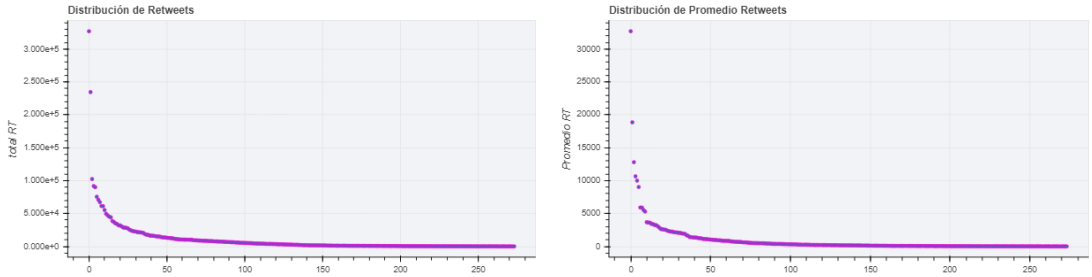
Figura 26: Análisis de tópico según cantidad de *hashtags* y largo del hilo

4.4.5. Análisis de impacto

Se desea analizar el efecto que genera sobre el impacto el largo del hilo, además de estudiar cómo se evoluciona el impacto a través del progreso del hilo. Debido a que ya se demostró la relación directa entre RT y FV, sólo se realiza el estudio según los Retweets.

Primero se desea comparar el comportamiento entre el total de los RT y el promedio de los RT para cada hilo. Se recuerda que la diferencia entre estas 2 métricas es que para cada hilo el promedio RT es el total RT dividido por el total de Tweets (largo). Un primer acercamiento a la comparación de estas dos métricas se observa en ambos gráficos de la figura 27, los cuales presentan de forma descendente ambas métricas en relación al hilo. Se observa que no existe una gran diferencia en el comportamiento, sólo en la escala del eje Y, siendo del orden de 10^5 en el gráfico 27a vs un orden de 10^4 en el gráfico 27b. Se debe dejar claro que en el proceso de calcular el promedio, los hilos no necesariamente toman la misma posición en ambas representaciones.

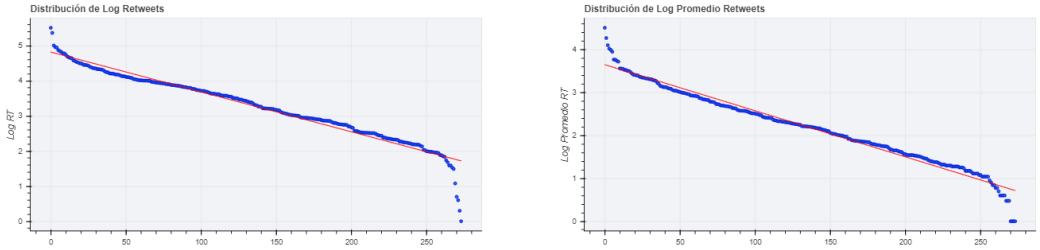
Debido al comportamiento presentado en ambos casos de la figura 27, se procede con un análisis de la distribución del logaritmo de las variables Total RT y Promedio RT, el cual se



(a) Total de RT según hilos. (b) Promedio de RT según hilos.

Figura 27: Análisis de impacto según RT por hilos

presenta en la figura 28. En este caso se comprueba que el comportamiento de los RT sigue siendo similar en ambas variables, sólo presentando una leve diferencia en la altura de la distribución, en conjunto con una pequeña diferencia en el inicio y final de los datos. Se tiene que la ecuación 6 representa a la linea de tendencia del gráfico 28a, mientras que la ecuación 7 pertenece al gráfico 28b. De ambas ecuaciones se puede observar que la pendiente es muy similar, variando en aproximadamente un $5\,\%$, mientras que el intercepto disminuye aproximadamente un $30\,\%$.



(a) Logaritmo de total de RT por hilo. (b) Logaritmo del promedio de RT por hilo.

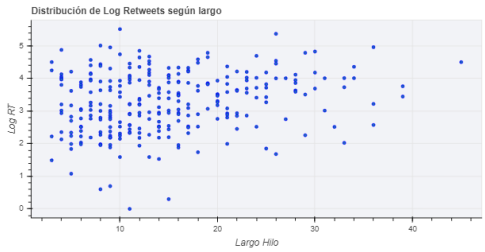
Figura 28: Análisis de impacto según el logaritmo de RT

$$Y = -0{,}011365 \cdot X + 4{,}824049 \tag{6}$$

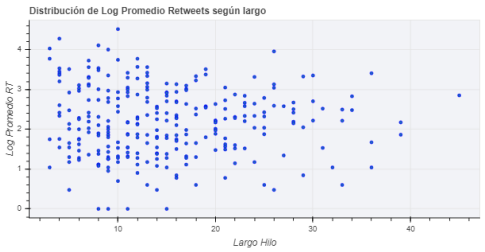
$$Y = -0{,}010761X + 3{,}655273 \tag{7}$$

Se prosigue con el análisis de los Retweets en respecto al largo con la figura 29, en donde se ve la relación entre ambas variables. Se observa el mismo comportamiento en las figuras 29a y 29b: mientras mayor sea el largo, mayor es el mínimo de RT por cada hilo, siendo la diferencia en ambos el rango de la variable dependiente.

Finalmente, se concluye del análisis de las figuras 27, 28 y 29 que si bien las 2 métricas de impacto presentan fenómenos similares, se puede considerar como mejor métrica compa-



(a) Logaritmo del total de RT según largo del hilo.

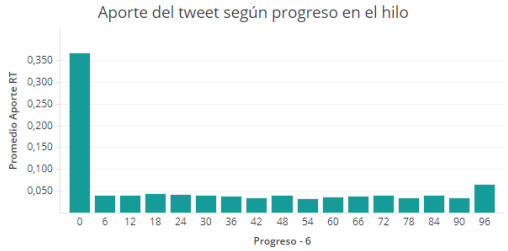


(b) Logaritmo del promedio de RT según largo del hilo

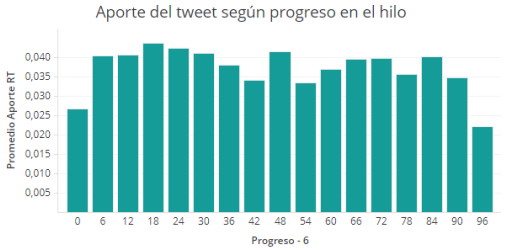
Figura 29: Análisis de impacto según largo del hilo.

rativa de los hilos el Promedio RT, puesto que elimina la ”ventaja inicial” que poseen los hilos más largos por el simple hecho de tener más Tweets donde acumular Retweets.

Luego del estudio del efecto del largo del hilo sobre el impacto, se procede a estudiar el comportamiento que presenta el impacto según el progreso del hilo. Se recuerda que el progreso del hilo es la posición relativa del Tweet sobre el hilo, que se calcula dividiendo el número del Tweet por el total de Tweets del hilo, teniendo un ”porcentaje de progreso” del hilo. Para estudiar el impacto según el progreso se debe utilizar la métrica ”Aporte RT”, la cual es el porcentaje que representa el total de RT del Tweet sobre el total de RT del hilo.



(a) Total de RT por Tweet según progreso del hilo, incluyendo primer y último Tweet.



(b) Total de RT por Tweet según progreso del hilo, excluyendo primer y último Tweet.

Figura 30: Análisis de impacto según progreso del hilo.

De la figura 30a se observa que el Tweet más importante de los hilos es el primero, con un un aporte promedio del $37\,\%$ del total de todos los RT que recibe un hilo, mientras que el segundo mayor valor se encuentra al final del hilo, con un aporte promedio de $6\,\%$. Luego de reconocer la importancia del primer y último Tweet de los hilos, se procede a excluir estas posiciones, quedando finalmente la distribución presentada en la figura 30b. De la figura 30b se observa que a lo largo de los hilos se posee un comportamiento similar entre Tweets, con un rango $[2{,}2\,\%, 4{,}4\,\%]$ en el promedio aporte RT, valores muy pequeños como para que una posición sea reconocida como más importante que las demás.

4.4.6. Impacto vs influencia

Se recuerda que la reputación del usuario, medida a través del puntaje sigmoídeo, se considera baja cuando se encuentra en el rango [0.7 , 0.8], una reputación normal entre (0.8 , 1.3) y una reputación alta entre [1.3 , 3.0] (al aproximar los rangos para estudiar los gráficos se tienen limites abiertos o cerrados, según sea el caso). Para analizar la influencia que genera la reputación del usuario sobre el promedio RT se presenta en la figura 31 el porcentaje de promedio RT según el puntaje sigmoídeo del usuario.

Se observa en la figura 31 que los usuarios con una reputación alta poseen mejores valores de impacto, mientras que los usuarios con peor reputación presentan malos resultados, en relación a los usuarios con reputación normal se observa que presentan un impacto regular. Se nota que ambos extremos presentan malos resultados.

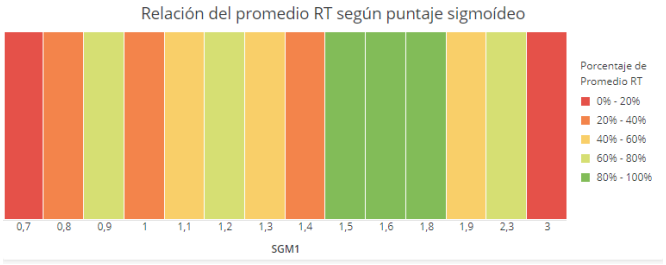
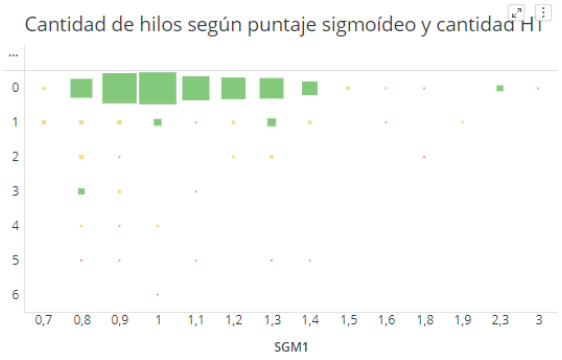


Figura 31: Relación entre promedio RT y puntaje sigmoídeo

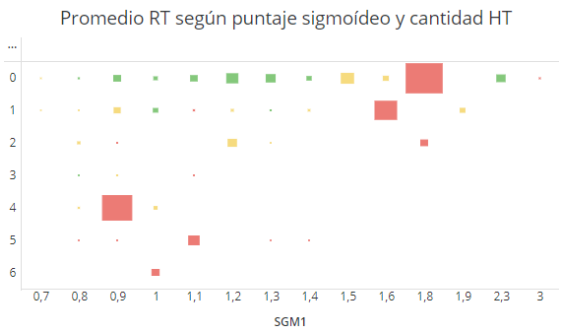
En el gráfico 32a se presenta la distribución de los hilos según la cantidad de *hashtags* utilizados y el puntaje sigmoídeo, mientras que en el gráfico 32b se presenta la distribución del promedio RT. En ambos casos los colores indican la distribución de los hilos, es decir, el porcentaje que representa esta cantidad en relación al total. El color rojo indica un porcentaje entre [0 , 35], el color amarillo [36 , 65] y el verde [66, 100], mientras que el tamaño depende de la variable a estudiar.

Del gráfico 32b se observa que sólo las combinaciones con una baja cantidad de hilos poseen valores superiores de promedio RT. Aún así, no deja de ser interesante observar que hilos con 4 HT publicado por usuarios con puntajes promedio obtengan una cantidad similar de promedio RT que los publicados por usuarios de mejor reputación.

Luego de analizar el comportamiento entre la reputación y la cantidad de *hashtags*, se procede a estudiar la relación del promedio RT según la reputación y la cantidad de Followers a través de la figura 33. Se observa que según aumenta la cantidad de FWR, también lo hace el promedio RT que presentan los hilos; también es posible comprobar el comportamiento mencionado en la figura 31 donde se indica que los usuarios con baja reputación no poseen buenos resultados de Retweets. Se puede comparar la figura 33 con la figura F1 presente en el anexo, en el que se observa que la gran mayoría que los usuarios se encuentran en la esquina superior izquierda, la cual indica una baja popularidad y reputación. Finalmente se puede concluir que la influencia del usuario cumple un rol importante en el promedio RT.



(a) Distribución de hilos según puntaje sigmoídeo y cantidad de *hashtags*



(b) Total de RT según puntaje sigmoídeo y cantidad de *hashtags*.

Figura 32: Comportamiento de promedio RT según puntaje sigmoídeo y HT.

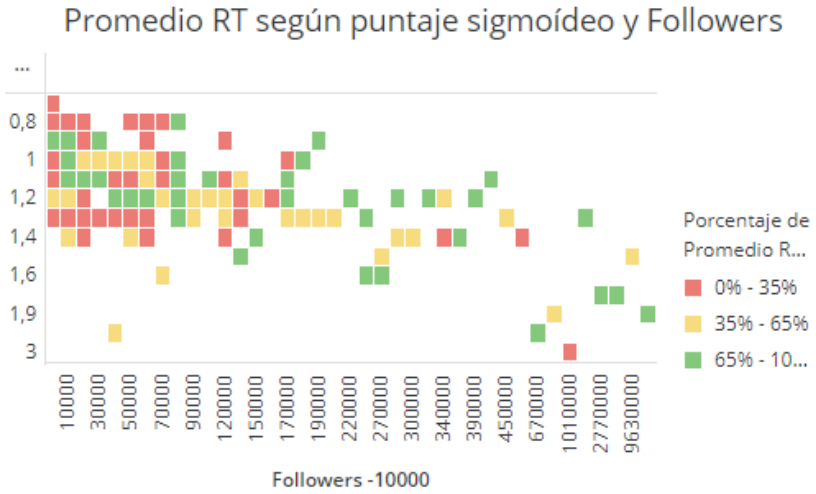


Figura 33: Promedio RT según cantidad de Followers y Puntaje sigmoídeo.

4.4.7. Impacto vs emoción

Se recuerda de la figura 20, que las 2 emociones más populares en la plataforma son *Joy* y *Fear*, aunque en la figura 34 se ve que ambas emociones son las que poseen un promedio inferior, mientras que *Sadness* y *Surprise* son las emociones con mejor promedio RT. Eso sí, al analizar la tupla de emociones en la figura F4 presente en el anexo, se observa como ambas emociones principales poseen el mayor valor en la combinación con *Fear*, por lo que finalmente se concluye que las tres emociones principales para ese gráfico son *Sadness*, *Surprise* y *Fear*.

Al comparar el promedio RT con la emoción y la cantidad de *hashtags* utilizados, a través de la figura 35 se puede observar que si bien la mayoría de las emociones no presentan HT, las que sí lo hacen no ven afectadas de manera negativa sus RT, sino más bien presentan un

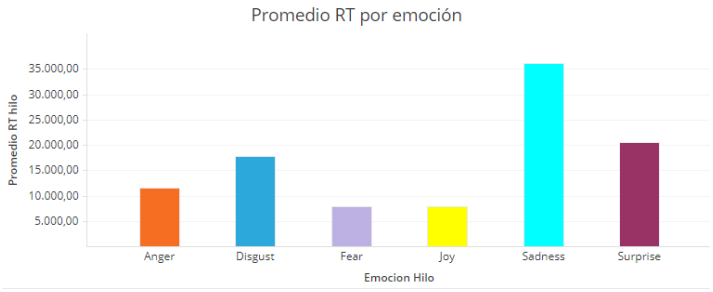


Figura 34: Promedio RT según emoción principal.

comportamiento levemente constante.

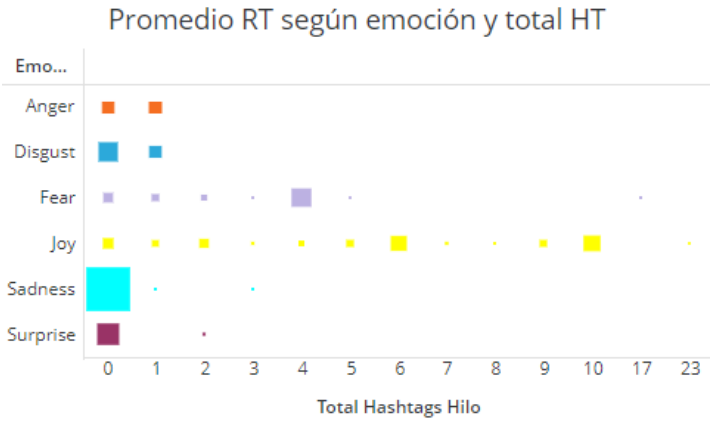


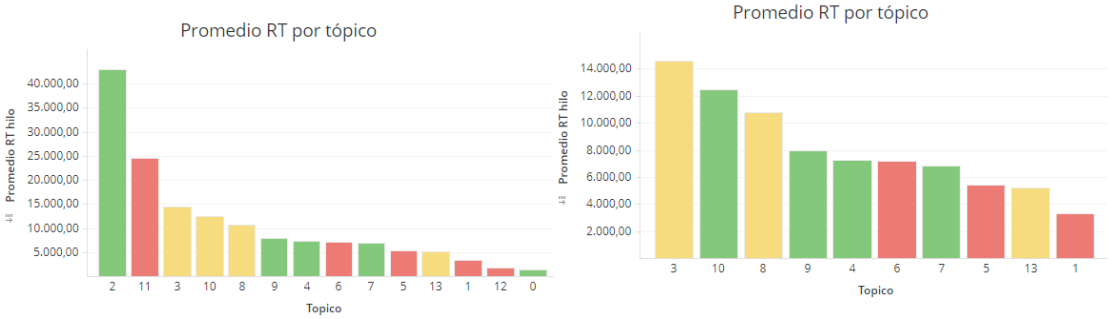
Figura 35: Promedio de RT según emoción principal y cantidad de *hashtags*.

4.4.8. Impacto vs tópico

Se presentan los gráficos de la figura 36 para comparar los tópicos según el promedio RT; el color de las barras indica el porcentaje del total de la variable estudiada que posee cada tópico, la asignación entre los colores y el porcentaje se presenta en la leyenda adjunta a la figura.

Se observa de la figura 36a que los tópicos con mejor promedio RT son los tópicos 2 (familia, hijos y salud) y 11 (escándalo rusia/elecciones Trump) con 43115 y 24529 RT, respectivamente, mientras que los peores son los 12 (economía, tecnología y negocios) y 0 (politica estadounidense, partidos políticos y reformas) con 1731 y 1444 RT cada uno. Luego de excluir estos 4 tópicos, se tiene el gráfico 36b donde se observa que el rango del promedio RT sigue siendo amplio $[3320, 14565]$, con los nuevos 2 máximos los tópicos 3 (noticias y novedades) y 10 (guerra e intervencionismo EEUU en el medio oriente), con 14565 y 12483 RT, mientras que el nuevo mínimo es 1 (Racismo) con 3320 RT.

Finalmente, se observa que en general los temas con mayor impacto son los de sucesos más cercanos al común de la gente, como lo son familia y salud, en conjunto con los escándalos.



(a) Promedio RT según tópico.

(b) Promedio RT según tópico, excluyendo los 2 mayores y los 2 menores.

Figura 36: Comportamiento de promedio RT según tópico.

4.4.9. Tópico vs emoción

A partir los resultados de la sección 4.4.8, se procede a analizar el comportamiento del promedio RT según la emoción principal para los mejores y peores tópicos:

- Mejores tópicos: 2, 11, 3 y 10
- peores tópicos : 0, 12 y 1

Mientras que las mejores y peores emociones según el gráfico 34 son:

- Mejores emociones: *Sadness*, *Surprise* y *Disgust*.
- Peores emociones: *Joy* y *Fear*.

Al igual que en los gráficos anteriores, en la figura 37 los colores indican el comportamiento de la variable de estudio en relación a todos los casos presentados, siendo verde el de valores superiores y rojo inferiores, según la escala presentada a la derecha del gráfico.

En la figura 37 se observa que para los 2 mejores tópicos (2 - Familia y salud y 11 - Rusia/Trump) todas las emociones que poseen presentan un buen promedio RT. Para el tercer mejor tópico (3 - Noticias y novedades) todas las emociones entregan buenos resultados, exceptuando el *Anger*. Para el cuarto tópico (10 - Medio oriente) se tiene que las emociones con mejores resultados son *Anger* y *Fear*.

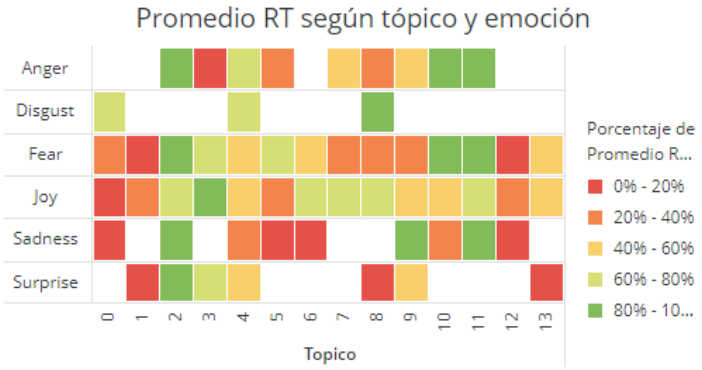


Figura 37: Promedio de RT según tópico y emoción

Para dos de los 3 peores tópicos (12 - Economía y 1 - Racismo) no existe una emoción con buenos resultados, mientras que para el último tópico (0 - Política) sólo la emoción *Disgust* entrega buenos resultados.

En relación a los demás tópicos, todos poseen comportamientos promedio, exceptuando el tópico 8 - Feminismo, el cual tiene un buen desempeño sólo en *Disgust* (*Disgust*).

Se observa que si bien *Sadness* y *Surprise* son las mejores emociones según la figura 34, en la figura 37 estas presentan resultados pobres en la mayoría de los tópicos, excluyendo solamente los tópicos 2, 9 y 10, los cuales presentan buenos resultados, para la segunda emoción sólo existe un tópico destacable, 2; en relación a la tercera mejor emoción, *Disgust*, se observa que todos los tópicos que presenta tienen buenos resultados. Al analizar las peores emociones, se observa que tienen un comportamiento muy variado, con combinaciones buenas y otras malas.

Figura 38: Distribución del promedio de RT según tópico y emoción.

Continuando con el análisis anterior, es posible ver en la figura 38 como se distribuye el total del promedio RT según cada tópico y emoción para los tópicos analizados. Se observa que para la mayoría de los tópicos existe una emoción principal, presentada en la tabla 6. Para

el resto de los tópicos se encuentra el gráfico F2 y la tabla T2, ambos en el anexo. Por como se definieron las variables de impacto, por ejemplo Promedio RT, se asume que la emoción predominante de un tópico es la emoción que la mayoría del público posee sobre este tópico.

Topico	Descripción del tópico	Emoción
2	Familia, hijos y salud	*Sadness*
11	Escándalo sobre la manipulación de Rusia sobre la elección de Donald Trump	*Sadness*
3	Hilos sobre noticias, novedades y escándalos recientes	*Joy*
10	Guerra e intervencionismo estadounidense en el medio oriente	*Anger*

Topico	Descripción del tópico	Emoción
1	Racismo a través de la historia y en la sociedad actual	*Joy*
12	Economía, tecnología y negocios	*Joy*
0	Política actual estadounidense, discusión sobre partidos políticos y reformas migratorias	*Disgust*

Tabla 6: Mejores y peores tópicos según emoción principal.

4.4.10. Tópico vs influencia

Se desea estudiar si existen temas predilectos según la reputación del usuario, por lo que se tiene en el gráfico 39 El promedio del puntaje sigmoídeo según cada tópico, se agregan también 2 líneas de referencia que dividen en los 3 grupos de reputación definidos (bajo, normal y alto). Se observa que si bien la mayoría de los tópicos son discutidos por usuarios con reputación normal, existe un tópico mencionado principalmente por cuentas de mayor reputación, que corresponde a 11 - Escándalo de manipulación de elecciones Trump.

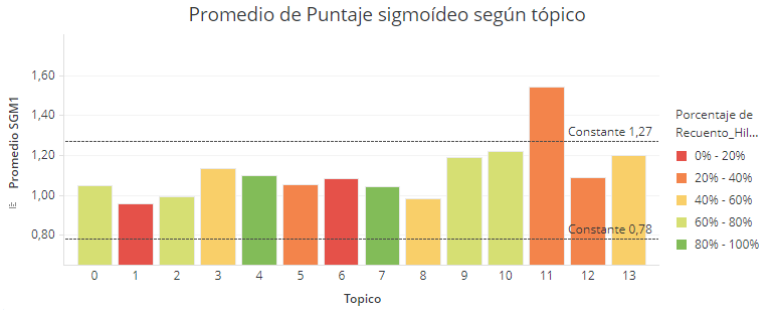


Figura 39: Promedio de puntaje sigmoídeo según tópico.

Luego de reconocer la distribución de la reputación de los usarios por tópico, se decide analizar el impacto que generan los usuarios según el tópico. En la figura 40 se divide el total

de promedio RT según el puntaje sigmoídeo y los tópicos; los colores indican qué porcentaje aporta cada segmento al total de promedio RT. Se observa que mientras menor sea la reputación del usuario, se tiende a hablar de casi todos los tópicos, mientras que los usuarios con mayor reputación sólo hablan de una cantidad mucho menor de temas. También se observa que, al igual que con otras dimensiones, entre mayor sea la reputación, mejor será el impacto promedio.

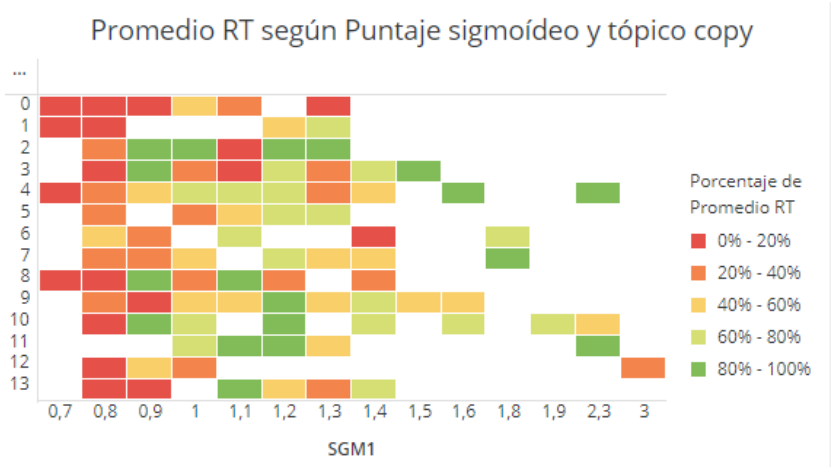


Figura 40: Promedio RT según tópico y puntaje sigmoídeo.

Se recuerda que las cuentas con una reputación mayor son principalmente entidades, las cuales, como ya fue mencionado, abarcan temáticas específicas, por lo hablan de una cantidad menor de tópicos, lo que explica el por qué se presentan menos tópicos mientras va aumentando el puntaje sigmoídeo. Es posible observar una tendencia menos mencionada a lo largo del análisis, la cual indica que para cada cuenta existe un tipo de audiencia distinta, puesto que si una cuenta define un conjunto acotado de tópicos que compartir, sus FWR desean consumir ese tipo de temas.

4.4.11. Emoción vs influencia

Para estudiar si el impacto se ve afectado en conjunto por la reputación y la emoción principal del hilo se construye el gráfico de la figura 41, en el cual el color indica el promedio RT de cada combinación en relación al total, la asociación entre el promedio RT y el color se presenta en la leyenda del lado derecho de la gráfica. En esta figura es posible apreciar que según aumenta el puntaje sigmoídeo, aumenta el promedio RT, mientras que disminuyen las emociones utilizadas. Este fenómeno es similar al observado en relación a la influencia y los tópicos.

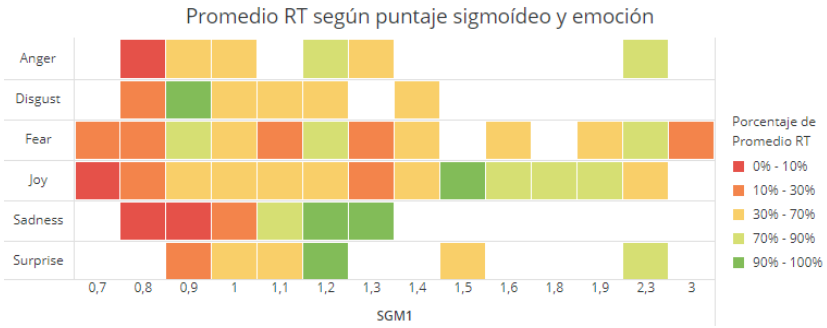


Figura 41: Impacto (Promedio RT) según reputación y emoción principal

4.5. Análisis de Relación entre Variables

Luego de estudiar el comportamiento de las variables y métricas, se decide analizar a través de diversos modelos de minería de datos el efecto que poseen éstas en relación al impacto. Para realizar esto se construye un *workflow* en la herramienta Orange Canvas, el cual se puede observar el diagrama en la figura F3 presente en el anexo.

Para el entrenamiento de estos modelos se utiliza el mismo subconjunto de datos del análisis gráfico, mientras que para la evaluación de los resultados se utiliza el resto del dataset. Para trabajar con distintos modelos se tienen que clasificar las variables según el tipo de dato, quedando finalmente de la forma presentada en la tabla 7.

Variable	Tipo	Rol
Hilo	Numérico	Metadato
Topico	Categórico	Característica
Total_tweet	Numérico	Característica
log_promedio_rt	Numérico	Objetivo
Total_hashtags	Numérico	Característica
Tupla_emocion_hilo	Categórico	Característica
User_total_fwr	Numerico	Característica
User_puntaje_sigmo	Numérico	Característica
User_verificado	Categórico	Caracterísitca

Tabla 7: Clasificación de las variables del hilo según tipo de dato y rol en el modelado.

Los modelos utilizados son máquinas de vectores de soporte (SVM por sus síglas en inglés), *random forest*, redes neuronales artificiales, k-vecinos cercanos (k-NN) y árboles de decisión.

De manera inicial se utilizan los parámetros por defecto que posee cada modelo a utilizar, luego se procede a optimizar los valores de modo que aumente el índice de correlación R^2, para luego excluir una variable y evaluar nuevamente el valor R^2, realizando este proceso para cada variable. Los resultados de estas iteraciones se presentan en la tabla 8; en ésta

tabla se observa que en todos los casos el modelo que presenta mejores resultados es SVM, con un valor R^2 cercano a 0.49 en los mejores casos.

Método	SVM	Random Forest	Redes Artificiales	kNN	Árbol
Inicial	0.347	0.203	0.135	0.197	0.000
Optimizado	0.493	0.348	0.255	0.193	0.000
Sin tópico	0.431	0.307	0.241	0.113	0.000
Sin Puntaje Sigmoídeo	0.490	0.359	0.303	0.176	0.036
Sin log FWR	0.437	0.292	0.194	-0.062	0.183
Sin tupla de emociones principales	0.405	0.341	0.361	0.224	0.000
Sin hashtags	0.361	0.293	0.125	-0.134	-0.027
Sin verificado	0.489	0.291	0.368	-0.061	0.016
Sin largo	0.493	0.351	0.406	0.331	0.116

Tabla 8: valor R^2 de cada modelo entrenado según cada caso.

Por otro lado, se tiene que al eliminar la variable largo del hilo del modelo, éste no empeora su resultado, lo que significa que no aporta mayor información. Esta conclusión se condice con el análisis de gráficos realizado.

También es digno de destacar que al excluir los *hashtags* del análisis, todos los modelos bajan su predicción de forma considerable, por lo que se considera esta variable como de gran importancia, aún cuando en el análisis gráfico no se pudo reconocer de forma concluyente un efecto real, sólo tendencias leves.

Otra variable de interés es la tupla de emociones principales, puesto que dependiendo del modelo, ésta afectará de forma positiva o negativa al resultado, eso sí; en SVM se observa que al excluirla es la segunda variable que más disminuye la precisión del modelo.

4.6. Análisis de relaciones obtenidas

Luego del análisis gráfico realizado, se procede a realizar un resumen de las relaciones obtenidas, en conjunto con un breve análisis de cada una y en general:

1. Impacto:

   a) Se comprueba que el largo no afecta sobre el promedio RT.

   b) Se reconoce que el primer y último Tweet son los más importantes en relación al aporte RT.

   Se observa que finalmente el hilo por sí solo no posee suficientes atributos para realizar un análisis muy específico sobre el impacto que posee, por lo que las recomendaciones a realizar deben ser enfocadas principalmente en las demás dimensiones de estudio.

2. Tópico:

    a) Se reconocen los temas tratados en cada tópico, aunque no es posible definir específicamente el tópico 4 puesto que su vector de palabras es demasiado amplio semánticamente.

    b) Ignorando el tópico 4, todos los tópicos presentan una cantidad de hilos similar.

    A través del análisis de las palabras principales de los tópicos se puede reconocer un tema para cada uno, el cual después se clasifica en tema general o específico, dependiendo qué tan amplia es la definición que se le entregó. Considerando esto, es posible realizar diversos análisis en otras dimensiones.

    Se puede explicar la imposibilidad de categorizar el tópico 4 debido a la gran cantidad de hilos que posee. Este fenómeno ocurre principalmente por la calidad del modelo entrenado, el cual si bien fue entrenado a través de un código que fue trabajado durante un periodo de tiempo considerable, no se logró que entregara modelos con mejores valores de *Coherence score*, el cual evalúa la similitud semántica de estos.

3. Impacto vs Influencia:

    a) Los hilos creados por cuentas con mejor reputación reciben mejores resultados de impacto.

    b) Los hilos creados por cuentas con mayor cantidad de Followers reciben mejores resultados de impacto.

    c) Se observa que los hilos de cuentas verificadas obtienen mayor cantidad de RT.

    Se observa que tanto con la reputación como con la popularidad ocurre el fenómeno de que al aumentar el valor aumenta el impacto, por lo que se encuentra necesario recalcar que si bien la reputación es una métrica que depende de los Followers, también depende de los Friends. Se identifica que la relación entre ambas es logarítmica, por lo que entre mayor la cantidad de seguidores, mayor debe ser la diferencia entre ambas variables para que la reputación aumente.

    Tal como se menciona en el marco conceptual y en el apartado de análisis de influencia, existen cuentas de entidades, las cuales por lo general poseen una reputación más elevada y una cuenta verificada, por lo que es esperable que las cuentas verificadas tengan un comportamiento similar a las cuentas con mayor reputación.

4. Impacto vs Emoción:

    a) Las mejores emociones para los hilos son *Sadness*, *Surprise* y *Disgust*.

    b) Las peores emociones para los hilos son *Joy* y *Fear*.

    Se encontró que existe una clara diferencia en el impacto cuando se estudian las emociones que presenta el hilo, por lo que se reconoce que la emoción con la que se presenta el contenido debe ser considerada al momento de generara recomendaciones de viralización.

5. Impacto vs Tópico:

    a) Los tópicos con mejores resultados son 2, 11, 3 y 10.

    b) Los tópicos con peores resultados son 0, 12 y 1.

    c) En general, son los más enfocados a la gente y su dia a dia (como lo son la familia y salud, las noticias y los escándalos) los que presentan mejores resultados.

    Se observa que existe una clara diferencia en el impacto cuando se estudia el tópico tratado en el contenido, por lo que se reconoce que además de las emociones, se debe considerar el tópico del contenido al momento de realizar recomendaciones de viralización.

6. Tópico vs Emoción:

    a) Se detectan emociones principales para cada tópico.

    Al dividir el impacto que genera cada tópico en todas las emociones que éste presenta en sus diversos hilos, se observa que en la mayoría existe una emoción claramente predominante, por lo que se genera la tabla T2 del anexo, la cual presenta esta emoción para cada tópico.

    Se reconoce que tanto el tópico como la emoción con la que se presentan los hilos definen el contenido compartido, por lo que considerando lo que se mencionó anteriormente sobre la definición de las métricas del impacto, como lo son los RT y FV (Total, Promedio y Logaritmo) indican que las acciones son realizadas cuando una persona está de acuerdo con el contenido del Tweet y desea compartirlo. Finalmente se concluye que la emoción predominante detectada en este apartado es la misma que poseen los usuarios, por lo que se debe recomendar estudiar cual es la visión que tienen los Followers sobre los contenidos que se desean compartir.

7. Tópico vs Influencia:

    a) El tópico 11 es mencionado principalmente por cuentas con mejor reputación.

    b) Se observa una tendencia que las cuentas con mejor reputación hablan de menos tópicos.

    Tal como se ha mencionado en reiteradas ocasiones, existe una tendencia a que las cuentas con mayor reputación sean entidades, las cuales hablan de una cantidad menor de temas específicos, por lo que se reconoce que el tópico 11 es mencionado principalmente por este tipo de cuentas.

8. Emoción vs Influencia:

    a) Se observa que entre las cuentas con mejor reputación se presenta una gama menor de emociones en comparación con los usuarios con menor puntaje sigmoídeo.

Como se mencionó en reiteradas ocasiones, las cuentas con mejor reputación son mayoritariamente entidades, las cuales mencionan una cantidad menor de tópicos, también se recuerda que se concluyó que cada tópico posee una emoción principal asociada, por lo que se puede concluir por transitividad que las cuentas con mayor reputación usarán una gama menor de emociones, esto explica el comportamiento descubierto.

Se debe mencionar también que si bien, a lo largo de todos los análisis gráficos, no se pudo comprobar una tendencia clara sobre la importancia de utilizar *hashtags* para el aumento del impacto, en varios de ellos se encontró una relación leve, por lo que es digno de mencionar el uso de estos en las futuras recomendaciones a realizar.

4.7. Guía de conductas para la viralización de alto impacto de contenidos en Twitter a través de hilos.

El mundo en el que vivimos, la gran mayoría de las personas pertenece a alguna red social, ya sea para estar conectado con sus seres queridos, compartir actividades y pensamientos, o simplemente estar informado de lo que ocurre día a día. Cada persona tiene un interés distinto sobre lo que comunica y lo que quiere consumir en las plataformas a las que pertenece. Esta guía de buenas conductas tiene como objetivo principal entregar pautas o recomendaciones básicas para la correcta publicación de contenidos, en búsqueda de que estos se viralicen sobre la plataforma Twitter.

La viralización de contenidos es un fenómeno al que todas las personas conectadas a alguna plataforma social han estado expuestas; éste se refiere a que una publicación sea compartida por una gran cantidad de gente en muy poco tiempo, ya sea un video de una catástrofe, un perro persiguiendo su cola. Una noticia en desarrollo, entre muchos otros ejemplos posibles. Cuando alguien realiza una publicación, lo más probable es que su objetivo sea que la mayor cantidad de gente lo consuma.

Se debe destacar, eso sí, que este fenómeno es extremadamente complejo, puesto que depende en gran parte del ambiente en donde el contenido se publica, se debe considerar el público al que se le comparte, el estado actual de la sociedad, y una inmensidad de parámetros imposibles manipular, por lo que esta guía busca plantear una base mínima para que los contenidos a publicar tengan mayores posibilidades de viralización.

Finalmente, se define como viralización de un contenido cuando éste posee un impacto superior al comportamiento del resto. En este caso, el impacto es definido por la cantidad de Retweets que recibe la publicación.

Antes de presentar las diversas recomendaciones, se encuentra útil definir dos variables muy imporantes en relación al público de una cuenta, la popularidad y la reputación. La popularidad de una cuenta es la cantidad de Followers que tiene, mientras que la reputación se define como la relación entre sus Followers y sus Friends, siendo una reputación mayor cuando la cantidad de Follower es mayor que la cantidad de Friends. Esta relación es a veces definida logarítimicamente, lo que significa que el crecimiento de Followers debe ser cada vez mayor cuando va creciendo Friends.

Tal como se mencionó anteriormente, en Twitter se encuentran cuentas de personas y cuentas de entidades, ya sean gobiernos, agencias de noticias, memes, etc. Dependiendo de esto se verán comportamientos muy distintos de la reputación y popularidad. También como todos pueden crear una cuenta y ponerle cualquier nombre, Twitter tiene un sistema para la verificación de la identidad de cuentas que dicen ser entidades/personas conocidas. Un usuario puede solicitar se reconocido por Twitter como una cuenta oficial, y luego de dicho proceso de verificación se decide si darle una insignia que lo reconozca.

A lo largo de esta guía se presentan 4 recomendaciones generales, las cuales indican una idea amplia sobre como mejorar el impacto en la plataforma, para cada una de éstas se entrega una serie de consejos para seguir de buena manera la recomendación general.

La **primera recomendación** es definir el perfil de la cuenta y reconocer su público. Luego de diversos estudios de trabajos anteriores, en conjunto con los análisis realizados, se observa el éxito de los contenidos publicados depende en parte debido al enfoque que posea la cuenta y la relación que posea entre sus Followers y Friends. Una cuenta personal es muy distinta a una cuenta de noticias, no sólo por la naturaleza de su identidad, sino que también los Followers de cada cuenta esperarán un comportamiento específico muy distinto para cada tipo de perfil.

El primer consejo para esta recomendación es definir si la cuenta a utilizar es una persona o una entidad, puesto que en diversos estudios se reconoce que el mejor comportamiento para las cuentas que representan entidades, es el de **enfocarse en los temas inherentes a la personalidad definida**, es decir, una cuenta de noticias sólo debiese tocar temas de noticias, mientras que una página de memes sólo debería hablar de memes, puesto que esto mejorará su reputación y su popularidad, dos atributos que a través del análisis realizado en este trabajo se reconocieron como importantes para la viralización de contenidos. Si se define la cuenta como el de una persona, es posible publicar sobre cualquier tema (cosa que no es tan recomendable según la literatura), pero lo más importante en esta arista es preocuparse de poseer un buen crecimiento de público, es decir, **aumentar la popularidad sin dañar la reputación**. Muchas personas, con el fin de aumentar la cantidad de Followers, siguen todas las cuentas que encuentren, esperando que éstas las sigan de vuelta, pero diversos estudios demuestran que **la reputación es más importante que la popularidad**.

Si la identidad de la cuenta es popularmente reconocida fuera de la plataforma (ya sea una persona conocida o una entidad de renombre) se aconseja **solicitar la verificación de la cuenta**, puesto que se reconoció que las cuentas verificadas poseen un impacto promedio superior por sobre las no verificadas en todas las dimensiones de estudio.

A lo largo de la primera recomendación se mencionó varias veces sobre ”los temas” o tópicos que se tocan en las publicaciones, por lo que la **segunda recomendación** es planificar el contenido de la publicación. Se realiza esta recomendación puesto que a través de los análisis realizados se observó que existe una clara influencia en el contenido de la publicación, en conjunto sobre la estructura de éste sobre el impacto generado.

En la literatura se pueden encontrar diversos trabajos que descubren tópicos a través del tiempo, que estudian cómo se comportan, sus ciclos de vida y cómo éstos van mutando, por lo que se reconoce que no se pueden recomendar temas específicos sobre los que hablar, sino que más bien, cómo decidir de cuales temas se desea hablar y como plasmarlos.

Antes de continuar con los consejos para esta recomendación, se tiene que reconocer que existen 2 casos distintos; el primero es donde ya se sabe sobre lo que se quiere hablar y sólo se necesita descubrir como hacerlo, mientras que el segundo es en donde se desea

encontrar un tópico para hablar y como hacerlo. Debido a esto, se asumirá que se desea buscar un tópico y también como tratarlo.

Cuando se realizaron los estudios sobre los tópicos que se tratan en el conjunto de datos analizado, se descubrió que la gran mayoría de los tópicos son los que están asociados principalmente a sucesos en concreto, es decir, eventos distintos que van ocurriendo esporádicamente y de forma distanciada en el tiempo, por lo que si no se tiene un tema del cual hablar, un consejo es **encontrar tópicos que sea tendencia en el momento**. Twitter posee los llamados *Trending Topics*, que son los ***hashtags* y palabras más utilizados en ese momento**.

De los temas encontrados, se observa que 2 de los primeros 3 que más impacto generaron fueron temas más genéricos, como lo son ”Familia, hijos y salud” como el tema con mejor resultado, e ”Hilos sobre noticias, novedades y escándalos recientes” como el tercero mejor, por lo que se aconseja **hablar de tópicos que generen un interés a un espectro amplio de la población**. Eso sí, se observa que otros temas amplios se encuentran entre los peores resultados, como lo es ”Economía, tecnología y negocios”, fenómeno del cual se hablará más adelante.

Luego de definir el tópico a tratar, se debe reconocer la mejor forma de presentar esta información. Primero que todo, se asume que se desea publicar el completamente en Twitter en formato de texto, es decir, se escribirá toda la información, análisis, comentarios y todo lo pertinente en múltiples Tweets consecutivos, formando así un hilo de Tweets, siendo posible adjuntar imágenes que apoyen el texto, pero el efecto de éstas no se evaluó en esta guía. Al momento de escribir el hilo, hay que tener en cuenta que tanto **el primer y último Tweet son los más importantes**, siendo por lejos el primero quien genere el mayor aporte al impacto total del hilo, por lo que se debe realizar una **introducción llamativa y que invite a la lectura**, además de escribir una **conclusión impactante y que genere diálogo** en el último Tweet. Ambas recomendaciones buscan maximizar el comportamiento natural que poseen ambos Tweets.

Se mencionó el largo del hilo pero no se declaró nada sobre valores recomendados; esto es debido a que no se encontró ningún comportamiento concluyente que afecte el impacto, por lo que se sugiere que el hilo se extienda lo necesario para que el tema se explaye de la mejor forma posible y que permita seguir las demás recomendaciones entregadas en esta guía.

Cuando se comunican contenidos a través de un medio es posible plasmarlos de diversas formas utilizando ciertas palabras por sobre otras para generar una opinión o sentimiento específico al receptor y, dependiendo de esta emoción generada, lograr que el receptor responda como el autor lo desee, por lo que la **tercera recomendación** es comunicar con una emoción acorde al tópico tratado.

Cada persona proviene de una realidad distinta, con vivencias personales y opiniones generadas a través de sus procesos de vida, por lo que no todos pensarán lo mismo sobre temas específicos. Eso sí, el ser humano vive en sociedad, lo que significa que tendrán experiencias

en común, lo que genera ciertas opiniones parecidas entre personas con vivencias similares.

En este punto ya se reconoció tanto el público al cual se le está hablando, como el tópico que se desea comunicar, por lo que el paso natural es definir cómo se desea comunicar. En esta guía se reconocen 6 emociones para comunicar contenidos: *Joy*, *Disgust*, *Anger*, *Fear*, *Surprise* y *Sadness*.

Debido a que se estudió un conjunto de datos estáticos, las conclusiones presentadas en este trabajo se consideran como los promedios para los tópicos y públicos, por lo que para cada cuenta puede variar levemente el resultado.

Algunos tópicos estudiados dignos de destacar debido a las emociones con las que se presentan y el impacto que generaron, estos son ”2 - Familia, hijos y salud”, ”3 - Hilos sobre noticias, novedades y escándalos recientes”, ”8 - Feminismo, disidencia sexual e identidad de género”, ”10 - Guerra e intervencionismo estadounidense en medio oriente” y ”12 - Economía, tecnología y negocios”.

Se obtuvo que el tópico 2 genera un gran impacto en las 5 emociones que presenta en los diversos hilos (No presenta hilos con *Disgust*); esto quiere decir que el impacto que genera sobre el público no depende tanto de la emoción sino que por sí mismo ya es interesante. Se tiene también que el tópico 3 presenta un buen comportamiento en *Joy*, *Fear* y *Surprise*, mientras que al ser presentado con *Anger* entrega malos resultados. Al analizar el tópico 8 se observa que *Joy* y *Disgust* presentan un impacto considerable, caso contrario ocurre con *Surprise*, *Fear* y *Anger*. En respecto al tópico 10 se observan buenos resultados tanto en *Anger* como en *Fear*, y malos resultados en *Sadness*. Finalmente, se tiene que el tópico 12 presenta malos resultados en las 3 emociones que presenta (*Joy*, *Fear* y *Sadness*).

Tal como fue mencionado anteriormente, los tópicos 2, 3 y 12 son temas más generales, de los cuales los primeros dos presentan buenos resultados y el tercero sólo malos resultados. Al analizar las emociones que presentan, se observa que si bien el tópico 12 presenta emociones en donde los otros dos le va bien, no logra generar un impacto deseado, lo que se puede deber a que el tema no es interesante. Finalmente, **al hablar de tópicos generales no es prioritario enfocarse en una emoción específica.**

En contraste con los otros tópicos mencionados, se observa que tanto los números 8 como 10 son mucho más específicos, referenciando temáticas completamente distintas y con distinto tipo de público. Lo interesante de ambos tópicos es que es posible visualizar que dependiendo del tópico una emoción será mejor que otra, puesto que se tiene por ejemplo *Fear* y *Anger* generando un impacto positivo en el tópico 10 pero uno negativo en el 8; caso contrario ocurre con *Joy*. Se reconoce finalmente que cada tópico específico tiene una carga emocional específica, por lo que **al hablar de tópicos específicos se recomienda hablar con la carga emocional acorde al público definido.**

Si se desea hablar de un tópico específico, pero no se reconoce una emoción principal para el público definido, o si el objetivo es publicar contenidos de un tópico general que no sea

muy interesante y se desea apelar a los sentimientos del público, se tiene que de manera estadística las primeras dos emociones con mayor impacto promedio son *Sadness* y *Surprise*, en algunos casos apareciendo en conjunto con *Fear*, por lo que **cuando no es posible definir una mejor emoción, se aconseja utilizar *Sadness* o *Surprise***.

Es digno de rescatar también que para la técnica de estudio utilizada en esta guía de buenas conductas, la cantidad de *hashtags* utilizados no es determinante para el impacto; pero aún así permitirá aumentar el público que visualice el contenido, por lo que se recomienda **utilizar a lo menos 1 *hashtag* referente al tópico**.

Finalmente, se recuerda que las recomendaciones realizadas están basadas en casos específicos, por lo que si bien se espera un aumento en el impacto de los contenidos compartidos si se siguen las observaciones realizadas, éstas no aplican a todos los casos existentes ni por existir, por lo que se considera como **cuarta y más importante recomendación** hacia el autor de los contenidos a viralizar, el **estudio constante de su público, las tendencias en los tópicos y la carga emotiva que posee cada uno**.

Se resumen las recomendaciones con sus respectivos consejos en la tabla 9.

Recomendación	Consejos
Definir el perfil de la cuenta y reconocer su público	▪ Enfocarse en los temas inherentes al perfil definido. ▪ Aumentar la popularidad sin dañar la reputación. ▪ Solicitar la verificación de la cuenta.
Planificar el contenido de las publicaciones	▪ Encontrar tópicos que sean tendencia en el momento de publicación. ▪ Utilizar *hashtags* y palabras tendencia en el momento de publicación. ▪ Hablar de temas específicos que generen interés a un espectro amplio de la población. ▪ Desarrollar en el primer Tweet una introducción llamativa que invite a la lectura. ▪ Escribir en el último Tweet una conclusión impactante que genere diálogo.
Comunicar el contenido con una emoción acorde al tópico tratado	▪ Al hablar de tópicos específicos se recomienda hablar con la carga emocional acorde al público definido. ▪ Al hablar de tópicos generales no es prioritario definir una emoción específica. ▪ Cuando no se logra definir una mejor emoción, se sugiere utilizar *Sadness* o *Surprise*.
Estudio constante de la plataforma.	▪ Estudiar constantemente el público definido. ▪ Estar pendiente de las tendencias de los tópicos. ▪ Reconocer las emociones para cada tópico estudiado.

Tabla 9: Resumen de las recomendaciones y sus respectivos consejos

CAPÍTULO 5
CONCLUSIONES

Este trabajo tuvo como objetivo principal el diseño y validación de una guía de conductas para la viralización de alto impacto de contenidos en la plataforma Twitter a través de hilos, por lo que se realizaron diversos análisis sobre un conjunto de datos recolectados específicamente para esto.

Para realizar este objetivo se definió que los contenidos presentes en Twitter presentan dimensiones interrelacionadas, como lo son la influencia del usuario, el tópico de la publicación y la emoción con la que estos se presentan. Para reconocer patrones de viralización, se tuvo que definir un objetivo a maximizar, por lo que se presentó como impacto de una publicación el alcance que estos logran dentro de la plataforma. Todas estas definiciones fueron inspiradas sobre diversos estudios anteriores presentes en la literatura. Además de lo anterior, se diseñó una metodología basada en estos trabajos mencionados.

Se reconoce que el objetivo principal fue cumplido completamente, puesto que a través de los resultados obtenidos durante esta metodología se pudo construir la guía planteada.

Para el diseño de la guía, se presentaron diversos objetivos que buscaban encaminar el trabajo, el primero de estos fue estudiar el impacto generado según el largo, la emoción y el tópico de los hilos, para identificar factores en común en los mensajes populares. Al momento de definir la estructura de los datos de estudio, se reconocío como atributo ”el largo”, siendo este la cantidad de Tweets que conforman los hilos, los cuales fueron definidos como los mensajes de estudio. Tanto el tópico como la emoción fueron dimensiones de estudio con sus propios atributos. Estas dimensiones fueron estudiado tanto en conjunto como por separado, en ambos casos obteniendo relaciones clave para la realización del objetivo principal. Se reconoce que este objetivo también fue completado exitosamente.

El segundo objetivo específico fue estudiar la relación entre el autor y el impacto de los hilos, para reconocer la importancia del autor en relación al impacto del contenido. A lo largo del trabajo se definió como otra dimensión de estudio la influencia del usuario, la cual para el análisis realizado dependió tanto de su reputación como de sus seguidores. Se logró reconocer que el autor si presenta un efecto sobre el impacto, logrando descubrir los patrones de ese fenómeno. Debido a los resultados obtenidos en estos análisis, en conjunto con la utilidad que presentaron los patrones, se puede considerar que este objetivo también fue cumplido completamente.

El tercer objetivo específico menciona la necesidad de analizar la relación de Likes, Retweets y respuestas del Tweet principal y los siguientes en un hilo, para estudiar el comportamiento viral de estos. Como se mencionó anteriormente, se definió la métrica de impacto a través de la cantidad de Retweets, también se demostró la relación que estos poseen en conjunto con los Likes a través de una regresión lineal, por lo que finalmente se consideró sólo necesario

el estudio del primero. Respecto a las respuestas, no fue posible estudiar este valor, puesto que la plataforma no entrega herramientas para la recolección de este. Además de definir el impacto y sus métricas, fue posible descubrir la influencia que poseen el primer y último Tweet. Se considera que no fue cumplir en totalidad el objetivo, pero si en gran porcentaje.

EL último objetivo específico definido fue la necesidad de examinar y asociar relaciones descubiertas entre variables de estudio, para proponer buenas prácticas de difusión de contenido. Luego de realizar los diversos análisis, se resumieron las relaciones descubiertas, se elaboraron asociaciones entre ellas y se construyó la guía de conductas planteada, por lo que se considera que este objetivo también fue cumplido exitosamente.

Uno de los aportes generados a través del trabajo realizado es la metodología que se diseñó para enfrentar este desafío. Esta fue elaborada a partir del análisis de múltiples trabajos encontrados en la literatura y fue puesta a prueba en el trabajo en sí. Se observa que se lograron cumplir completamente la mayoría de los objetivos, tanto el general como 3 de los 4 específicos, obteniendo resultados interesantes, por lo que se considera que esta metodología planteada es útil para posibles trabajos futuros.

Se debe mencionar que se utilizaron diversos *Software* y herramientas para el proceso de trabajo. En relación a la base de datos, se puede decir que MongoDB fue un gran aporte para el manejo de los datos, puesto que estos poseían una estructura acorde a como el motor maneja los registros. En relación al análisis gráfico, se puede concluir que el uso de la herramienta Microstrategy fue una decisión correcta, puesto que gracias a la funcionalidad incluida de conexión con la base de datos utilizada, no se presentaron mayores problemas con el manejo de la información; también es digno de destacar que gran parte del trabajo fue basado sobre análisis de gráficos, el uso de esta herramienta permitió un avance expedito. En relación al uso de Orange, se puede decir que las funcionalidades entregadas fueron suficientes para el trabajo realizado, su modalidad de *drag and drop* facilitó la construcción de los modelos y su análisis respectivo, eso sí, se considera que si el trabajo fuese principalmente basado en este tipo de técnicas, puede que esta no sea la mejor herramienta, puesto que al aumentar de forma drástica la cantidad de datos a analizar, este presentaba algunos problemas de rendimiento y estabilidad. Se concluye también que el trabajo realizado es un aporte tanto en los resultados obtenidos como en relación a la estructura del trabajo en sí.

A lo largo del trabajo realizado se observaron algunas influencias que no se lograban plasmar completamente a través de las dimensiones definidas, como lo era el público de cada cuenta generadora de contenido. Debido a que un usuario decide seguir a otra cuenta, el público de cada una presenta características distintas. Este fenómeno no fue considerado a lo largo del estudio, puesto que las únicas métricas que se poseían era más que nada sobre la cantidad de estos, y no su comportamiento en sí. También se reconoce que otro fenómeno similar ocurre cuando se habla de las emociones, puesto que en esta dimensión sólo se habla de la presentación del contenido a través de estas, pero en ningún momento se estudia el sentimiento que finalmente genera sobre el público el contenido presentado. Se dejan planteados estos problemas como consideraciones para posibles trabajos a futuro.

Una observación importante de destacar sobre el alcance de este trabajo es que todos los estudios realizados fueron enfocados a un público angloparlante, puesto que sólo se recolectaron hilos escritos en inglés. Si bien el idioma en sí no es considerado una variable de estudio, este si condiciona los autores seleccionados, los tópicos presentes y el público que los consume, por lo que no es posible afirmar de forma automática que los resultados obtenidos a través de los análisis realizados sean correctos para todo el mundo, ya que que existen una inmensidad de diferencias entre culturas, eso sí, se debe destacar que la guía de conductas fue presentada de tal forma que la mayoría de las recomendaciones pueden ser extrapolables a distintos ambientes gracias a la generalidad que se dio a las recomendaciones.

Es posible concluir que la realización del trabajo fue de gran aporte para el futuro, puesto que permitió la profundización de contenidos de interés y que cada día toman más fuerza en el mercado laboral, como lo es la minería de datos. Se reconoce también que múltiples conocimientos adquiridos a lo largo de la carrera fueron de gran aporte para la elaboración de esta obra. Si bien no es posible mencionar un solo curso dictado por malla que hay generado un gran aporte a este trabajo, se reconce que la estructura de la mayoría de los cursos entrega una de las habilidades más importantes de todas, el diseño y planificación para la resolución de conflictos. Frente a todos los problemas que fueron surgiendo a lo largo del trabajo nunca se llegó a un punto en donde no se supiese como proseguir, puesto que se poseía la experiencia necesaria para poder replantearse el problema, reconocer que estaba ocurriendo y lograr aprender como solucionarlo.

En relación a los ramos electivos que fueron tomados, se considera que base de datos avanzadas entregó tanto los conocimientos necesarios para definir el manejo de los datos, como también alternativas para la concepción de estos.

También se considera muy importante haber cursado inteligencia de negocios, puesto que entregó los conocimientos tanto para la visualización de los datos, como la interpretación de estos y el manejo de las herramientas utilizadas en esta memoria.

Otro ramo de gran utilidad fue minería de datos, fue aquí donde se conoció por primera vez la gran utilidad e información que pueden llegar a entregar datos que a primera vista no significan mucho para el fenómeno de estudio. Sirvió mucho también formar la mentalidad para realizar las preguntas más acertadas frente a los datos.

Finalmente, el último curso que se desea mencionar es tecnologías de búsqueda en la Web. En este curso se aprendió que es posible realizar una gran cantidad de estudios sobre texto plano, obteniendo una gran cantidad de conocimientos que fueron aplicados en este trabajo. También es muy interesante destacar que fue en este curso donde por primera vez nació la idea de realizar un trabajo como este, por lo que se considera que ha sido uno de los ramos que mayor impacto generó en el transcurso universitario.

REFERENCIAS BIBLIOGRÁFICAS

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

[Canales y Martínez-Barco, 2014] Canales, L. y Martínez-Barco, P. (2014). Emotion detection from text: A survey. En *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pp. 37–43.

[Cha *et al.*, 2010] Cha, Meeyoung and Haddadi, Hamed and Benevenuto, Fabricio and Gummadi, P Krishna and others (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30.

[Colnerič y Demsar, 2018] Colnerič, N. y Demsar, J. (2018). Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, pp. 1–1.

[Drakopoulos *et al.*, 2017] Drakopoulos, G., Kanavos, A., Mylonas, P., y Sioutas, S. (2017). Defining and evaluating twitter influence metrics: a higher-order approach in neo4j. *Social Network Analysis and Mining*, 7(1):52.

[Eysenbach, 2011] Eysenbach, G. (2011). Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. *J Med Internet Res*, 13(4):e123.

[Garcia *et al.*, 2017] Garcia, D., Mavrodiev, P., Casati, D., y Schweitzer, F. (2017). Understanding popularity, reputation, and social influence in the twitter society. *Policy & Internet*, 9(3):343–364.

[Hasan *et al.*, 2014] Hasan, M., Rundensteiner, E., y Agu, E. (2014). Emotex: Detecting emotions in twitter messages.

[Lau *et al.*, 2012] Lau, J. H., Collier, N., y Baldwin, T. (2012). On-line trend analysis with topic models:\# twitter trends detection topic model online. *Proceedings of COLING 2012*, pp. 1519–1534.

[Mohammad, 2012] Mohammad, S. M. (2012). # emotional tweets. En *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 246–255. Association for Computational Linguistics.

[Petrović *et al.*, 2010] Petrović, S., Osborne, M., y Lavrenko, V. (2010). Streaming first story detection with application to twitter. En *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pp. 181–189. Association for Computational Linguistics.

[Roberts *et al.*, 2012] Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., y Harabagiu, S. M. (2012). Empatweet: Annotating and detecting emotions on twitter. En *LREC*, volumen 12, pp. 3806–3813. Citeseer.

[Tan, 2018] Tan, P.-N. (2018). *Introduction to data mining*. Pearson Education India.

ANEXOS

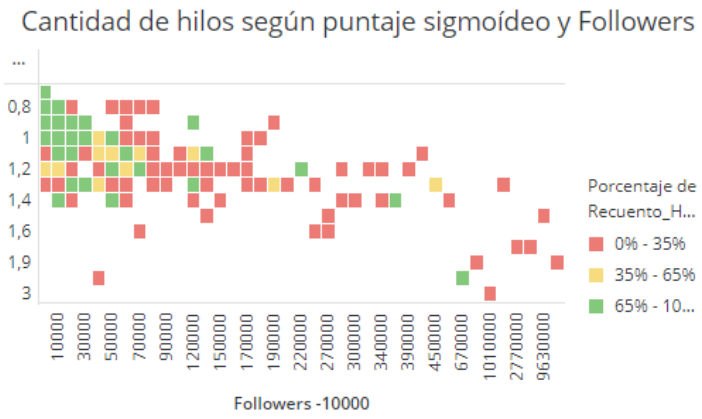


Figura 1: Distribución de hilos según cantidad de Followers y Puntaje sigmoídeo.

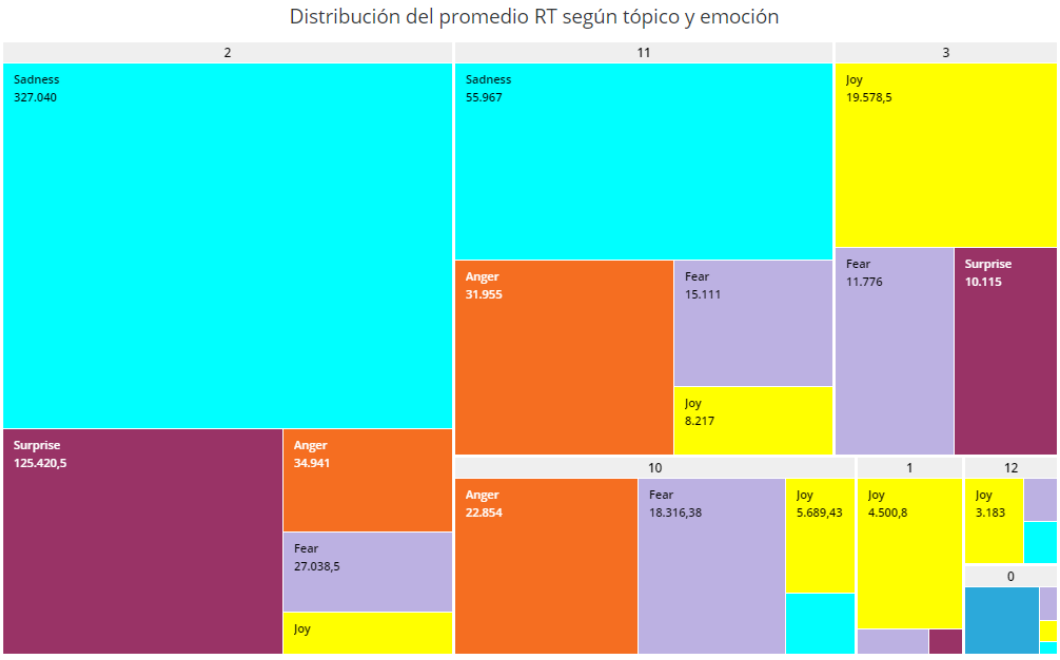


Figura 2: Distribución del promedio de RT según tópico y emoción.

Tópico	Hilos	Posible tema según palabras	6 primeras palabras
Tópico 0	23	Política actual estadounidense, discusión sobre partidos políticos y reformas migratorias.	American, Conservative, People, Democrat, Republican, Vote.
Tópico 1	15	Racismo a través de la historia y en la sociedad actual.	Black, White, Love, Life, Truth, Power.
Tópico 2	27	Familia, hijos y salud.	Child, Kid, Day, Work, Remember, Story.
Tópico 3	17	Hilos sobre noticias, novedades y escándalos recientes	Article, Thread, Report, Twitter, News, Claim.
Tópico 4	127	No es posible reconocer un tema en específico.	People, Make, Work, Time, Good, Thing.
Tópico 5	19	Escándalo sobre la influencia de rusia sobre la política de EEUU.	FBI, Russian, Agent, Report, Russia, Inteligence.
Tópico 6	11	Cambio climático y políticas públicas sobre energías renovables.	Power, Tax, Neuron, Air, High, Policy.
Tópico 7	35	Política internacional y gobierno.	Law, State, President, Order, Official, Office.
Tópico 8	16	Feminismo, Disidencia sexual e identidad de género.	Woman, People, Sex, Man, Point, Entry.
Tópico 9	31	Noticias sobre el reporte Mueller sobre actos ilícitos de Donald Trump.	Investigation, Report, Muell, Flynn, Crime, Mueller .
Tópico 10	22	Guerra e intervencionismo estadounidense en el medio oriente.	Kill, Jew, Back, People, War, Man.
Tópico 11	13	Escándalo sobre la manipulación de Rusia sobre la elección de Donald Trump.	Trump, Campaign, President, Russian, Russia, Email.
Tópico 12	17	Economía, tecnología y negocios.	Money, Year, Pay, Company, Sell, Make.
Tópico 13	18	Guerra comercial y conflictos políticos en medio oriente e India.	Regime, Iran, Datum, Iranian, India, Build.

Tabla 1: Clasificación de los tópicos modelados.

Topico	Emoción	Descripción
2	Tristeza	Familia, hijos y salud
11	Tristeza	Escándalo sobre la manipulación de Rusia sobre la elección de Donald Trump
3	Alegría	Hilos sobre noticias, novedades y escándalos recientes
10	Enojo	Guerra e intervencionismo estadounidense en medio oriente
8	Disgusto	Feminismo, disidencia sexual e identidad de género
9	Tristeza	Noticias sobre el reporte de Mueller sobre actos ilícitos de Donald Trump
4	No	No se reconoce tema
6	Alegría	Cambio climático y políticas públicas sobre energías renovables
7	Alegría	Política internacional y gobierno
5	Miedo	Escándalo sobre la influencia de Rusia sobre política de EEUU
13	Miedo	Guerra comercial y conflictos políticos en medio oriente e India
1	Alegría	Racismo a través de la historia y en la sociedad actual
12	Alegría	Economía, tecnología y negocios
0	Disgusto	Política actual estadounidense, discución sobre partidos políticos y reformas migratorias

Tabla 2: Tabla de tópicos ordenados de forma descendente según impacto, se presenta su emoción principal y la descripción del tópico.

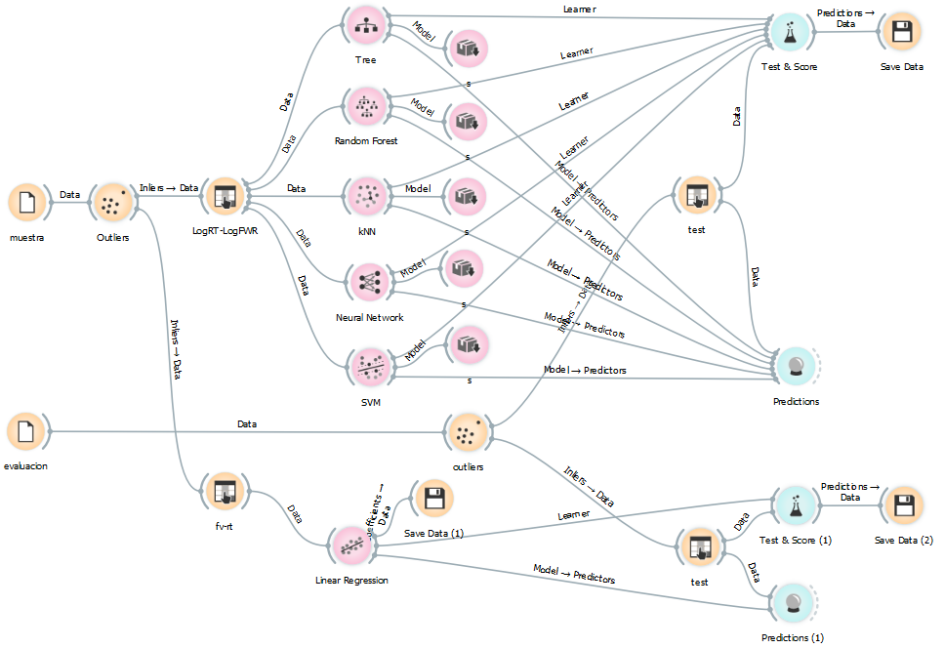


Figura 3: Workflow de Orange utilizado para la construcción y evaluación de diversos modelos de minería de datos.

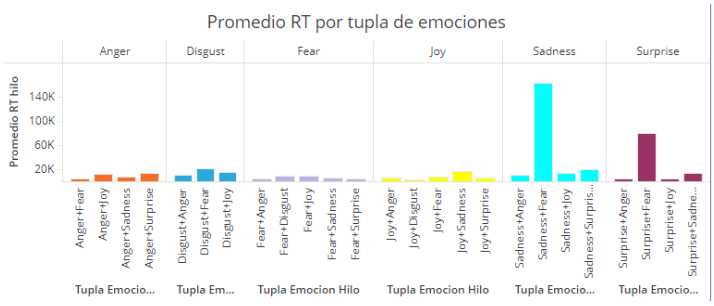


Figura 4: Promedio RT según tupla de emociones