Facultad de Ciencias Económicas y Empresariales

Hacia un análisis de reputación online, mediante el modelado de tópicos de la información en redes sociales, de la Universidad Pontificia de Comillas

Autor: 201707977

MADRID | junio 2022

ÍNDICE

LISTADO DE ABREVIATURAS

ACA	Análisis de contenido automatizado
API	Application Programming Interfaces. (Interfaz de programación de aplicaciones)
AS	Sentiment Analysis. Análisis de Sentimientos
CRISP-DM	Cross Industry Standard Process for Data Mining
KDD	Knowledge Discovery in Databases
K-NN	K-vecinos
K-M	K-Medio
NB	Naibe Bayes
PLN	Procesamiento de Lenguaje Natural
SOM	Mapas Autoorganizados
WWW	World Wide Web

RESUMEN

En el análisis de reputación online se busca extraer conocimiento a partir de datos principalmente textuales, mediante el modelado de tópicos de la información, sean "estructurados" y "no estructurados". En este caso, documentos recopilados de la red social Twitter a través de su servicio de la API y técnicas de web scraping, con la finalidad de agrupar, identificar y clasificar las temáticas más relevantes de la comunidad de la Universidad Pontificia de Comillas en Twitter. Se sigue la metodología CRISP-DM aplicada a la minería de texto junto con un procesamiento de lenguaje natural con los recursos disponibles. La finalidad de este proyecto es analizar la reputación de la universidad (análisis de sentimientos).

Palabras clave: red social, Twitter, Procesamiento de Lenguaje Natural, Minería de texto, Análisis de Sentimiento, Web scraping.

ABSTRACT

Online reputation analysis seeks to extract knowledge from mainly textual data by modelling information topics, both "structured" and "unstructured". In this case, documents collected from the social network Twitter through its API service and web scraping techniques, in order to group, identify and classify the most relevant topics of the community of the Universidad Pontificia de Comillas on Twitter. It follows the CRISP-DM methodology applied to text mining together with natural language processing with the available resources. The aim of this project is to analyze the reputation of the university (Sentiment Analysis),

Key words: social network, Twitter, Natural Language Processing, text mining, Sentiment Analysis, Natural Language, Text mining, Web scraping.

CAPÍTULO I: INTRODUCCIÓN

1.1 Motivación

La principal característica que se puede observar de una marca, es su reputación. En un mercado cada vez más competitivo, es necesario el poder recibir la información necesaria de los consumidores para transformarla y poder mejorar la eficacia del alcance de la marca o institución en cuestión.

La información es poder, y puesto que la reputación afecta a las decisiones de compra de los consumidores, nunca se debe descuidar. Como se establece en la jornada organizada por ESIC en colaboración con Global Alliance for Public Relations and Communication Management, "Reputación, RSC y Comunicación en el ecosistema digital", en España, si una marca u institución ostenta una reputación excelente, la predisposición a comprar en la misma aumenta en un 77%, si se trata de una marca con reputación media, esta predisposición disminuye a un 23%, mientras que si la marca ostenta una reputación sólida, será del 39%. Como se puede observar, la importancia de tener una reputación excelente afecta a la marca en todos sus ámbitos.

A través de este análisis, se van a poder observar diversas reacciones ante posibles crisis de reputación, mejorar la atención al cliente y usuarios, e incluso facilitar la capacidad para comprender a los consumidores en relación con los diversos discursos y directrices que se entregan.

En relación al canal en el que efectuar los análisis de sentimientos y modelados de tópicos para analizar la reputación de una marca, es interesante centrarse en las redes sociales. Éstas son muy consultadas para determinar la una visión negativa o positiva de la empresa o institución en cuestión, por lo que nos centraremos en ellas para la elaboración de este proyecto. Según el "Estudio de Redes Sociales 2019" realizado por la agencia de digital commerce marketing Elogia e IAB Spain y patrocinado por Adglow, más de un 62% de

los españoles confían en Facebook para consultar opiniones sobre marcas y empresas. YouTube, Instagram y Twitter le siguen, con un 37%, 24% y 18% respectivamente.

Las redes sociales de micro blogueo se clasifican como redes donde se publican mensajes de corta longitud, de no más de 300 o 400 caracteres, según los autores (Bermingham A. & Smeaton A., 2012), "los documentos de micro blog significan que pueden ser fácilmente publicados y leídos en una variedad de plataformas y modalidades. Esta restricción de brevedad ha llevado al uso de artefactos textuales no estándar como emoticones y lenguaje informal"**.** Su corta longitud de documento sugiere que cualquier información que contiene cada cuadro de texto, por lo general es de naturaleza ruidosa, incluyendo en ellos simbologías y emoticonos. (Los emoticonos expresan actividades, alegrías u otras emociones).

Una de las redes sociales de micro blogueo más famosas actualmente es Twitter, fundada en el año 2006 y actualmente con millones de usuarios activos cada día a nivel mundial, la plataforma Twitter goza de alta penetración en España, unos 4,2 millones de usuarios activos en 2021, encontrándose entre los 25 países con la mayor cantidad de usuarios activos a nivel mundial (Fernández, R., 2022). *Twitter: número de perfiles de la red social Twitter en España 2014-2021)*

Las redes sociales más destacadas en contenido textual, utilizadas con alta frecuencia en nuestro país y con enfoque público son Facebook y Twitter (excluyendo casos muy particulares), otras redes como Flickr, Behance, Instagram o YouTube, están más orientadas a contenido de imágenes y video respectivamente. Twitter representa para el país una ventana importante de interacción, con abundante contenido textual en información "no estructurada" y "semiestructurada", que es posible obtener tanto en flujo directo como en archivo histórico a partir del uso de los servicios gratuitos y pagados que ofrece la red a los desarrolladores, con alto contenido a múltiples.

En la investigación realizada por (Lozares C., 1996), sobre la teoría de las redes sociales, se afirma que "las redes sociales están presentes hace varias décadas, con investigaciones desde los años treinta, pero tomando fuerza en los cincuenta y sesenta cuando se realizaron numerosos pruebas para diseñar métodos, estudiar minuciosamente las relaciones sociales y descubrir sus pautas aunque muchos de estos intentos fueran

relativamente rudimentarios y no condujeran a métodos suficientemente atractivos y de sencilla comprensión para los investigadores". En buena medida todo cambia en los últimos sesenta y en los setenta con un mayor desarrollo de la base matemática, concretamente de la teoría de grafos (1965; Harary, 1969), la llegada de los algoritmos de computación hace además posible sus inicios en la implantación práctica.

En las últimas métricas publicadas por Global Digital 2018 por (We Are Social & Hootsuite, 2018), Simon Kemp muestra que la cantidad de personas en todo el mundo que utilizan las redes sociales constantemente creció en más de 100 millones en los primeros tres meses de 2018, alcanzando casi 3 mil millones para fines de marzo. (KEMP, S., 2018) "Con 4 mil millones de usuarios ahora mismo en línea (unos 3 mil millones lo desde dispositivos móviles), ya somos capaces de ver nuevos patrones de comportamiento en el ámbito digital. Los teclados se sustituirán por las tecnologías con reconocimiento de voz. El contenido audiovisual dominará las redes y los mensajes. Las nuevas tecnologías ofrecerán una mejor experiencia digital al consumidor. Las empresas necesitarán reconsiderar sus estrategias, desarrollar nuevas competencias y tener la facilidad de adaptarse rápidamente". Estos datos muestran que más de la mitad de la población mundial utiliza redes sociales, con creciente aumento en horas diarias por usuario conectado y constantemente.

1.2 Objetivos.

El objetivo de este trabajo es extraer información automáticamente sobre los temas de conversación en redes sociales, que involucran a la Universidad Pontificia de Comillas, de cara a un análisis de la imagen y la reputación online de la institución. Esta información se basará en tweets. Así, a través de la extracción de tweets en un período de más de 2 años, este proyecto servirá para infinidad de utilidades: desde determinar la reputación online de la universidad, hasta desarrollar mejores estrategias empresariales.

Aun así, y después de analizar con detenimiento este proceso, la mayor utilidad del trabajo va a ser poder acercar de una forma óptima a los jóvenes y captar así nuevos alumnos además de afianzar tanto los que se encuentran estudiando en la institución como los que ya han pasado por ella.

El estudio se dividirá en las siguientes secciones de interés:

1. Análisis descriptivo del corpus de tweets.
2. Identificación de los temas de conversación de la Universidad.
3. Semántica y clasificación de los tópicos.

1.3 Descripciones de los siguientes capítulos.

A continuación, se procederá a describir el contenido y utilidad de los siguientes capítulos:

El capítulo II, recopila toda la información teórica del análisis de sentimientos y el modelado de tópicos. Se incluye una descripción de todo lo utilizado en este proyecto, estableciendo un procedimiento teórico CRISP-DM. Además, se incluyen definiciones y explicaciones sobre el proceso de depuración de datos agrupado en *tokenización*, exclusión de palabras (*stopwords*) y determinación de palabras equivalentes o *stemming*. Se ha facilitado una breve explicación del paquete computacional R, ya que, aunque se trate de un conocimiento básico, ha sido el utilizado a través de scripts para la realización del proyecto.

En el capítulo III, se incluye la metodología a seguir para la realización del análisis de la reputación de la Universidad Pontificia de Comillas, en donde se irá haciendo mención de las librerías, paquetes computacionales y funciones de R y RStudio para la realización de este examen. Se dividirá en: preprocesamiento, depuración de datos, modelado (clustering o análisis de conglomerados) y por último análisis de sentimiento.

En el capítulo IV, se procederá a analizar los resultados generados a través de las fases de la metodología. En este apartado, se efectúan varios análisis de los gráficos generados, determinando cual ha sido la reputación de la Universidad Pontificia de Comillas desde el 1 de enero de 2018 hasta el 31 de mayo de 2022. Se discutirán y debatirán los resultados de nubes de palabras generadas, tópicos de la universidad e incluso las diversas respuestas emocionales que los usuarios de Twitter tienen en relación con la universidad y se encontrarán uniones relacionales entre ellos.

En el último capítulo, se comentarán unas conclusiones finales tanto sobre el proceso y análisis del proyecto, como sobre los resultados finales que se han obtenido de la

universidad. Se podrá determinar así la reputación de la Universidad Pontificia de Comillas.

CAPÍTULO II: ELEMENTOS TEÓRICOS

2.1 Servicios de APIs de Twitter

El acrónimo API (Application Programming Interface) o Interfaz de Programación de Aplicaciones es descrito por (Jacobson D. & Brail G., 2011), en su trabajo sobre APIs como, "una relación, conectada por módulos de software que se comunican o interactúan con otros a través de un conjunto de comandos, funciones y protocolos informáticos que permiten a los desarrolladores crear programas específicos para ciertos sistemas operativos. Una vez que se establece una relación de ese tipo, los desarrolladores se ven tentados a usar la API porque saben que pueden confiar en ella. La relación aumenta la confianza y, por ende, el uso".

De igual manera hace que la conexión entre proveedor y consumidor sea mucho más eficiente ya que las interfaces están documentadas, son consistentes y predecibles. Las API de Twitter y Facebook son ejemplos muy famosos. Hay APIs que están abiertas a cualquier desarrollador, APIs que están abiertas solo a socios y APIs que se usan internamente para ayudar a administrar mejor el contenido y facilitar la colaboración entre equipos, algunas con acceso gratis y otras no.

Twitter proporciona sus servicios de APIs estándar gratuitas y públicas, que proporcionan la funcionalidad de consulta básica y acceso fundamental a datos de Twitter, y APIs empresariales (Gnip), que brindan datos históricos y en tiempo real para impulsar las empresas a gran escala, recientemente se ha introducido API Premium, por supuesto con un acceso mucho mayor, enfocado en relaciones a gran escala. Estos servicios proporcionados por la plataforma constantemente realizan cambios para mejorar sus conexiones con los desarrolladores y para desviar en buena medida el alto contenido spam y malware generado cada día. (JACOBSON D., BRAIL G. & WOODS D., 2011), *(*DAU, A.,2017)

El contacto con los servicios de APIs está disponible para varios lenguajes de programación a través de sus librerías, las cuales están en modalidad pública en la documentación de desarrolladores de la red social. Nosotros utilizaremos R y RStudio para la generación del dataset con el que vamos a trabajar el modelado de tópicos.

Las librerías computacionales proporcionadas por Twitter para el contacto con los desarrolladores que interactúan con la red para sus fines personales y/o empresariales, están disponibles gratuitamente para todos, aunque en algunos casos con mayores privilegios. Estas bibliotecas de trabajo, aunque no necesariamente todas fueron creadas o probadas por Twitter, son compatibles con la API estándar de Twitter.

La conexión entre estas APIs de Twitter y cualquier desarrollador, requiere tener en cuenta las reglas predefinidas por las políticas de privacidad de la red, que cambian constantemente, como, por ejemplo, en su última actualización el 25 de mayo de 2018. Los lenguajes de programación más destacados son: C++, PHP, Objective-C, Swifter, Java, Tembo, .Net entre otros, con acceso a las librerías: python-twitter, tweepy, TweetPony, twitter-gobject, TwitterSearch, twython, TwitterAPI, twitter-ads, twitter-console, tweetstream, entre otras, mediante los sistemas operativos Apple, Windows y Linux.

2.2 Procedimiento para el desarrollo de proyectos en minería de datos (CRISP-DM)

Para describir la metodología CRISP-DM, según (Kalev L., 2012), es necesario indagar sobre el análisis de contenido, el cual es descrito como “Un ejercicio analítico cuyo objetivo es obtener información de cierto conjunto de datos, generalmente textos o grabaciones”. Históricamente, el análisis de contenido se ha servido de otras técnicas que mejoran su alcance y se ha venido aplicando en marcos de investigación cuantitativos, cualitativos y mixtos. A través de los métodos computacionales y del análisis de contenido automatizado (ACA) se logra mejorar los análisis de contenido tradicionales de manera más acertada a pesar de trabajar con diferentes escalas de datos.

“Los orígenes de CRISP-DM, se remontan hacia el año 1998 cuando un importante consorcio de empresas europeas tales como NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, SPSS, y Daimer-Chrysler, proponen a partir de diferentes versiones de KDD (Knowledge Discovery in Databases), el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining”, tal como lo menciona la guía de (GALLARDO ARANCIBIA, J. A. 2013). La metodología CRISP-DM sigue los lineamientos necesarios para la elaboración de un proyecto de minería de texto, siguiendo seis fases, descritas a continuación.

Fase de comprensión del problema

Determinar los objetivos del problema. Este es el primer paso para desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar minería de datos y definir los criterios de éxito. Los problemas pueden ser diversos cómo, por ejemplo, detectar fraude en el uso de tarjetas de crédito. En este caso queremos analizar la reputación de la Universidad Pontificia de Comillas, en todo caso analizado anteriormente.

Fase de comprensión de los datos

La comprensión de los datos, comprende la recolección inicial de datos con el objetivo de establecer un primer contacto con el problema. Así, debemos familiarizarnos con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Siguiendo con nuestro autor (GALLARDO ARANCIBIA, J. A. 2013) “Las principales tareas para desarrollar en esta fase del proceso son:

1. Recolección de datos iniciales. La primera tarea en esta segunda fase del proceso de CRISP-DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.

2. Descripción de los datos. Después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

3. Exploración de datos. A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.

4. Verificación de la calidad de los datos. En esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea en este punto, es asegurar la completitud y corrección de los datos."

Fase de preparación de los datos

En esta fase del proceso, una vez que se ha realizado la recolección inicial de datos de una manera eficiente, según (GALLARDO ARANCIBIA, J. A. 2013) se deben preparar, limpiar y acotar para las técnicas que se utilicen posteriormente, como son los denominados análisis de sentimientos y modelado de tópicos. Como establecen (Dávila Hernández, F. y Sánchez Corales Y, 2022), La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Fase de modelado

"En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

1. El modelado que se realice deberá ser apropiado para el tema.
2. Los datos deben estar correctamente depurados y establecidos.
3. Se debe atener el cumplimiento de los requisitos del problema.
4. El tiempo para la generación del modelo no puede ser excesivo, debe ser adecuado.
5. Debe tenerse un buen conocimiento de la metodología." (GALLARDO ARANCIBIA, J. A. 2013)

Fase de evaluación

En este paso del proceso, se debe efectuar una buena evaluación del modelo. Gracias a los criterios mencionados anteriormente, se deben analizar tanto gráficos como tablas que provienen del modelado. La fiabilidad en la que se basa el modelo recae sobre los datos que se han filtrado y depurado con anterioridad, por lo que es de vital importancia efectuar los pasos anteriores de la forma más precisa posible. Una vez obtenidos los resultados, se vuelve a revisar el proceso.

Fase de implementación

En esta fase se transforma el conocimiento obtenido en acciones dentro del proceso de negocio. El encargado de efectuar el análisis debe ser consciente de todas las observaciones hechas a lo largo del estudio y por ende recomiende acciones que ejecutar. Por ejemplo, en el caso de que se efectúe un análisis de reputación de una marca concreta, una vez efectuado, el analista deberá indicar los pasos a seguir para que esa marca determinada aumente su reputación.

2.3 Teoría de preprocesamiento (Minería de texto)

La minería de texto, es una extensión de la minería de datos, también conocida como minería de datos de texto o descubrimiento de conocimiento de textual, como lo manifiesta (Ridge K., 2010) en su artículo sobre minería de textos, "se refiere al proceso de extracción de patrones interesantes y no triviales o conocimiento de los documentos de texto. Se resaltan dos puntos a tener en cuenta, primero la refinación de texto que transforma documentos de texto "no estructurados" en una forma intermedia; y posterior a ello la destilación del conocimiento que deduce patrones y similitudes".

Existen varias técnicas de estudio para afrontar la exploración de datos con la minería de textos, bien sea para grandes cantidades o pequeños cuadros de texto. Según, (Torres D., 2013), los ítems a seguir son indispensables para este tipo de estudios: tokenización, stopwords y stemming, descritos en detalle a continuación.

2.3.1 Eliminación de caracteres y simbologías (Tokenización)

Para (Torres D., 2013), "el proceso de tokenización es altamente dependiente del lenguaje sobre el cual se está trabajando y las reglas específicas de este, por lo que esta tarea resulta trivial para una persona familiarizada con la estructura del lenguaje. Para identificar las distintas palabras (o tokens) es necesario primero definir los delimitadores de los tokens, los que generalmente corresponden a los signos de puntuación y otros caracteres distintos a las letras del alfabeto. Luego los delimitadores de tokens se separan de las palabras y son reemplazados por un espacio blanco simple. De esta forma cada palabra queda separada por un espacio blanco simple y facilita la tokenización." Por lo cual el primer paso para el manejo de texto es separar la lluvia de caracteres y símbolos de las palabras contenidas en los cuadros de texto, comúnmente encontradas en textos categorizados como "no estructurados" o "semiestructurados".

2.3.2 Exclusión de palabras (Stopwords)

Stopwords, con traducción al español como "detener palabras", descrito por (Torres D., 2013), "es un método de filtración que consiste en remover palabras comunes que resultan inútiles para la caracterización de un documento. La idea de la aplicación de una

lista de stopwords es remover palabras que ocurren muy poco o no contienen información útil, como artículos, pronombres, conjunciones, preposiciones, etc." Las palabras que aparecen en la mayoría de los documentos aportan poca información para distinguir los distintos tipos de documentos, así como las palabras que ocurren muy raramente es probable que no tengan una relevancia estadística y pueden ser descartadas del diccionario. Al descartar las palabras de la lista de palabras se puede lograr una reducción significativa del tamaño del diccionario, pero no existe una lista fija de palabras que sea universalmente utilizada.

En cada país y sus diferentes regiones las sociedades de su cultura utilizan frecuentemente palabras no lingüísticas en textos clasificados como "no estructurados" y "semiestructurados". Alimentar la lista de stopwords resulta de gran aporte para el proceso de análisis de los datos.

2.3.3 Palabras equivalentes (Stemming)

Después de que el documento de texto haya sido separado en una secuencia de tokens, el siguiente posible paso es convertir cada token a una forma estandarizada, proceso llamado stemming. Para (Torres D., 2013), "la aplicación de stemming se basa en la observación de que las palabras en los documentos a menudo tienen muchas variantes morfológicas. Las palabras que tienen una misma raíz lingüística pueden ser tratadas como una única palabra, la cual entrega probablemente una mejor descripción del contenido del documento, lo que podría no ocurrir si se utilizara cada palabra individualmente. El objetivo de stemming es reconocer los conjuntos de palabras que pueden ser tratadas como equivalentes. Muchas veces no hay necesidad para mantener el singular y plural de una misma palabra, así como los verbos pueden ser almacenados es su forma infinitiva. También se puede extender el concepto hacia los sinónimos."

Algunos de los efectos de la aplicación de stemming es la reducción del número total de atributos dentro del texto (o reducción del tamaño del diccionario) y el incremento de la frecuencia de ocurrencia de algunos atributos. Así como para stopwords sí, no hay un algoritmo de stemming que sea universalmente usado, donde el idioma tiene un papel clave. Por otro lado, la aplicación de stemming debe ser implementada con cautela para

no eliminar palabras del diccionario que puedan resultar relevantes, dado que no se considera la semántica de las palabras.

2.3.4 Paquete computacional R

Evaluar con éxito un proyecto sobre minería de texto, requiere indispensablemente tanto de los datos, la meta trazada en ellos y el apoyo de paquetes computacionales. Como lo describe la documentación de su sitio informativo, "R es un lenguaje y entorno para computación y gráficos estadísticos. Es un proyecto de GNU que es similar al lenguaje y al entorno S, desarrollado en Bell Laboratories (anteriormente AT & T, ahora Lucent Technologies) por John Chambers y sus colegas". R se puede considerar como una implementación diferente de S. R está disponible como Software Libre bajo los términos de la Licencia Pública General GNU de la Free Software Foundation en forma de código fuente. Se compila y se ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluidos FreeBSD y Linux), Windows y MacOS

2.4 Clustering

El análisis clúster o análisis de conglomerados divide la base de datos generada a través del procedimiento CRISP-DM en distintos grupos relacionales. Así, puede localizarse diversos tópicos dentro de un data frame determinado. Estas agrupaciones de datos o clústeres, tienen que capturar la estructura de la base de datos que se quiere analizar. Este agrupamiento de efectúa a través de similitudes.

Para poder hablar de similitudes, en el análisis de conglomerados se suele establecer algún tipo de distancia.

Jon Kleinberg (J. Kleinberg, "*An impossibility theorem for clustering*" en Proceedings of the 15th International Conference on Neural Information Processing Systems, 2002) establece varios puntos necesarios que un clúster debe presentar, siendo estos los establecidos a su vez por (Fernando Sancho Camparrini, 2020. "Algoritmos de clustering".)

-No se deben dar unos resultados distintos si se cambia la escala del algoritmo.

-Debe tener consistencia, por lo que, si las distancias dentro de cada clúster se reducen o incluso si aumentan, el algoritmo no debería cambiar de resultados.

-Debe tener riqueza.

Pese a que estos requisitos deberían ser determinantes, el estudio de Kleinberg determina que no pueden cumplirse los tres simultáneamente, solamente se pueden crear algoritmos de conglomerados que incumplan uno de ellos, pero a su vez cumplan con los otros dos requisitos.

K-means

El algoritmo K-means (modelo de centroides) es aquel que asigna cada punto al clúster cuyo centro o centroide se encuentra más cerca. Es así como lo determinan (PUNITHAVALLI M., PUNITHA S. C., NATHIYA G. en "An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm").

Así, el algoritmo va a intentar buscar una partición de los datos en *k* agrupaciones. Es un algoritmo eficiente que requiere de un procedimiento simplificado.

Según (PUNITHAVALLI M., PUNITHA S. C., NATHIYA G. *"An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm")*, los pasos para la generación del algoritmo son:

1. Determinar el número de clústeres *k.*
2. Generar k clústeres de manera aleatoria y determinar los centroides de los clústeres.
3. Asignar cada punto a su clúster más cercano.
4. Recalcular los nuevos centros del clúster
5. Repetir los dos pasos anteriores hasta que se cumpla algún criterio de convergencia.

2.5 Análisis de Sentimientos

Los estudios sobre Análisis de Sentimiento, denominados en inglés con el acrónimo AS (Sentiment Analysis), que se han ido efectuando desde hace varias décadas y con un gran crecimiento actual, abarcan investigaciones bastante amplias y en numerosas áreas de trabajo, enfocándose principalmente en redes sociales de micro blogueo y medios digitales en forma de periódicos con contenidos textuales

Los estudios realizados sobre análisis de sentimiento en español, siguen lineamientos principalmente de modelos realizados en otros idiomas, siguiendo diferentes metodologías como por ejemplo la CRIS-DM (metodología europea aplicada a proyectos de minería de datos), K-Medio (K-M), K-Vecinos más cercanos (K-NN), Naibe Bayes (NB) y mapas autoorganizados (SOM), entre otras.

Explorar y extraer conocimiento de datos textuales en redes sociales, representa para (Dubiau L. & Ale J., 2013), en su proyecto sobre análisis de sentimiento en español, que, "el análisis de sentimiento consiste en determinar la actitud de un escrito ante determinados productos, situaciones, personas u organizaciones (objetivo); identificar los aspectos que generan opinión (características); quien las posee (titular); y cuál es el tipo de emoción (me gusta, me encanta, lo valoro, lo odio) o su orientación semántica (positiva, negativa, neutra)".

Las cargas emocionales que pueda tener un texto pueden evaluarse para extraer conocimiento, bien sea para datos estructurados, como, por ejemplo: libros, revistas, artículos, publicaciones periodísticas o bien para datos no estructurados y/o semi-estructurados, como es el caso de publicaciones en las redes sociales, sea en datos netamente textuales o en transcripciones a texto de audio o video.

En el artículo publicado por (Nakov P., 2017), el autor manifiesta que, "el análisis de sentimiento aplicado a medios sociales requiere del uso de técnicas de procesamiento del lenguaje natural para identificar y categorizar las opiniones expresadas en cada cuadro de texto, con el fin de determinar la actitud del autor hacia un tema en particular o en general. Normalmente, las etiquetas discretas tales como positiva, negativa, neutral y objetiva se

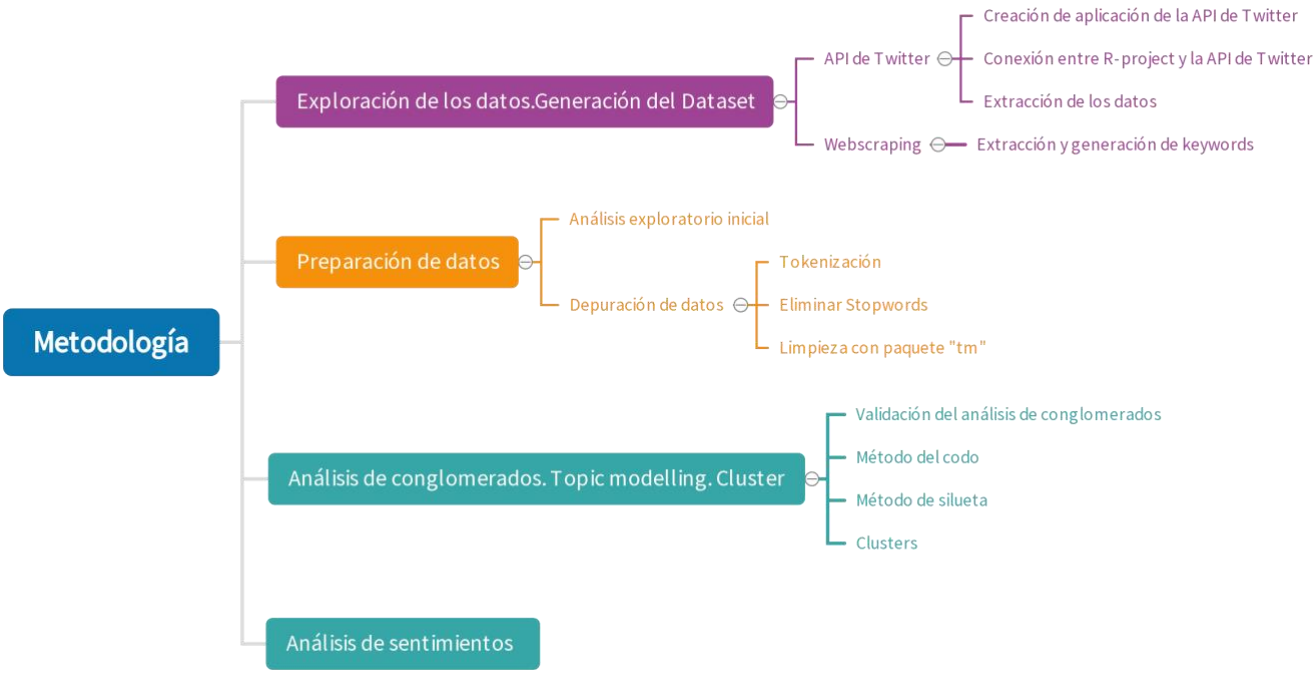
utilizan para este propósito, pero también es posible usar etiquetas en una escala ordinal, o incluso valores numéricos continuos.”

CAPÍTULO III: METODOLOGÍA

La metodología para el desarrollo de esta investigación se basa en el método y modelo de procesos CRISP-DM para la elaboración de proyectos de minería de datos, este modelo como se ha explicado en el capítulo anterior, comprende las siguientes etapas: comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

¿Por qué se utiliza el modelo de procesos CRISP-DM y no otro? Este modelo garantiza seguir pasos ordenados para hacer tanto una toma como un análisis de los datos suficiente. Se ha escogido este modelo por su simplicidad explicativa que veremos a continuación.

A continuación, se muestra un pequeño esquema como base de este capítulo:



3.1 Librerías y conexiones con lenguajes de programación

Las librerías computacionales proporcionadas por Twitter para el contacto con los desarrolladores que interactúan con la red para sus fines personales y/o empresariales, están disponibles gratuitamente para todos, aunque en algunos casos con mayores privilegios. Estas bibliotecas de trabajo, aunque no necesariamente todas fueron creadas o probadas por Twitter, son compatibles con la API estándar de Twitter.

La conexión entre estas APIs de Twitter y cualquier desarrollador, requiere del respeto de varias reglas predefinidas por las políticas de privacidad de la red, constantemente cambiantes, con su última actualización el 25 de mayo de 2018. Entre los lenguajes de programación más destacados se encuentran: C++, PHP, Objective-C, Swifter, Java, Tembo, .Net entre otros, con acceso a las librerías python-twitter, tweepy, TweetPony, twitter-gobject, TwitterSearch, twython, TwitterAPI, twitter-ads, twitter-console, tweetstream, entre otras, mediante los sistemas operativos Apple, Windows y Linux. Toda su documentación se muestra actualizada y a disposición de todos los usuarios en su sitio web de desarrolladores.

3.2 Exploración y comprensión de los datos. Generación del dataset

Se requiere extraer tuits de la red social a una base de datos local, utilizando los servicios de la API de Twitter y un proceso de raspado de datos (web scraping).

El primer paso para la realización de un modelado de tópicos es la exploración del dataset. Aquí, se extraerán los todos los tweets que han sido escritos en la red social y que tengan alguna relación con la universidad.

Cabe mencionar la importancia de generar un dataset completo con un acceso profundo a los tuits relacionados con la Universidad Pontificia de Comillas, puesto que sienta las bases para poder efectuar uniones relacionales y conclusiones finales. Así, la comprensión de los tuits que a continuación se va a explicar con detenimiento se efectuará a través de dos formas distintas: a través de la denominada API de Twitter y a través de un web scraping más superficial gracias a la librería “Intro. to academictwitteR” (r-project.org).

La razón por la cual se efectúa esta generación del dataset a través de dos formas independientes (aunque finalmente se combinen y se filtren tuits únicos) es debido a la política de privacidad de Twitter, en donde se indica que solamente podrán extraerse tuits de un período de treinta días. (developer.twitter.com, "*Developer terms*"). Lo que verdaderamente aporta valor en este proyecto, es el poder analizar cualquier mes venidero usando e incluyendo lo visto aquí. Así, aunque se obtenga un dataset de dos fuentes distintas, y una de ellas sea solamente de un mes, es importante enriquecer el proyecto y se estima de vital importancia efectuar este análisis de reputación de la forma más completa posible.

3.2.1 API de Twitter

Creación de aplicación de la API de Twitter

Como paso inicial se necesita tener acceso a una cuenta en Twitter, con un número de teléfono móvil verificado en la red social y que dicha cuenta no presente ningún tipo de restricción o suspensión. Una vez siendo usuario de la red social se crea una aplicación en la plataforma de desarrolladores, Twitter (2022), la cual generará cuatro claves o códigos de acceso con los cuales se conecta desde R-project para realizar las peticiones de datos. Una vez verificada la información por parte de la red social, se habilita el funcionamiento de las claves. Los cuatro códigos que Twitter proporciona para establecer la conexión son: "consumerKey", "consumerSecret", "accessToken" y "accessSecret".

Conexión entre R-project y la API de Twitter

Una vez instalado el software R-project, en su versión gratuita R-project 3.4.1 con el entorno visual Rstudio 1.1, se procede a cargar la librería "TwitteR 1.1.9" (Gentry, 2019), la cual es descrita en su sitio web como: "TwitterR es un paquete R que proporciona acceso a la API de Twitter. La mayoría de las funciones de la API son compatibles, las llamadas a la API son útiles en el análisis de datos e interacción en tiempo real." Instalado el software R-project y las claves de acceso de la API habilitadas, se realiza la conexión entre el ordenador y la plataforma.

Extracción de los datos

Se realiza la petición de búsqueda de tuits la cual se solicita a través de keywords para el caso en cuestión. Se ejecuta así una consulta a la API. Esta consulta se hace a través del comando "searchTwitter", fijando el idioma y el período de extracción.

Las keywords que se han utilizado para la llamada de la API son:

- Universidad Pontificia de Comillas
- UCOMILLAS
- ICADE
- Universidad de Comillas
- @ucomillas
- @ICADE_Derecho
- #CIDICADE

Así, a lo largo de un mes se ha efectuado este proceso siete veces. Se ha hecho una llamada a la API ese número de veces para tener un dataset representativo. Cabe mencionar que esto se podría efectuar para todos los meses venideros, y no solo de mayo de 2022. Este dataset que se obtiene incorpora variables interesantes, como pueden ser emojis o incluso número de retuits. Hablaremos más adelante sobre la selección de variables que más se adecúen al proyecto. Todos estos tuits se incorporan entonces al data frame.

Se procede a continuación a unir los tuits provenientes de las siete extracciones efectuadas. Así, todos los tuits que han sido extraídos de las diversas llamadas a la API se encuentran en una misma tabla de datos (a través de la función "rbind")

Se obtienen así datos desde el 1 de mayo de 2022 hasta el 31 de mayo de 2022, agrupados en una tabla la cual denominaremos "data_comillas_API.csv"

3.2.2 Web scraping

Puede ocurrir que la plataforma de la que se desea extraer información no disponga de una API de consultas, o bien que imponga limitaciones en cuanto a la cantidad de información que se puede obtener, como es el caso de Twitter. En estos casos, se puede intentar extraer la información directamente de la web, lo que se conoce como web scraping.

Según (Bo Zhao, 2017), el raspado de la web, también conocido como extracción de la web, es una técnica para extraer datos de la World Wide Web (WWW) y guardarlos en un sistema de sistema de archivos o base de datos para su posterior recuperación o análisis. Normalmente, los datos de la web se extraen utilizando el Protocolo de Transferencia de Hipertexto (HTTP) o a través de un navegador web. Esto se lleva a cabo de forma manual por un usuario o automáticamente por un bot o rastreador web. Es una técnica eficaz y poderosa para recopilar grandes volúmenes de datos, generalmente son codificaciones escritas a medida del sitio web que se desea extraer información.

A través de la librería Intro. to academictwitteR (r-project.org). se consigue un dataset completo con las siguientes keywords: "universidad pontificia comillas", "universidad comillas", "icai, comillas", "pontificia comillas", "#ICADE, comillas", "ICADEalumnos", "#ICADE universidad", "#ICADE #comillas".

Se obtienen así datos desde el 1 de enero de 2018 hasta el 31 de mayo de 2022, agrupados en una tabla la cual denominaremos “Tweets_Comillas.csv”

3.3 Preparación de los datos

En esta fase, se procede a efectuar la unión de las tablas extraídas a través de la API de Twitter y el web scraping a través de la función “rbind”. Se debe convertir todo en una tabla de R, que contenga todo el dataset, con duplicados y sin limpiar, simplemente importado.

La base de datos generada en R contiene un total de 57.045 observaciones o tuits con un total de 31 variables. Mencionar también que existe la presencia de tuits repetidos, por lo que, en la depuración de los datos, analizaremos como eliminarlos.

Se deben observar ahora las variables que se encuentran en la base de datos completa. Debemos determinar las variables óptimas para el análisis final de la universidad. Es por ello por lo que, de las 32 variables originales, solamente 6 son necesarias y de vital importancia. Para poder efectuar la selección de las mismas, se utiliza la función "dyplr".

Estas variables para el estudio inicial son:

- **Text:** Este es el tuit enviado por el usuario

- **Created_at:** variable sobra la fecha de publicación del tuit.

- **Like_count:** número de favoritos alcanzados por el tuit. Muy importante para una nueva variable que se incorporará a continuación denominada "Interacciones".

- **Retweet_count:** número de retuits alcanzados por el tuit. Igual que para el número de favoritos, esta variable es muy importante para una nueva variable que se incorporará a continuación denominada "Interacciones".

- **Tweet_id:** variable para poder identificar el tuit.

- **User_username:** nombre de usuario que ha escrito el tuit.

3.4 Análisis exploratorio inicial

Con la base de datos completa, se realiza una inspección inicial donde se explora su dimensión y todos los detalles posibles.

En contenido de los documentos resalta la aparición de conectores (el, la, que, para, del.…), signos de puntuación, direcciones URL, caracteres especiales, nombres de usuario, etiquetas de grupos, números y códigos que representan símbolos emoji.

Vemos que se incluyen enlaces a páginas web, retuits o tuits duplicados, por lo que hay que prestar extremada atención para eliminar este "ruido" que dificulta nuestro análisis. Cabe mencionar que todos los procesos de limpieza de tuits que se van a efectuar no son fiables al 100%, por lo que se tendrá que efectuar varios para asegurar una óptima depuración y filtrado

En esta fase del proceso, es importante efectuar una buena selección de variables. Como ya hemos mencionado anteriormente, esas seis mencionadas serán las elegidas para efectuar el análisis inicial y la depuración de los datos.

Es interesante añadir una nueva variable a nuestro análisis. Es por ello por lo que se crea la variable "Interacción", que servirá para poder saber cuál es el impacto que un tuit ha generado. Se construye a través del sumatorio de likes y retuits que posee un tuit. Es necesaria para la creación y generación de tablas de frecuencias y para filtrar las cuentas de usuarios que tengan mayor interacción, que veremos más adelante.

En esta fase del proceso, los términos más repetidos van a ser palabras de las cuales no podemos obtener mucha información. Es por ello por lo que en la próxima fase depuraremos los datos para facilitar el análisis de los mismos.

Para comprobar que efectivamente, se debe efectuar un depurado de la base de datos, se ha procedido a crear una nube de palabras con "stopwords". No se ha incluido en el trabajo debido a la escasa aportación que genera, pero su generación sí permite comprobar la necesidad de efectuar una limpieza. Se ha efectuado con anterioridad una *tokenización* para poder efectuar la nube.

Gracias a la creación de esta nube de palabras mencionada, además de la generación de una columna para comprobar las frecuencias de los términos en la base de datos, se ha podido observar que existe un número que, aunque no significativo, es necesario separar, de tuits o términos en otros idiomas que no sea el castellano. En este análisis exploratorio inicial es fundamental poder localizar estas particularidades para después efectuar un correcto análisis de reputación.

Se ha determinado entonces cuál es el segundo idioma más utilizado para la generación de tuits por parte de los usuarios, en relación con la Universidad Pontificia de Comillas, siendo este el inglés seguido del catalán.

Estos términos, aunque no vayan a ser considerados para la realización del análisis de sentimientos ni del modelado de tópicos, son objeto de mención puesto que forman parte de la base de datos inicial. Estos no van a ser tomados en consideración debido a su ínfimo efecto en el análisis de la reputación de la universidad.

3.5 Depuración de datos

Algunas variables contenidas en la base de datos, que especifican, por ejemplo: el dispositivo electrónico, el navegador o aplicación del cual fueron enviados los mensajes no son de importancia para este estudio. Se realiza el descarte de estas variables y se enfoca la limpieza en la variable “text”, “fecha” y “nombre de usuario”, las cuales representan el contenido textual de interés. Esto ya se ha efectuado con anterioridad, pero se menciona como comprobación y corrección.

En esta sección se procede con la depuración de los datos a través del proceso de tokenización (función “gsub”), filtrado de stopwords (diccionario de palabras en español) y niveles de embudo para reducir palabras a su raíz (stemming). Este proceso permite reducir la dimensionalidad de la base de datos y concentrar los datos textuales que poseen el contenido de interés. La característica principal de la depuración o limpieza de datos, es excluir información que no tiene valor para el análisis de la reputación de la Universidad Pontificia de Comillas.

La primera parte de esta depuración de datos se da a través de la *tokenización*. Es aquí donde se separan los textos de los tuits en términos individuales denominados tokens. Cabe mencionar que, en este momento, se efectúa una conversión de la base de datos, que pasa a encontrarse solamente en minúsculas.

Es ahora el momento en donde se efectúa una eliminación de los tuits repetidos que han sido originados debido a la utilización de dos bases de datos distintas provenientes de distintas fuentes.

Se efectúa entonces una primera limpieza, denominada limpieza inicial, donde se descartan términos que no interesan para la realización del análisis. A través de la función "gsub", se eliminan retuits, direcciones web, signos de puntuación, links, números y se limpian los denominados saltos de línea y tabulaciones.

Ahora se procede a realizar una segunda depuración, esta vez con apoyo de los paquetes de minería de texto 'tm', "tidytex" y "tibble". Se eliminan los tuits repetidos y se construye una matriz de frecuencia.

Esto se realiza a través de la función stopwords del paquete "tm". Aquí se incluyen una serie de palabras establecidas por un lingüista profesional en la lengua española que determina qué palabras no son relevantes para la comprensión y análisis de la lengua. Así, se determina qué palabras son depuradas y cuáles no. Esta librería se denomina "spanish", ya que estamos efectuando el análisis en español.

Para perfeccionar aún más la limpieza, se efectúan otras técnicas de depurado, ya que la eliminación de datos que no son relevantes siempre se puede mejorar. Es por eso por lo que se utiliza el paquete "str_replace_all". La razón de la utilización del mismo es la apreciación en la limpieza anterior de nomenclaturas griegas, saltos de línea exagerados e incluso apariciones de puntuaciones.

Además, se efectúa la aplicación de otro paquete de "tm" para efectuar una limpieza aún más profunda. Se efectúa también a través de la función "stopwords", pero esta vez utilizando la librería "nltk". La característica que diferencia a esta librería de la anteriormente mencionada (librería "spanish") es que en esta se incluye un léxico

proveniente de España, que no contiene léxico latinoamericano. Adicionalmente se trabaja con la codificación en UTF-8.

Una vez se han determinado los paquetes correspondientes de las stopwords, se procede a filtrar las mismas en la tabla.

Se logra así reducir la dimensionalidad de la base de datos de 57.045 tuits a 16.809. Una vez realizada la limpieza, los tuits que contienen imágenes y enlaces y no están acompañados de ninguna cadena de texto, quedan como casillas vacías que posteriormente son eliminadas.

Por último, una vez que la base de datos ha sido depurada, se determinan las frecuencias de todos los términos. Así, debemos determinar que, para facilitar el análisis que se efectúa, se ha determinado que los 5000 términos que tengan más frecuencia son los que se utilizarán para continuar con el modelado, que veremos a continuación.

A partir de este conocimiento se crea el corpus textual (función Corpus), posteriormente utilizando la matriz de documento término (TermDocumentMatrix), se verifica la cantidad de términos que posee el corpus textual, la cantidad de documentos y otros detalles adicionales.

3.6 Modelado: Clustering

Después de crear la matriz documento-término para el corpus textual (a través de la función “Corpus”) y seleccionar las palabras con mayor aporte a la identificación de la temática, se escoge el método de aprendizaje no supervisado de partición (k-means). Para la matriz documento-término se escoge la mejor técnica de análisis de conglomerados con el criterio del método Elbow y el coeficiente de silueta. El análisis permite crear grupos con los términos contenidos en los documentos que contienen temas similares y sobre los que se hace un posterior análisis para identificar la temática más destacada de los mismos. Se realiza el agrupamiento con las funciones “Ckmeans.1d.dp” y “fviz-nbclust”. El resultado permite identificar los conglomerados, términos y temáticas que caracterizan el contenido más relevante de los documentos en estudio.

3.6.1 Validación del análisis de conglomerados

Para encontrar el número de grupos aceptable que componen las temáticas relevantes de los usuarios de la Universidad de Comillas en Twitter, previamente filtrados, existen varios métodos, para efectos de este proyecto se usa el método de elbow y el método de silueta.

Método del codo (Elbow method)

Ambos métodos apoyan la búsqueda del valor de (k) para tomar la mejor decisión. A continuación, se muestra la ilustración del Método Elbow, que recomienda usar un valor de k=2, se observa en el gráfico que a partir de ahí sus puntos de inflexión parecen estabilizarse.

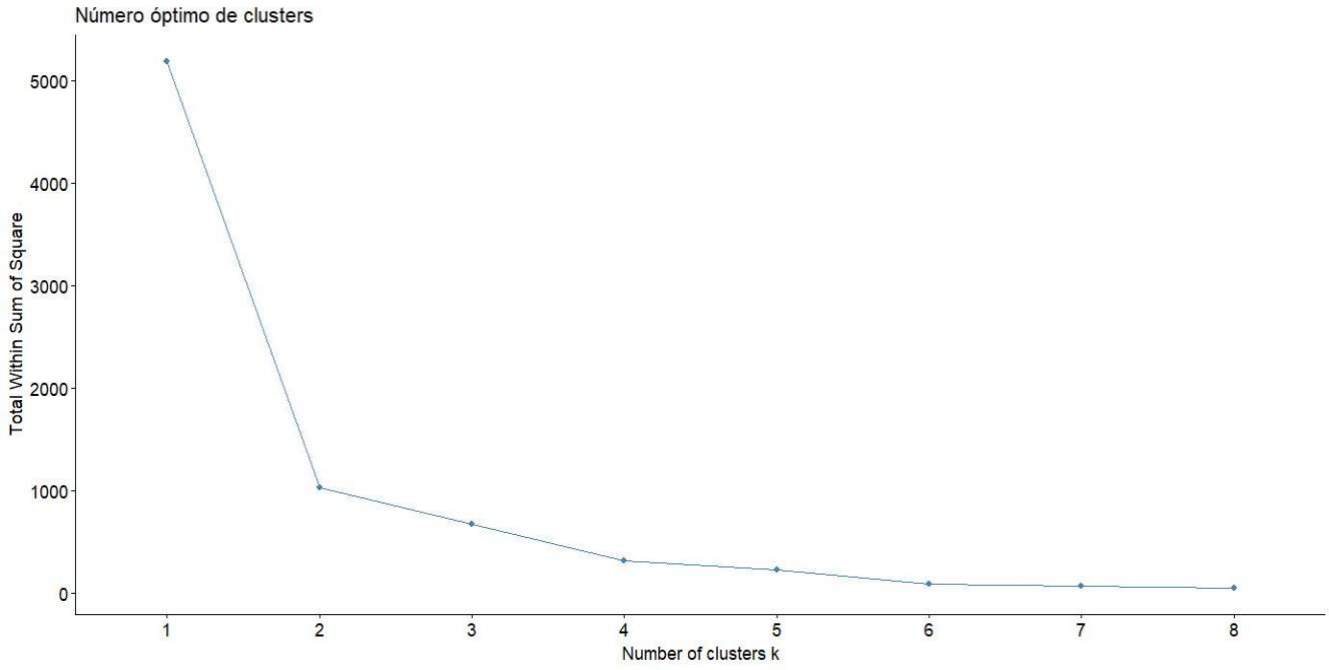


Ilustración 1: Gráfico con puntos de inflexión representando el método Elbow. Fuente: Elaboración propia. Datos obtenidos de Twitter.com

Método de Silueta

Se apoya la verificación realizando una evaluación con el método de silueta, que persigue encontrar la cantidad de grupos más adecuada. A continuación, se muestra la ilustración para k = 2 y k = 3, donde se verifica mejor resultado para k = 2.

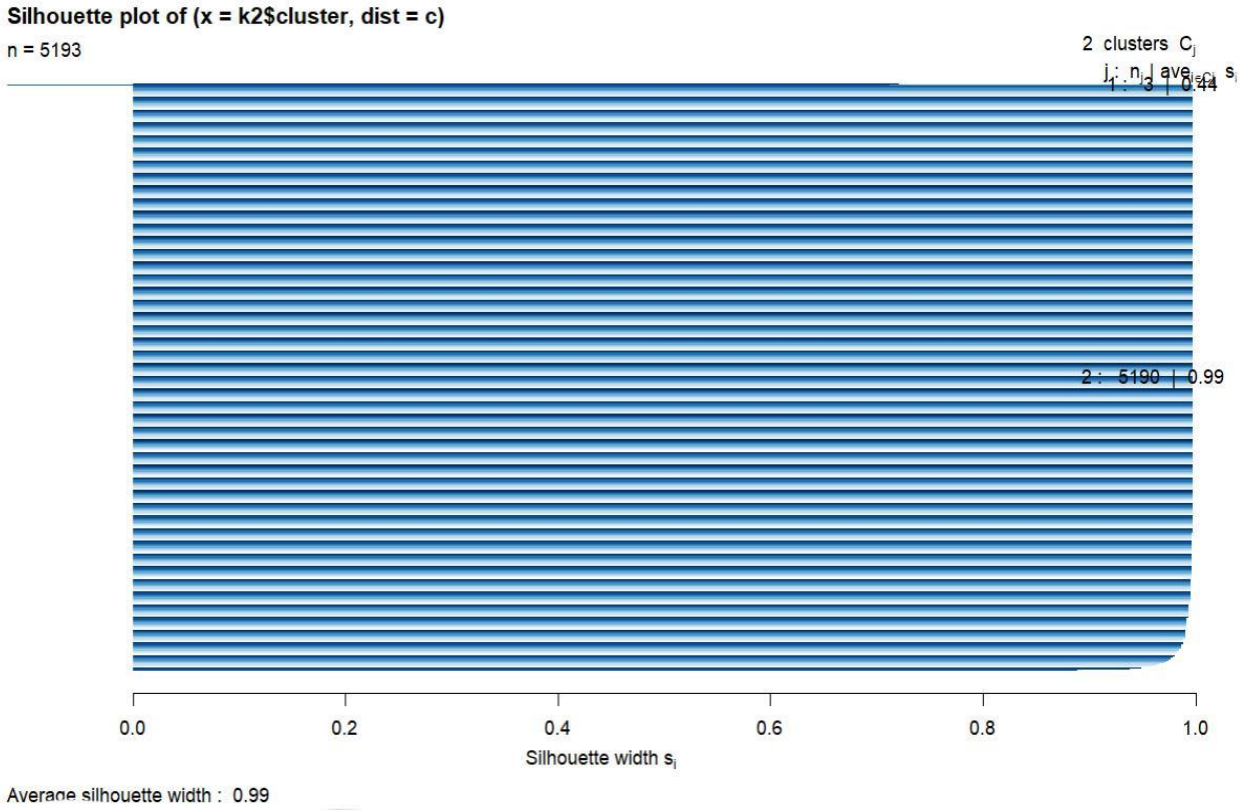


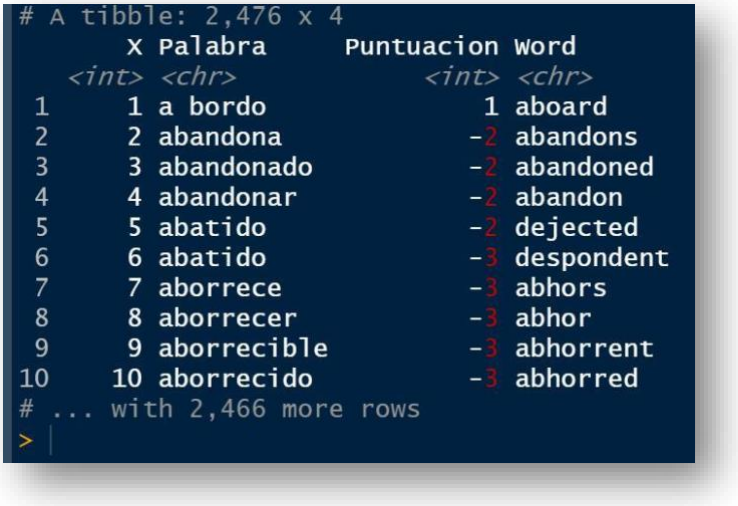
Ilustración II: Gráfico representando el método de Silueta. Fuente: Elaboración propia. Datos obtenidos de Twitter.com

Se utiliza entonces para el análisis de tópicos de la universidad, un análisis de conglomerados con un algortimo K-means, k=2. Por lo tanto, como veremos en el capítulo IV, existirán 2 tópicos perfectamente diferenciados sobre la universidad.

3.7 Análisis de sentimientos

En esta fase final incluida en la metodología, la base de datos debe estar completamente depurada, sin ningún tipo de "stopwords", como se ha efectuado anteriormente en la fase de depuración de los datos. Así, se puede realizar un análisis de sentimiento en base a los datos textuales depurados y se procede a identificar cuentas más relevantes, siguiendo los criterios del trabajo para la minería de datos textuales

Para realizar el análisis de sentimiento, se utiliza el léxico AFINN traducido al español. El léxico originalmente contiene 2476 términos raíz, los cuales se encuentran ponderados desde [-4 a 4], tanto para determinar el grado negativo y positivo, quedando 0 como neutral. Las funciones computacionales para realizar los filtrados están contenidas en los paquetes "tidyverse", "tidytext" y "tm". A continuación, se muestra una tabla con diez términos del léxico y su puntuación como ejemplo para configurar una idea más amplia de la generación de un análisis de sentimientos.



```
# A tibble: 2,476 x 4
       X Palabra     Puntuacion Word
   <int> <chr>            <int> <chr>
 1     1 a bordo              1 aboard
 2     2 abandona            -2 abandons
 3     3 abandonado          -2 abandoned
 4     4 abandonar           -2 abandon
 5     5 abatido             -2 dejected
 6     6 abatido             -3 despondent
 7     7 aborrece            -3 abhors
 8     8 aborrecer           -3 abhor
 9     9 aborrecible         -3 abhorrent
10    10 aborrecido          -3 abhorred
# ... with 2,466 more rows
>
```

Tabla I: se puede observar términos del léxico AFINN y su puntuación de -4 a 4.

Fuente: Elaboración propia. Datos obtenidos de Twitter.com

Una vez realizado el filtrado de términos con el léxico AFINN se muestra la dimensión de los sentimientos positivos y negativos contenidos en los 16.809 tuits ya limpiados. La cantidad de términos que coinciden con los incluidos en la lista AFINN es de 737. Existen algunos términos ponderados por el léxico AFINN que no encajan en el contexto, no se

puede garantizar que el 100% de los términos sea filtrado correctamente, sin embargo, es un léxico que contiene un universo de términos raíz que es suficiente para encontrar patrones significativos y relevantes.

	Tipo	Palabra	n
1	Negativa	no	5376
2	Positiva	si	3050
3	Negativa	problemas	1853
4	Positiva	mayor	1018
5	Positiva	gracias	928
6	Positiva	apoyo	846
7	Positiva	premio	605
8	Positiva	enhorabuena	534
9	Negativa	pobreza	511
10	Negativa	guerra	498
11	Positiva	mejor	488
12	Positiva	interesante	455
13	Positiva	oportunidad	422
14	Positiva	valor	414
15	Positiva	oferta	394
16	Positiva	premios	385
17	Negativa	puesto	375
18	Negativa	solo	355
19	Negativa	enfermedad	324
20	Negativa	residuos	324
66	Positiva	claro	105
67	Positiva	suerte	104
68	Positiva	cierto	100
69	Negativa	cobrar	98
70	Positiva	excelencia	96
71	Positiva	compartir	95
72	Negativa	peor	92
73	Positiva	bienvenida	92
74	Positiva	celebrado	91
75	Positiva	libertad	91
76	Negativa	vano	83
77	Positiva	responsable	83
78	Negativa	amenaza	81
79	Positiva	fe	79
80	Negativa	hambre	78
81	Positiva	confianza	78
82	Negativa	cancela	76
83	Positiva	sentido	76
84	Positiva	mejorar	75
85	Negativa	abusos	74
198	Negativa	advierte	24
199	Negativa	confundir	24
200	Positiva	compartido	24
201	Positiva	enorme	24
202	Positiva	facilitar	24
203	Positiva	popular	24
204	Positiva	saliente	24
205	Positiva	vision	24
206	Negativa	expulsado	23
207	Positiva	esperando	23
208	Positiva	fuerte	23
209	Positiva	salvar	23
210	Positiva	agradecido	22
211	Positiva	dulce	22
212	Positiva	estimado	22
213	Positiva	favorito	22
214	Positiva	justa	22
215	Positiva	promesas	22
216	Positiva	promovido	22
217	Negativa	peligro	21

Tabla II: Términos de la base de datos y si consideración negativa o positiva en base al léxico AFINN.

Fuente: Elaboración propia. Datos obtenidos de Twitter.com

A partir de aquí, se comenzará a graficar los diversos resultados para poder analizarlos y debatirlos. El análisis de sentimientos permitirá ver si los usuarios de Twitter que hablan o comentan sobre la universidad tienen opiniones positivas o negativas sobre la misma.

3.8 Determinación de cuentas más relevantes

Para la determinación de las cuentas más relevantes se debe utilizar la variable "Interacción" mencionada anteriormente. A continuación, se efectúa una segmentación por años (2018, 2019, 2021, 2022). A través de la función "ggplot" se generan las determinadas cuentas relevantes y finalmente se unen para generar el total.

CAPÍTULO IV: RESULTADOS Y DISCUSIONES

4.1 Resultados de la depuración de datos

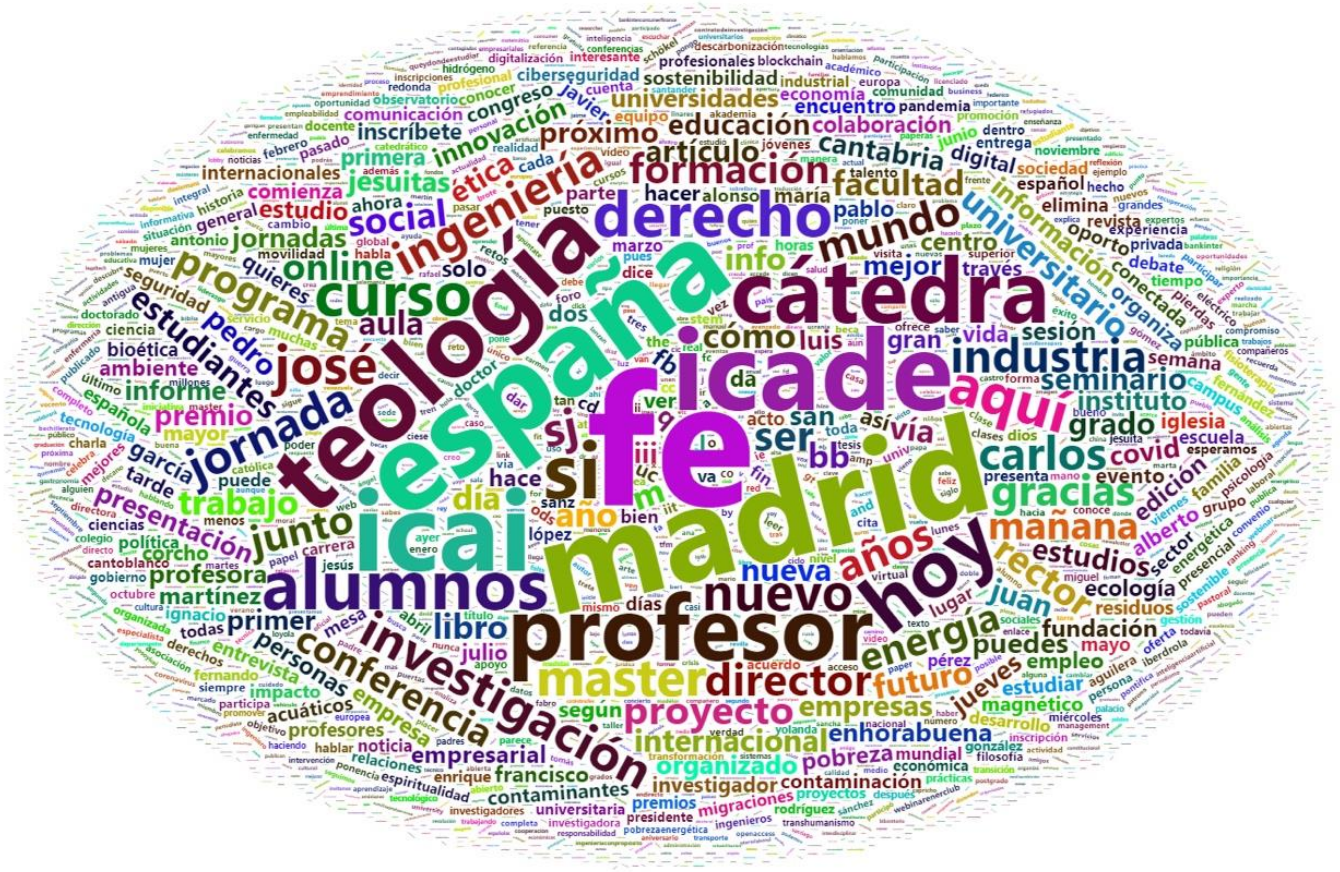


Ilustración III: Nube de palabras de términos más frecuentes sobre la Universidad Pontificia de Comillas.

Fuente: Elaboración propia. Datos obtenidos de Twitter.com

Esta nube de palabras representa todos los términos que se han podido localizar en la base de datos, cuando ésta ha sido depurada y limpiada. Se ha decidido efectuar un análisis a través de una nube de palabras para facilitar la comprensión y utilización de la información que se ha extraído. Como aclaración, cabe mencionar que no se han tenido en cuenta las "keywords" que se han utilizado para extraer los datos de los tuits sobre la universidad, puesto que en ese caso no se podría ver con facilidad los términos que sí se utilizan frecuentemente. Si se utilizaran "keywords" los términos representativos que se pueden observar quedarían ocultos tras esos términos, puesto que siempre van a ser los que más se repiten. No se tienen en consideración entonces palabras como "comillas" "universidad", "pontificia" entre otras.

Las palabras que se encuentran más cercanas o son de un tamaño mayor, son las que más se utilizan en los tuits que se han extraído. Es por esto por lo que palabras como “fe”, “Madrid”, “teología”, “España”, “derecho”, “ingeniería” o “formación”, entre otras resaltan por su tamaño.

Sin tener ninguna información sobre la Universidad Pontificia de Comillas, se puede empezar a tener una idea de cuáles son los valores que transmite. Es una universidad jesuita, por lo que términos como “fe” y “teología” tienen mucho peso dentro de la institución.

Palabras como “ingeniería”, “masters”, “formación” transmiten información sobre las ramas académicas que maneja la institución. Vemos que la Universidad Pontificia de Comillas incluye la Facultad de Derecho (ICADE) y la Escuela Técnica Superior de Ingeniería (ICAI), además de la posibilidad de efectuar másteres en la misma.

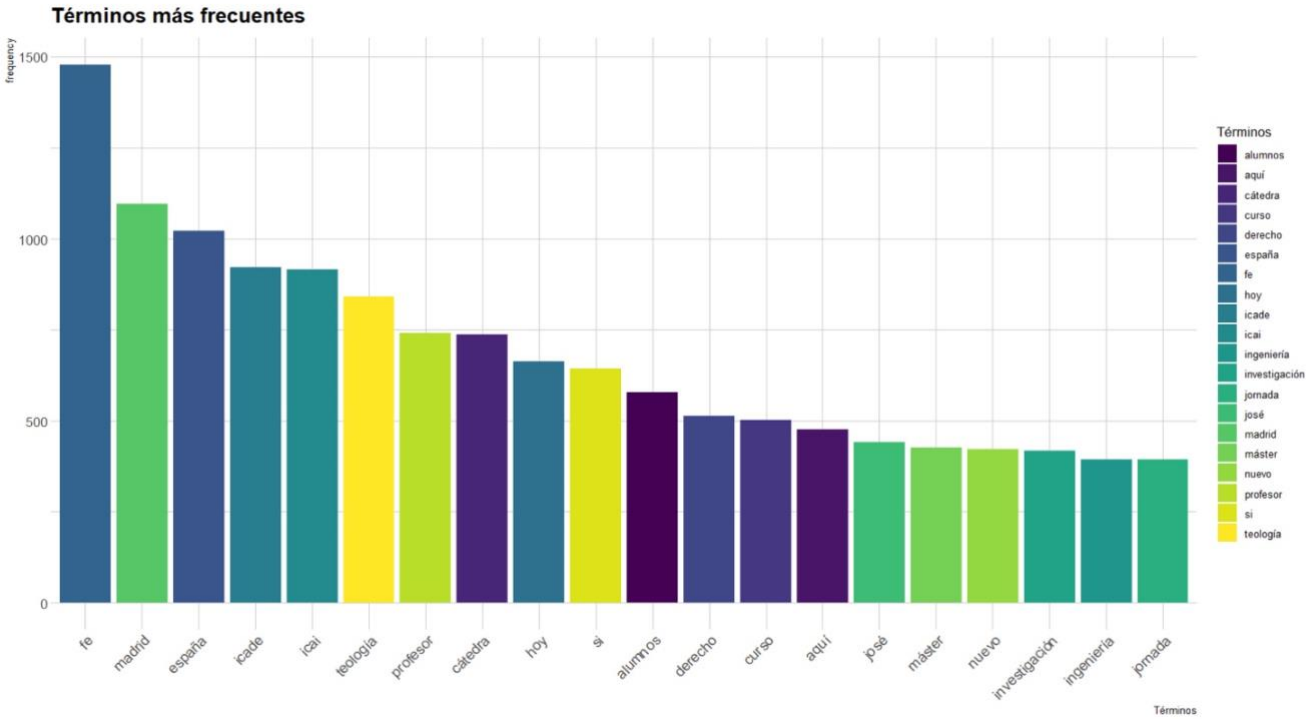


Ilustración IV: gráfico que muestra los términos más frecuentes depurados sobre la Universidad Pontificia de Comillas.

Fuente: Elaboración propia. Datos obtenidos de Twitter.com

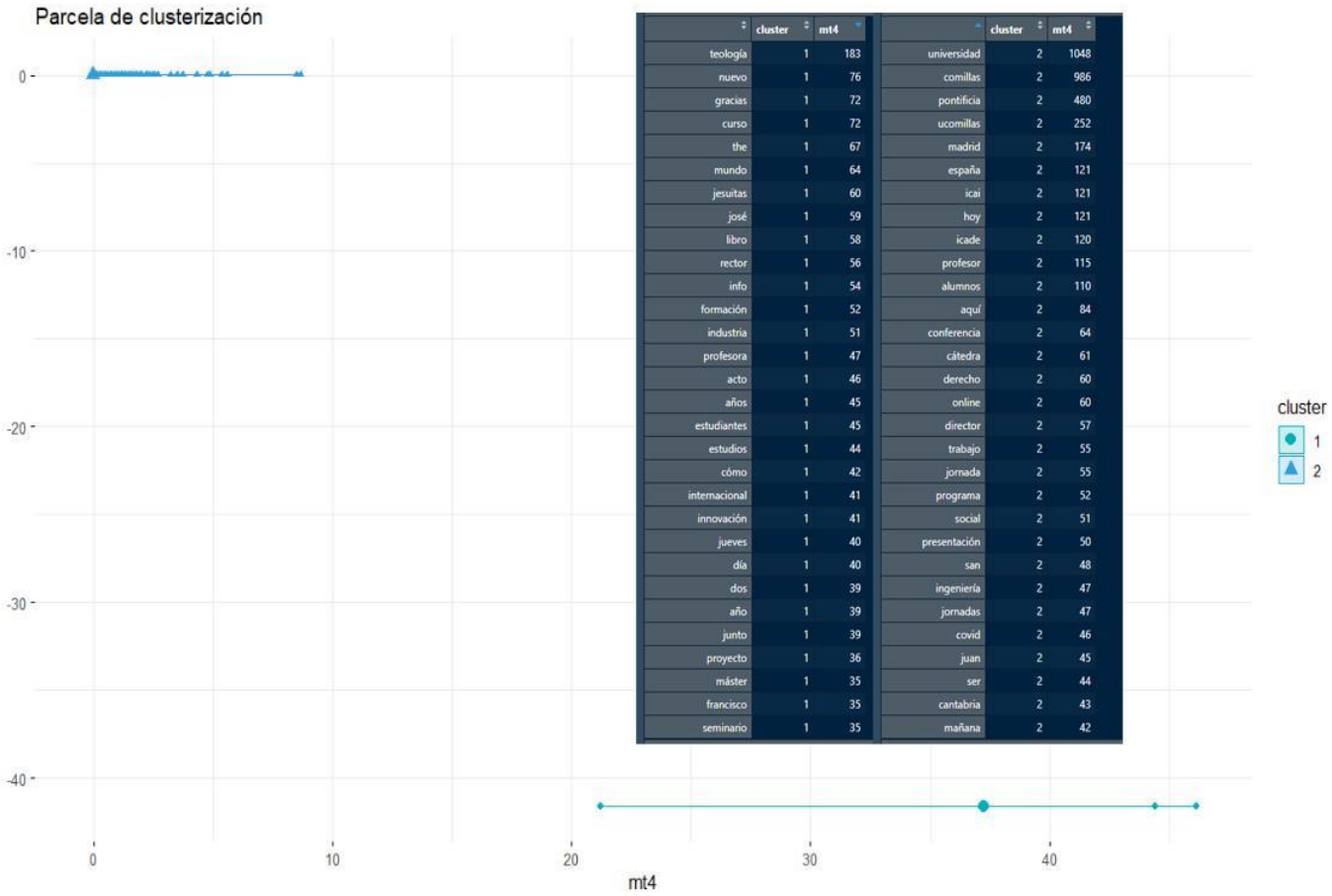
Continuando con lo visto en la nube de palabras (*Ilustración III*). En la *Ilustración IV* podemos observar un gráfico de barras con todos los términos más frecuentes que han sigo generados por usuarios de Twitter en relación con la universidad desde el 1 de enero de 2018 hasta el 31 de mayo de 2022.

Esta es una forma menos visual de mostrar los datos, pero más completa en términos de información. Como se puede ver, el eje vertical muestra la frecuencia de los términos, mientras que en el eje horizontal aparecen los 20 términos más frecuentes de la base de datos.

Liderando la tabla tenemos al término “fe”, el cual aparece en la base de datos más de 1400 veces. Es por ello por lo que deducimos que muchos de los tuits relacionados con la institución tienen una connotación religiosa.

Se puede observar que se encuentra la palabra “si” y “hoy” en la tabla de frecuencias. Si quisiéramos retirarlas, podríamos efectuarlo manualmente a través de la función “gsub” en RStudio.

4.2 Resultados del análisis de tópicos. Clustering



	cluster	mt4
teología	1	183
nuevo	1	76
gracias	1	72
curso	1	72
the	1	67
mundo	1	64
jesuitas	1	60
josé	1	59
libro	1	58
rector	1	56
info	1	54
formación	1	52
industria	1	51
profesora	1	47
acto	1	46
años	1	45
estudiantes	1	45
estudios	1	44
cómo	1	42
internacional	1	41
innovación	1	41
jueves	1	40
día	1	40
dos	1	39
año	1	39
junto	1	39
proyecto	1	36
máster	1	35
francisco	1	35
seminario	1	35

	cluster	mt4
universidad	2	1048
comillas	2	986
pontificia	2	480
ucomillas	2	252
madrid	2	174
españa	2	121
icai	2	121
hoy	2	121
icade	2	120
profesor	2	115
alumnos	2	110
aquí	2	84
conferencia	2	64
cátedra	2	61
derecho	2	60
online	2	60
director	2	57
trabajo	2	55
jornada	2	55
programa	2	52
social	2	51
presentación	2	50
san	2	48
ingeniería	2	47
jornadas	2	47
covid	2	46
juan	2	45
ser	2	44
cantabria	2	43
mañana	2	42

Ilustración V y Tabla III

Fuente: Elaboración propia. Datos extraídos de Twittter.com

A través de una clusterización con el algoritmo K-means, (k=2), todos los términos depurados y limpiados de la base de datos se han agrupado en 2 grupos bien diferenciados. La creación de dos clústeres se ha determinado siguiendo las bases de recomendación del Método de Silueta y el Método del Codo.

Vemos en la *Ilustración 5* como los términos se han agrupado en dos espacios del gráfico muy separados entre sí. Esto es debido a que los dos tópicos que han sido generados distan el uno del otro, lo cual es muy positivo para el análisis. Que se haya recomendado la utilización de solamente 2 agrupaciones, es ventajoso, puesto que ayuda a poder diferenciar fácilmente ambas. Si nos encontráramos con 4, 5 o incluso 6 agrupaciones, será mucho más difícil poder diferenciar los mismos.

En la *Tabla III* podemos observar como ejemplo palabras que forman parte de un clúster o del otro. Así, podemos efectuar un análisis de resultados y hablar de los dos tópicos que predominan a la hora de escribir tuits sobre la Universidad Pontificia de Comillas.

El primer grupo que se ha representado (clúster 1 en la *Tabla III*), en el que se recogen palabras como "mundo", "internacional", "proyecto", "seminario" y "acto" entre otras, tiene como tópico principal la creación de eventos o programas dentro y fuera del ámbito de la universidad. En este tópico se incluyen los valores de la institución y la forma en la que la universidad quiere evolucionar.

En el segundo grupo que se ha representado (clúster 2 en la *Tabla III*), en el que se recogen palabras como "profesor", "alumnos", "trabajo", "jornada", "COVID" y "online" entre otras, tiene como tópico principal los temas relacionados con la institución de la universidad en sí. Se incluyen elementos sobre la organización, clases, profesores e incluso problemas que ocurren para la lección de clases. En resumidas cuentas, tópicos relacionados con la propia Universidad Pontificia de Comillas en su sentido más estricto.

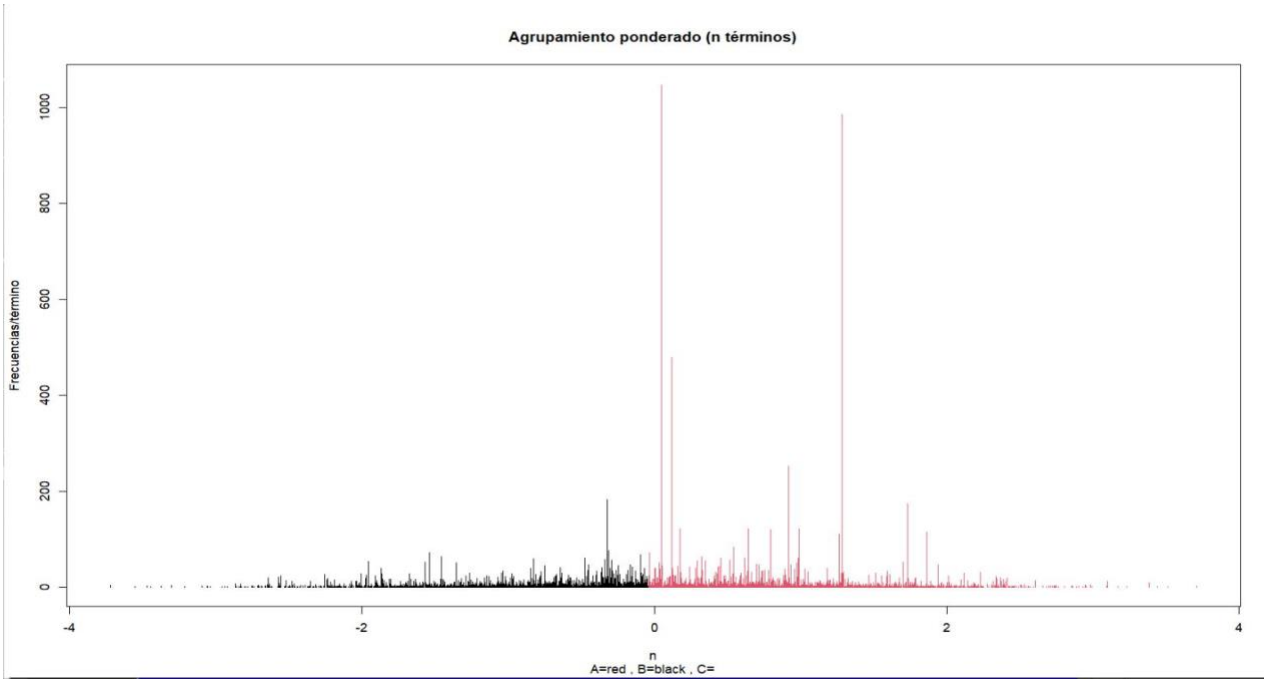


Ilustración VI: Agrupamiento ponderado para la totalidad de términos previamente limpios y ordenados

Fuente: Elaboración propia. Datos extraídos de Twitter.com

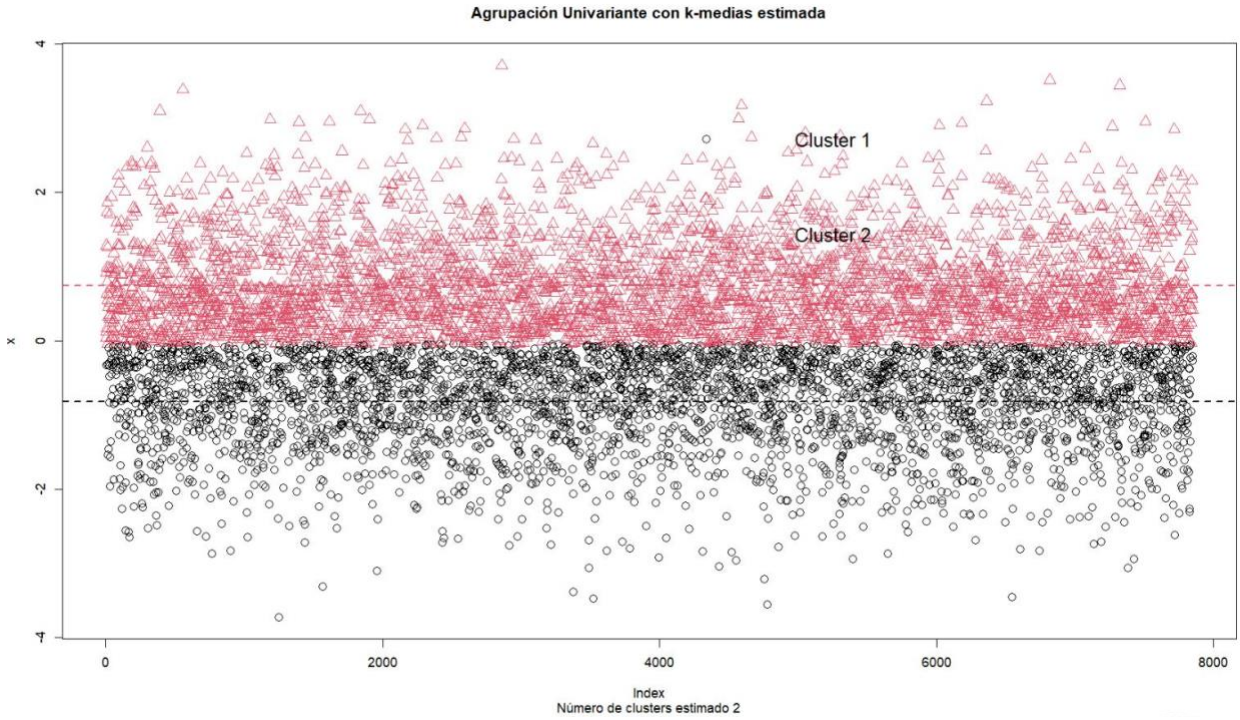


Ilustración VII

Fuente: Elaboración propia. Datos extraídos de Twitter.com

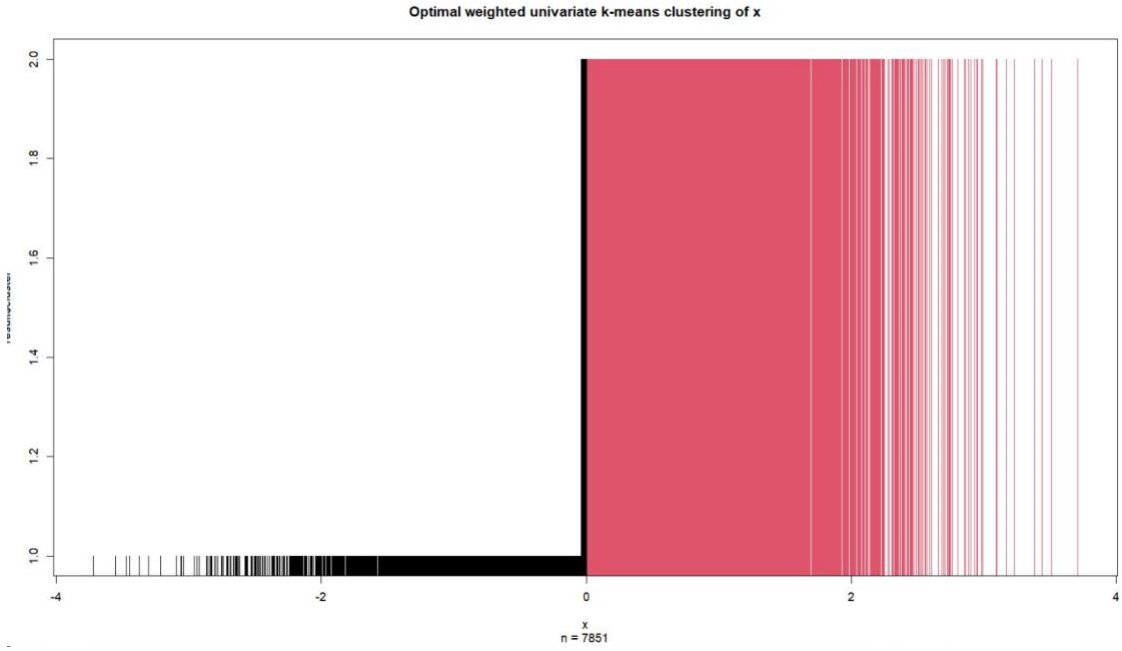


Ilustración VIII.

Fuente: Elaboración propia. Datos extraídos de Twitter.com

En la *Ilustración VI* y la *Ilustración VII*, podemos observar de una forma mucho más clara la distinción de tópicos que hemos mencionado anteriormente, esta vez con las frecuencias de cada término. En la *Ilustración VII* simplemente se completa la información explicada anteriormente.

La *Ilustración VIII*, es de especial importancia porque indica claramente que el clúster donde se encuentra el tópico relacionado con la Universidad Pontificia de Comillas y su organización tiene más peso que el clúster en donde se tratan temas de creación de eventos dentro y fuera del ámbito de la institución.

4.3 Resultado del Análisis de Sentimientos

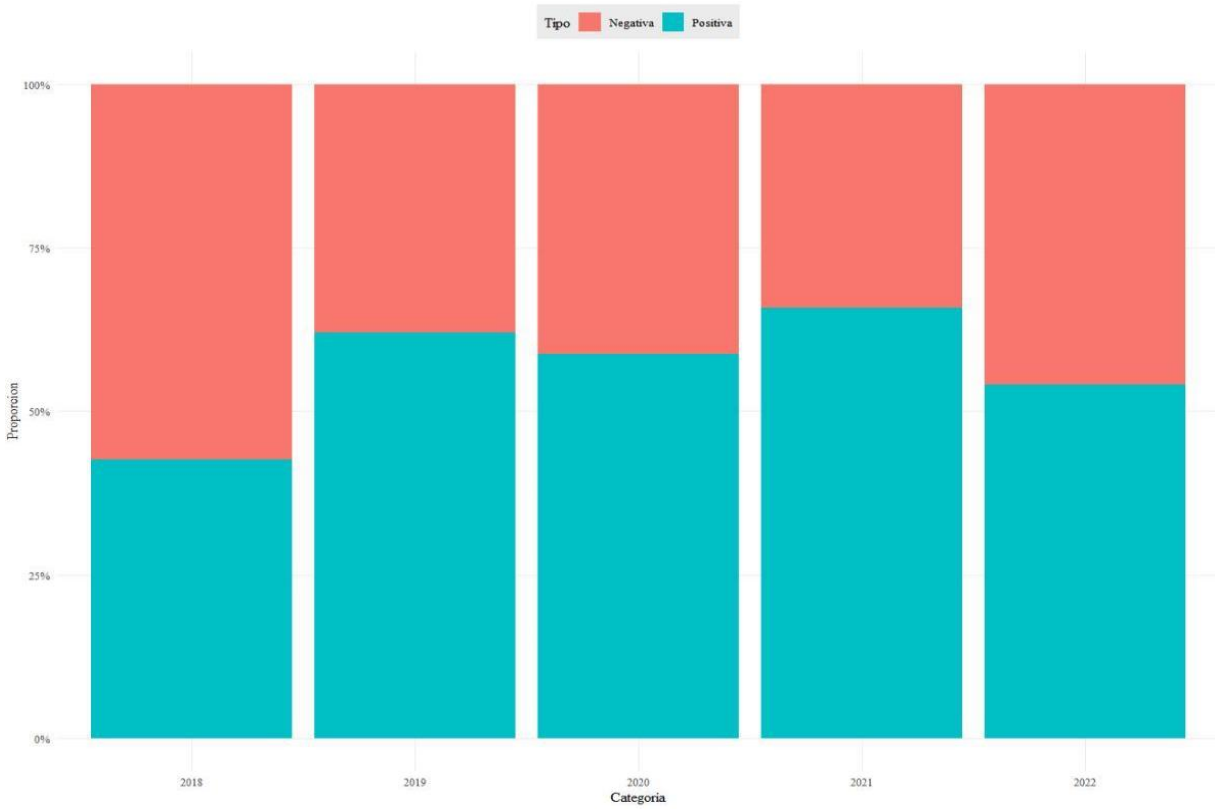


Ilustración IX

Fuente: Elaboración propia. Datos extraídos de Twitter.com

En la *Ilustración IX*, podemos observar una distribución box-plot en donde aparece el porcentaje de términos negativos o positivos que han comentado usuarios en Twitter sobre la Universidad Pontificia de Comillas categorizados por año.

Como podemos observar, en 2018 los comentarios han sido en más de un 50% negativos, mientras que en 2021 es donde los comentarios han sido más positivos. En general, al tratarse la Universidad Pontificia de Comillas en una institución educativa, nunca se podrá apreciar una cantidad superior de comentarios negativos que positivos, como puede ocurrir en el caso de empresas de trabajadores, puesto que nadie se refiere a la misma criticándola.

Aunque los resultados mostrados son bastante dispersos, podemos concluir que existe un ambiente bueno y positivo en relación a las opiniones de la institución.

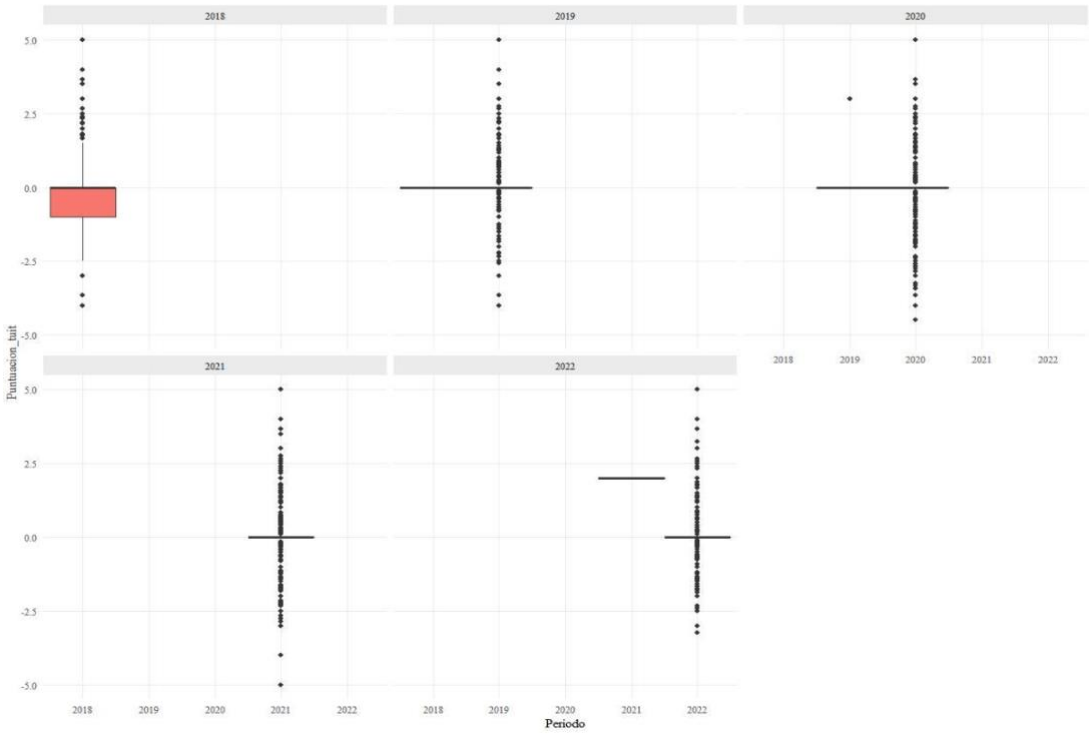


Ilustración X

Fuente: Elaboración propia. Datos extraídos de Twitter.com

En la *Ilustración X* se explora la dispersión con diagramas de caja o bigote a través con apoyo de la librería "ggplot2", sobre la distribución de los términos que expresan algún sentimiento: sobre la línea 0.0, sentimiento en positivo; por debajo de la línea, sentimiento negativo.

Se muestra el diagrama de caja, el cual no se logra divisar porque las puntuaciones están dispersas de -4 a 4 y no parecen concentrarse en algún rango específico, excepto los del año 2018, en donde observamos que las puntuaciones son prácticamente neutras.

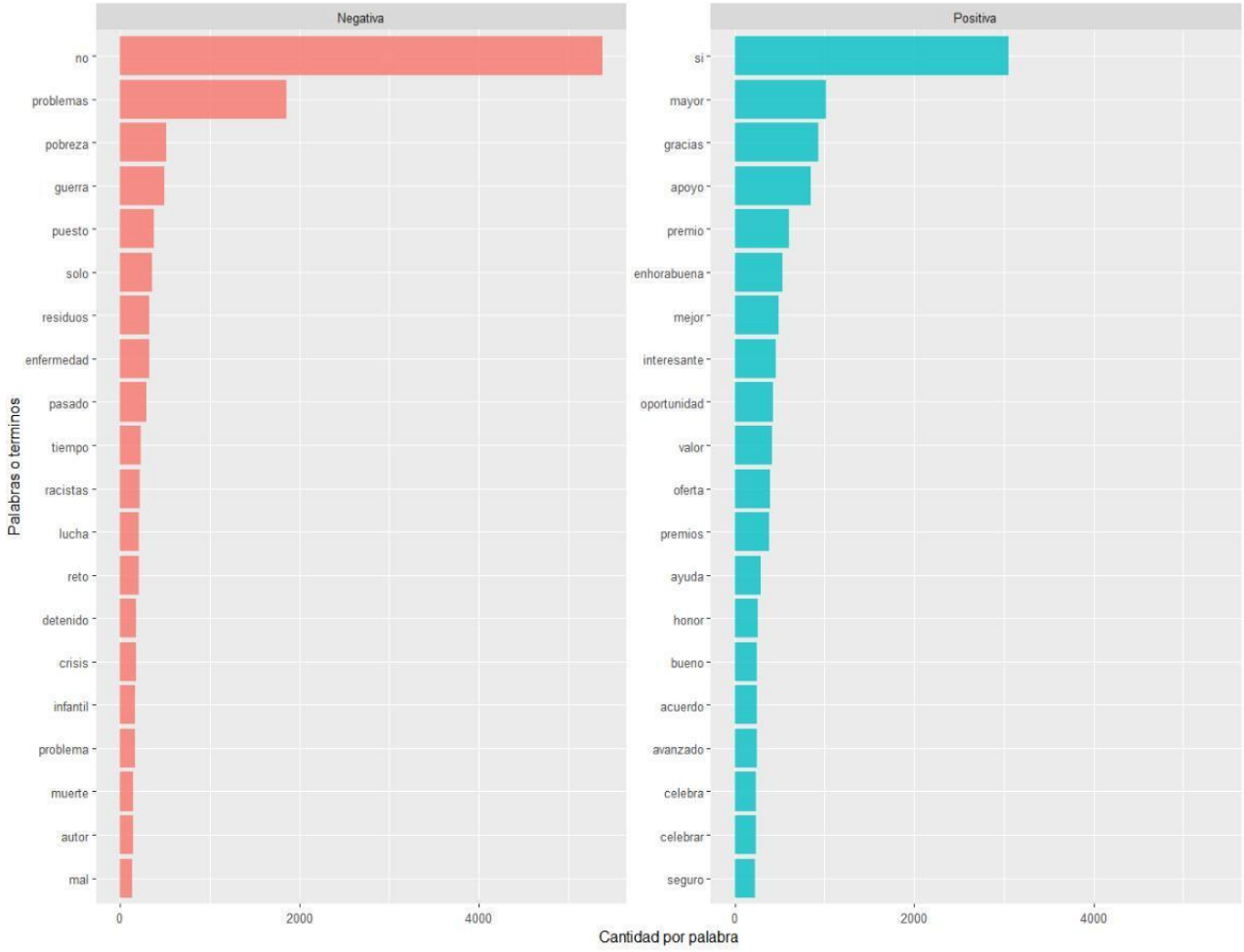


Ilustración XI

Fuente: Elaboración propia. Datos extraídos de Twitter.com

En la *Ilustración XI*, se muestran los términos con el sentimiento más significativo durante todo el período de la muestra: desde el 1 de enero de 2018 hasta el 31 de mayo de 2022.

Podemos observar que los términos positivos que tienen más peso en la muestra tienen que ver directamente con la universidad. Palabras como “gracias”, “enhorabuena”, “valor”, “oportunidad” e “interesante” entre otras, indican que la universidad está efectuando un correcto trabajo tanto en su organización como en su apoyo a alumnos.

Por otro lado, en los términos negativos se habla de:

- “enfermedad”: se ha observado que se hace mucha mención a este término debido a la pandemia COVID-19 que recientemente hemos superado.
- “guerra”: este término hace referencia a la guerra actual entre Ucrania y Rusia.
- “crisis”: se ha observado que la mención a este término tiene que ver con la situación política y económica en la que se encuentra España.
- “pobreza”

  …

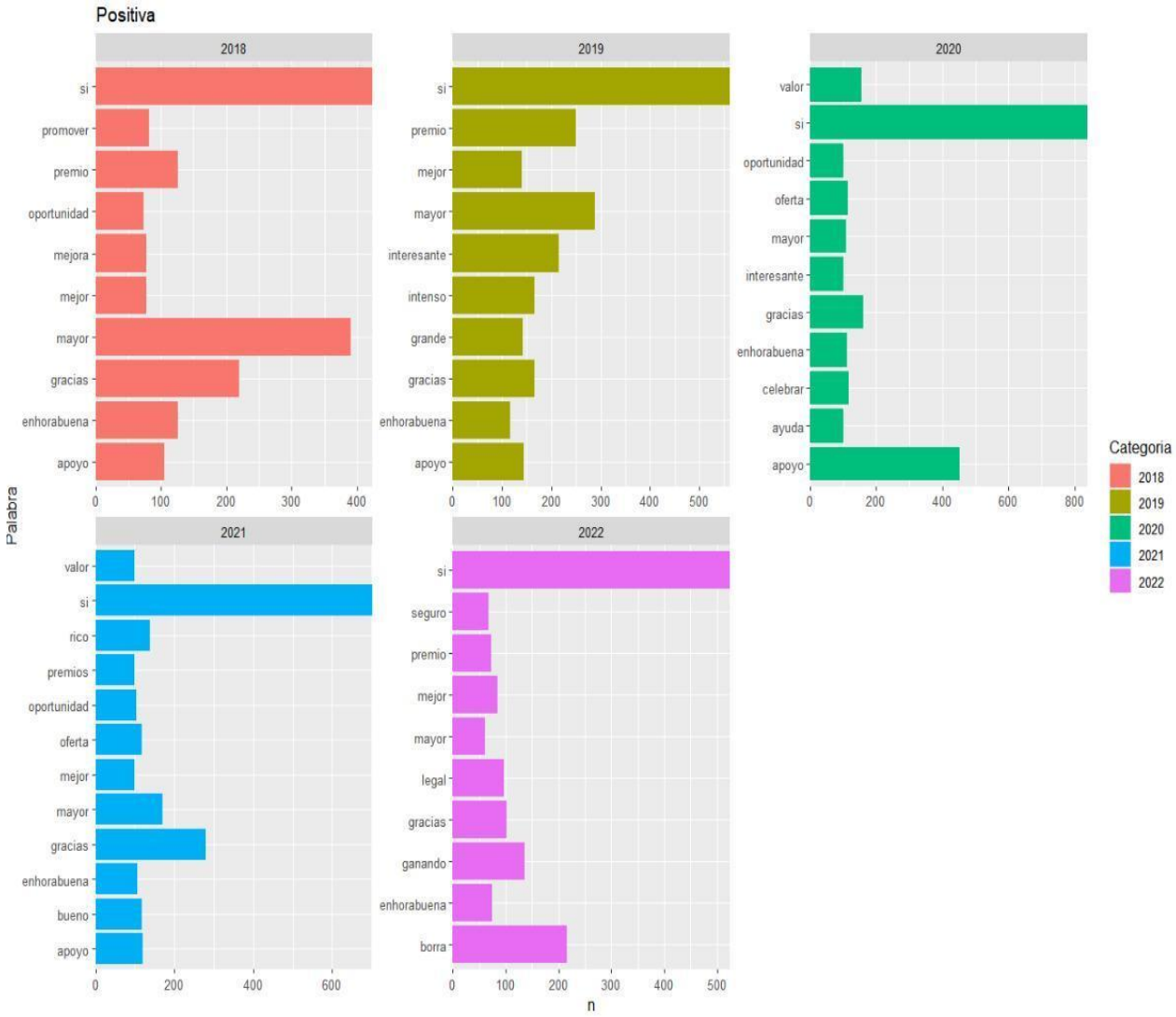


Ilustración XII

Fuente: Elaboración propia. Datos extraídos de Twitter.com

En la *Ilustración XII*, se representan los sentimientos positivos categorizados año a año. Se procede entonces a efectuar un análisis de los resultados en cada anualidad.

En el año 2018, podemos observar términos que se relacionan con agradecimientos y felicitaciones ("gracias", "enhorabuena", "apoyo", entre otros). Por lo que podemos deducir que la Universidad Pontificia de Comillas ha recibido buenos comentarios.

En los años 2019 y 2020, los sentimientos positivos más relevantes siguen la misma línea que en 2018. Esto quiere decir que la universidad es constante en su esfuerzo y dedicación para mantener e incluso mejorar su reputación. Cabe mencionar que en 2020 la frecuencia con la que se mencionan estos términos o sentimientos positivos se reduce, posiblemente debido a la pandemia del COVID-19.

En el año 2021, se pueden observar sentimientos parecidos que los años anteriores.

En el año 2022, aunque incompleto puesto que la toma de datos se ha efectuado hasta el 31 de mayo de 2022, se han incorporado nuevos sentimientos. En primer lugar, se incluyen palabras como "ganando", por lo que se entiende que la universidad está incrementando positivamente alguna función, o mismamente ha ganado algún premio. En segundo lugar, se incluye el término "legal", palabra que no tiene en principio ninguna aportación al análisis, simplemente sirva para indicarnos que la universidad tiene una facultad de derecho.

En un conjunto global y, como ya se ha mencionado anteriormente, la Universidad Pontificia de Comillas tiene un sentimiento positivo fuerte, cosa que se refuerza y mantiene a lo largo de los años.

A continuación, se mostrará una ilustración donde se incorporan los sentimientos negativos categorizados por año.

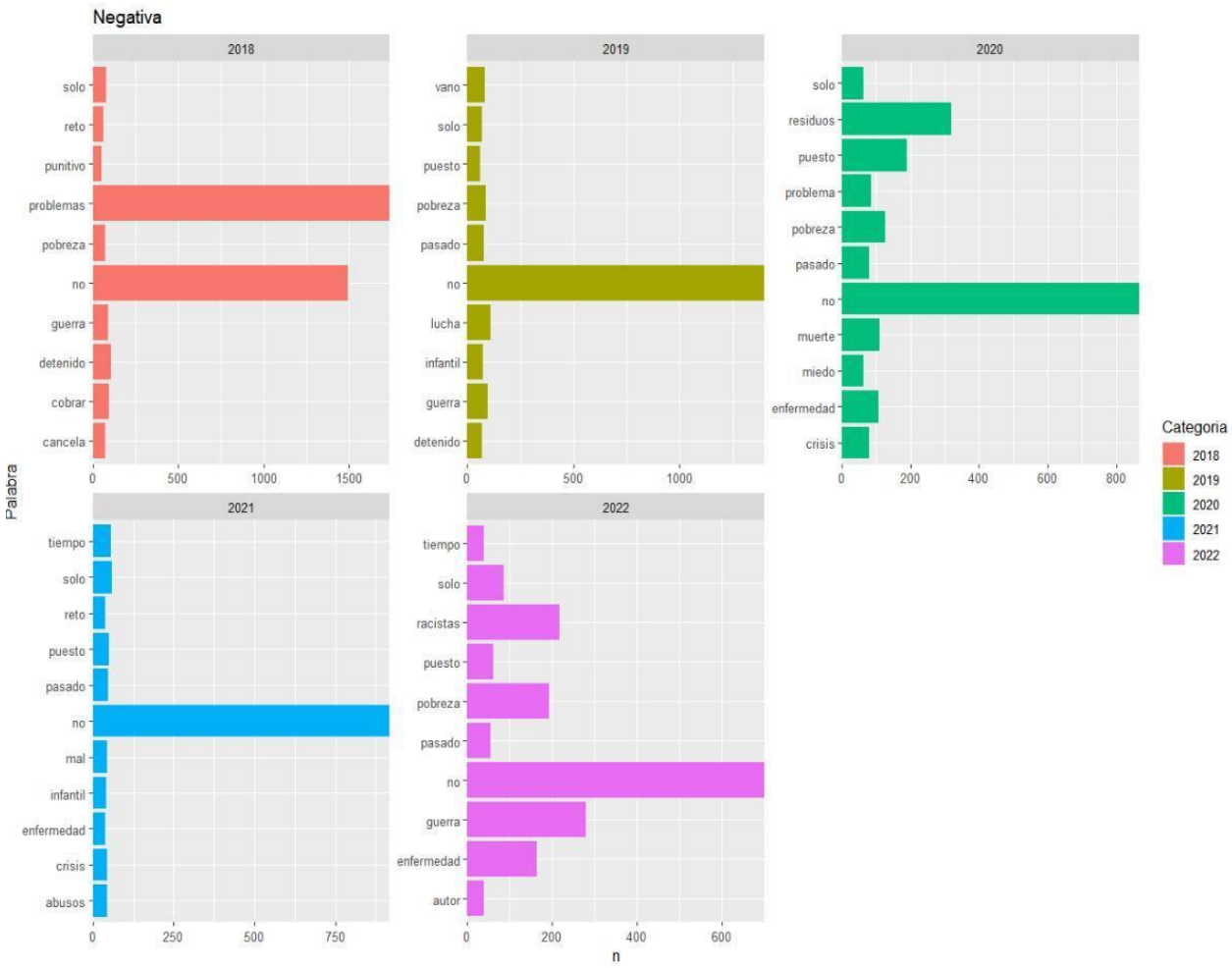


Ilustración XIII

Fuente: Elaboración propia. Datos extraídos de Twitter.com

Se procede a analizar los resultados provenientes de la *Ilustración XIII*.

En el año 2018, puede observarse que, salvo dos términos, los demás sentimientos han sido mencionados de una forma muy reducida. En relación con los términos que sí que tienen una frecuencia muy representativa, debemos centrarnos en la palabra "problemas". La mención reiterada de este término con mucha connotación negativa viene dada por dificultades y desafíos que aparecen a la hora de gestionar una institución académica tan importante como lo es la Universidad Pontificia de Comillas. No debemos centrarnos solamente en este término, pero sí es necesario tenerlo en cuenta.

En el año 2019 podemos observar sentimientos negativos como pueden ser "guerra", "pobreza" o incluso "infantil" entre otros. No debemos confundir estos resultados y determinar que la Universidad Pontificia de Comillas es responsable de una supuesta

“guerra”, pero sí podemos determinar que la institución se ha encargado de sacar a la luz temas conflictivos que ayudan a desarrollar las opiniones de los jóvenes. A su vez, se puede observar que la universidad ayuda a personas en situaciones desfavorecidas (término “infantil”), ya que, al tener una conexión muy directa con la Compañía de Jesús, siempre intentará ayudar al prójimo.

En el año 2020 hay tres términos que llaman la atención. Estos son “enfermedad”, “residuos” y “miedo”. La razón por la que aparecen estos sentimientos es la aparición de la pandemia COVID-19. Numerosos tuits han sido escritos en relación con la universidad, ya que en este año tuvo que cerrar sus puertas y apañárselas para efectuar clases en formato telemático.

En el año 2021, vemos hay una reducción en términos negativos, igual que en 2018 y 2019. Esto se debe a que se retoman las clases y las actividades vuelven parcialmente a la normalidad.

Finalmente, en 2022, aunque solamente se incluyan datos hasta el 31 de mayo, se puede observar un aumento de frecuencias en los términos negativos, aunque nada alarmante en comparación con los términos positivos mencionados en la *Ilustración XII*. La principal razón de este aumento se debe a la guerra entre Ucrania y Rusia que actualmente está teniendo lugar. Así, se puede ver perfectamente en términos como “guerra”, que en este año tiene una frecuencia mucho más alta de lo habitual, “pobreza” y “racistas” entre otros.

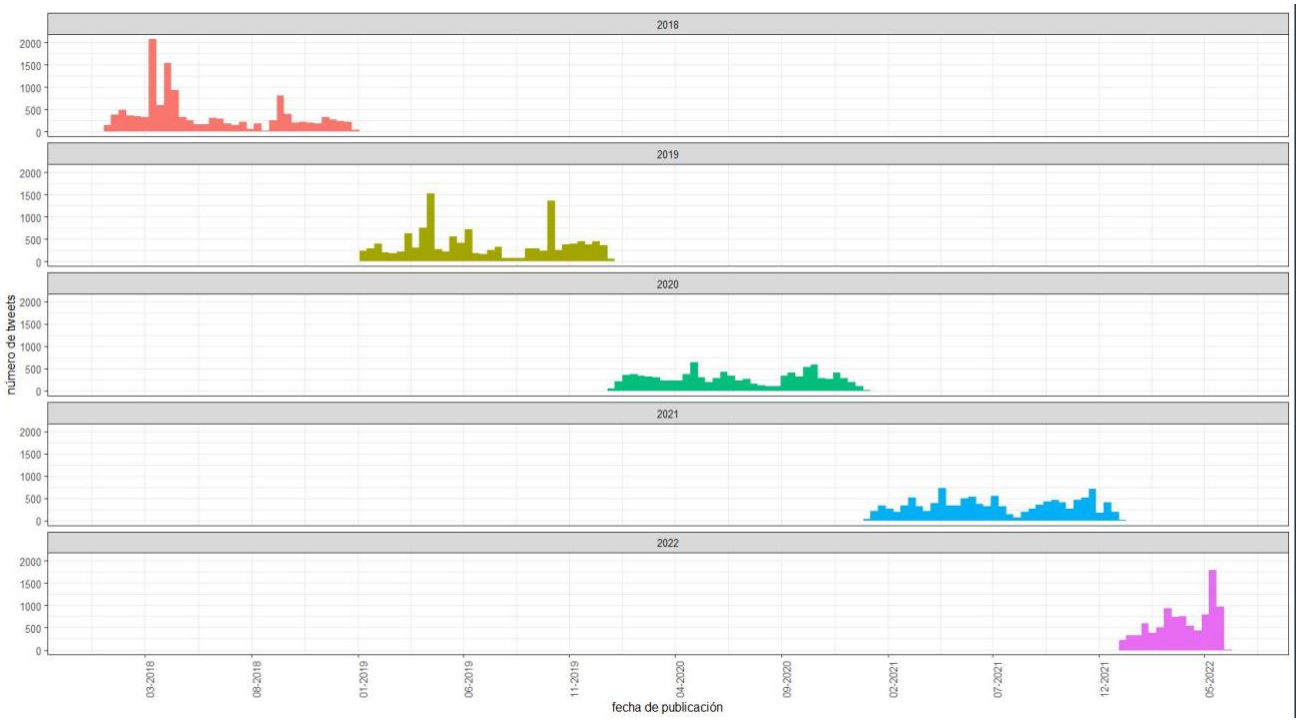


Ilustración XIV

Fuente: Elaboración propia. Datos extraídos de Twitter.com

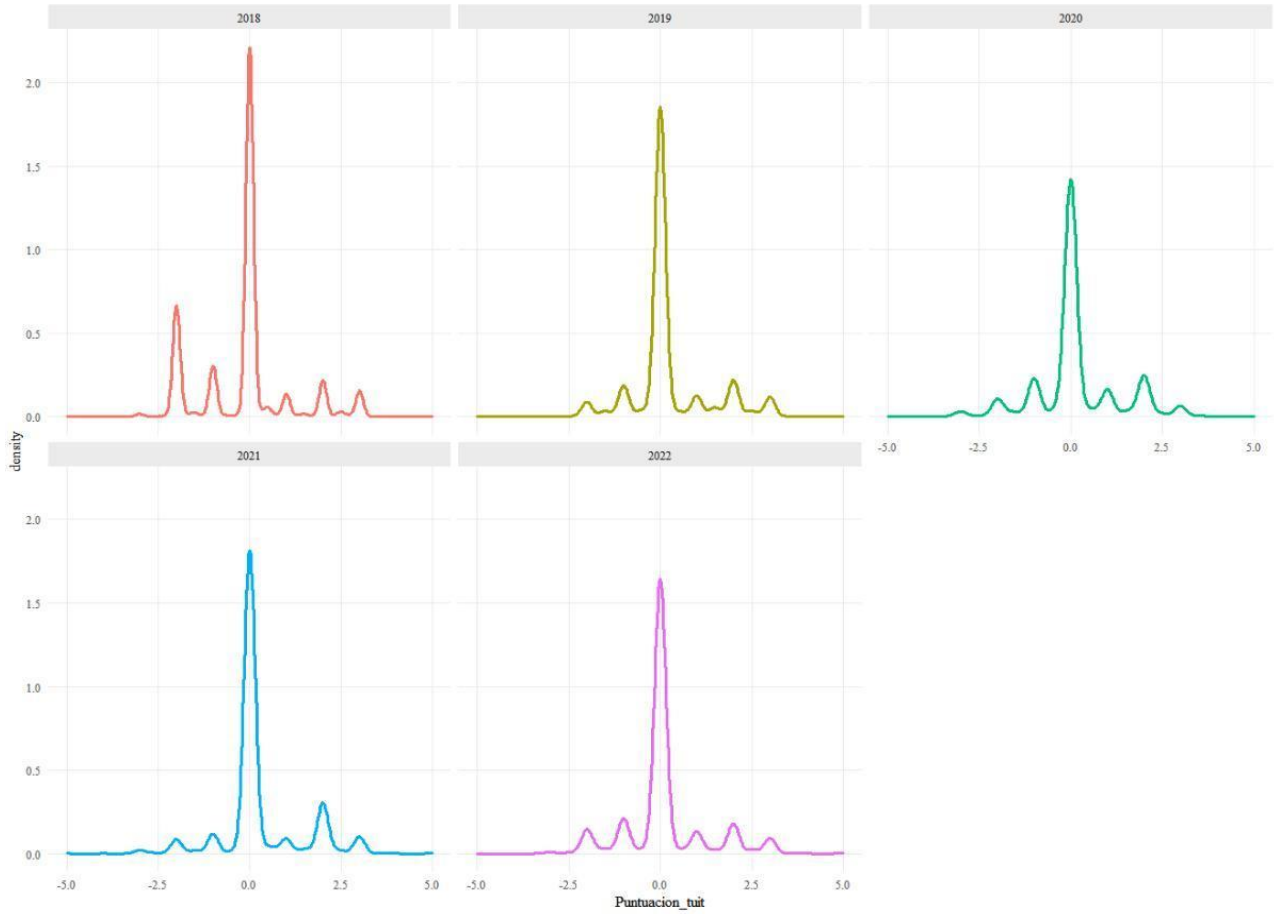


Ilustración XV

Fuente: Elaboración propia. Datos extraídos de Twitter.com

La *Ilustración XIV*, ayuda a entender mejor la distribución de los tuits enviados a lo largo de todo el período analizado. Como se puede observar, todos los tuits han sido escritos de una manera uniforme a lo largo de los años.

En 2018, ha habido un aumento de tuits relacionados con la universidad a principios de año, pero después se ha estabilizado.

En 2019, podemos ver que, tanto a principios como a finales de año, existen repuntes en el número de tuits sobre la Universidad Pontificia de Comillas.

En 2020 y 2021 la distribución de los tuits ha sido constante.

Finalmente, en 2022, vemos otra vez un repunte en abril o mayo.

La *Ilustración XV* muestra un gráfico de densidades distribuido por años, que indica de una forma más directa y cercana la carga de términos identificados para cada periodo.

4.4. Determinación de usuarios más relevantes

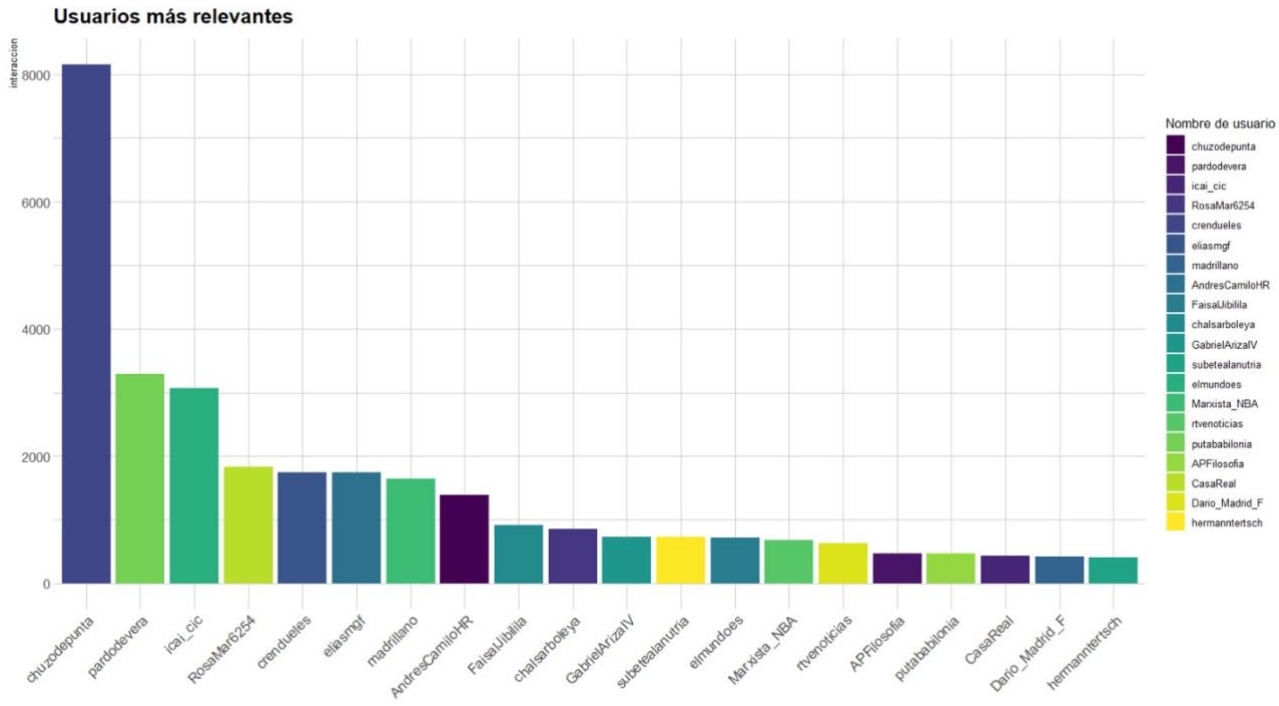


Ilustración XVI

Fuente: Elaboración propia. Datos extraídos de Twitter.com

A continuación, en la *Ilustración XVI*, se puede observar una lista de los 20 usuarios más relevantes. Para poder analizar qué individuos tienen más peso a la hora de hablar de la Universidad Pontificia de Comillas, se ha procedido a añadir una nueva variable en la base de datos denominada "Interacción", mencionada en el *Capítulo III*. Esta variable se determina a través del sumatorio de los retuits y likes que se han obtenido en todos los tuits escritos por los usuarios.

CAPÍTULO V: CONCLUSIONES

Este estudio ha analizado la reputación de la Universidad Pontificia de Comillas a través de un análisis temporal, puesto que se ha efectuado el análisis desde el 1 de enero de 2018 hasta el 31 de mayo de 2022, pero también se ha efectuado a través de un análisis espacial, ya que se han analizado todos los posibles tuits que tenían alguna conexión con la universidad.

El primer paso para la realización de este análisis ha sido la extracción de los tuits a través de dos formas distintas. La primera, gracias a sucesivas llamadas a la API de Twitter, que han permitido generar una base de datos pequeña, pero con mucha información. Esta base de datos se define como pequeña debido a las limitaciones que tiene Twitter en su política de privacidad para la extracción de información. La segunda forma se ha realizado a través de un web scraping, generando una base de datos con información contenida desde enero de 2018. Se han unido ambas bases de datos y se han eliminado los tuits repetidos.

En el segundo paso se ha efectuado un análisis inicial, en donde se ha podido observar como la base de datos contenía numerosas stopwords y elementos que no eran necesarios para la elaboración del análisis de sentimientos y el modelado de tópicos.

En el tercer paso del proceso, se ha procedido a la depuración y limpieza de la base de datos a través de la eliminación de caracteres y terminologías (*tokenización*), en donde se dividen los tuits en tokens, definiendo primero los delimitadores de los mismos, que suelen ser signos de puntuación y otros caracteres distintos a las letras del alfabeto. Seguido de esto, se ha procedido a la exclusión de palabras (*stopwords*) a través de varios paquets y librerías distintas. Después de haber efectuado una determinación de palabras equivalentes (*stemming*), ya se podrá hablar de una base de datos depurada, aunque nunca al 100%, ya que siempre hay elementos que se escapan a los filtrados. Y esto es una de las cosas que han sido más interesantes a la hora de efectuar el análisis de la reputación de la universidad, ya que por mucha limpieza o filtrados que se hagan, siempre van a quedar elementos sin depurar.

Como ejemplo se puede ver esto en el momento en el que, después de efectuar una primera exclusión de palabras, algunos de los tuits que ya esta había sido objeto de una primera limpieza, aún poseían terminologías en griego.

Así, se ha procedido a efectuar un modelado a través de algoritmo k-means (clustering) y a continuación se ha procedido a realizar un análisis de sentimientos.

Los resultados obtenidos en el *Capítulo IV* de este estudio han sido bastante satisfactorios y claros para poder tener una idea sobre cuál es la reputación de la Universidad Pontificia de Comillas durante los últimos años. Es importante mencionar que muchos de los términos con connotación negativa que se han extraído y analizado se mencionan con frecuencia en tuits relacionados con la institución debido a que la universidad posee un carácter muy altruista, quizás derivado de su historia o quizás debido a su estrecha vinculación con la Compañía de Jesús, en la que se defienden los principios de ayuda al prójimo y de respeto.

Para hablar sobre la connotación positiva que la universidad ostenta, hay que dejar claro que mantener una visión positiva de cualquier institución durante un largo período de tiempo genera muchas dificultades, y más en los años en los que se ha analizado la reputación de Comillas. Los años 2020 y 2022 han sido difíciles, y eso se ha podido notar en el análisis de este estudio. La pandemia generada por el COVID-19 ha creado una situación complicada para la universidad, por lo que esto ha sido observado a lo largo de este análisis de modelados y de sentimientos.

Cabe mencionar a su vez que, al igual que ha habido comentarios negativos durante esos años, en el año siguiente los usuarios de Twitter han seguido efectuando sentimientos positivos, felicitando a la institución. Esto nos indica una muy fuerte implicación por parte de la Universidad Pontificia de Comillas para sobreponerse a estímulos negativos. Debe continuar por este camino.

En relación a los tópicos que se han generado durante el estudio de la universidad, es importante recalcar que a través del Método de Silueta y del Método del Codo, se ha establecido un agrupamiento óptimo de dos clústeres. Esto es determinante pues permite

separar perfectamente todos los términos de la base de datos, incluyéndolos en dos grupos perfectamente diferenciados.

Así, los dos grupos podrían denominarse “Institución” y “Eventos” respectivamente. Como se ha mencionado anteriormente, uno de ellos está basado en el ambiente organizativo de la institución, mientras que el otro se basa en aspectos unos poco más externos.

El objetivo de este trabajo se ha conseguido con éxito, ya que gracias a este estudio se puede realizar una interpretación de la reputación de la Universidad Pontificia de Comillas. La siguiente fase del proceso, según el modelo CRISP-DM es la de implementación del mismo. Todas las marcas tienen un margen de mejora de sus reputaciones, y esta institución no iba a ser menos.

5.1 Limitaciones y trabajo en un futuro.

Para la realización de este estudio, siempre se pueden incorporar nuevos análisis para enriquecer la precisión del mismo.

Para una mejor implementación del modelado de tópicos y determinación de temas sobre la universidad, se pueden efectuar modelos basados en estadística como el análisis discriminante lineal (método matemático).

En relación con el análisis de sentimientos, se podría efectuar un código de programación que determinase o categorizase sentimientos como la tristeza y la alegría.

BIBLIOGRAFÍA

-LOZARES C. (1996). *La teoría de redes sociales*, Universitat Autònoma de Barcelona. Departament de Sociología, Barcelona; Obtenido en https://ddd.uab.cat/pub/papers/02102862n48/02102862n48p103.pdf

-DERCZYNSKI L., MAYNARD D. (2014), *Analysis of Named Entity Recognition and Linking for Tweets,* University of Sheffield, Sheffield, S1 4DP, UK.

-SUSTER S., TULKENS S. AND DAELEMANS W. (2017), *A Short Review of Ethical Challenges in Clinical Natural Language Processing,* University of Antwerp, Antwerpen, Bélgica.

-ZHAO B. (2017). *Web scraping*. *Encyclopedia of big data*, 1-3.

-DUBIAU L. Y M. ALE J. (2013), *Análisis de Sentimiento sobre un Corpus en Español*, Facultad de Ingeniería. Universidad de Buenos Aires, Argentina.

-RIDGE K. (2010), *Text Mining - The state of the art and the challenges*. Singapore. Disponible en https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.6973&rep=rep1&type=pdf

-NAKOV P. (2017), *Semantic Sentiment Analysis of Twitter Data,* Institute, HBKU, Doha, *Qatar*.

-TORRES D. (2013), *Diseño y aplicación de una metodología para análisis de noticias policiales utilizando minería de textos.* Universidad de Chile, Santiago de Chile, Chile.

-Twitter INC., (2022), *Twitter libraries — Twitter Developers,* San Francisco, California, Estados Unidos. Recuperado el 25 de mayo de 2022 https://developer.twitter.com

-GALLARDO ARANCIBIA J. A. (2013) *CRISP-DM Metodologías para el desarrollo de proyectos en minería de datos.* EPB 603 Sistemas del Conocimiento. Madrid, España.

http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037

-JACOBSON D., BRAIL G. & WOODS D. (2011), *Creating Channels with Aplication Programming Interfaces*, United States.

-LI C., WANG Y. & FAN T. (2018), *An Automatic Data Cleaning Procedure for Electron Cyclotron Emission Imaging on EAST Tokamak Using Machine Learning Algorithm.* Department of Engineering and Applied Physics, School of Physical Sciences, University of Science and Technology of China, No. 96 Jinzhai Road, Hefei, China.

-KALEV L. (2012). *Data mining methods for the content analyst: An introduction to the computational analysis of informational center*. New York: Routledge. ISBN: 978 0415895149, United States.

-FEINERER, I. (2013). *Introduction to the tm Package Text Mining in R*. Accessible en linea: https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

-JOACHIMS T. (2012). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* Universitat Dortmund Informatik, Germany. Disponible en https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

-ALLAIRE, J. (2012). *RStudio: integrated development environment for R.* Boston, Disponible http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.651.1157&rep=rep1&type=pdf#page=14

-MITCHELL T. (2015). *Machine Learning.* McGraw Hill, New York, United States.

-TONG, Z. y ZHANG, H. (2016, mayo). Una investigación de minería de textos basada en el modelado de temas LDA. En *Congreso Internacional de Ciencias de la Computación, Ingeniería y Tecnologías de la Información* (pp. 201-210).

-FERNÁNDEZ R., (2022). *Twitter: número de perfiles de la red social Twitter en*

España 2014–2021, Statista, s.f., Disponible en:

https://es.statista.com/estadisticas/520056/usuarios-de-twitter-en-espana/

-"Reputación, RSC y Comunicación en el ecosistema digital", jornada organizada por ESIC en colaboración con Global Alliance for Public Relations and Communication Management.

-BERMINGHAM A. and SMEATON A. (2012). *On Using Twitter to Monitor Political Sentiment and Predict Election Results*, Dublin, Ireland. Disponible en: https://www.researchgate.net/publication/267250109_On_Using_Twitter_to_Monitor_Political_Sentiment_and_Predict_Election_Results

-AGUILAR, M., (2020) *Análisis de sentimiento: Machine Learning para tu marca*. Topic flower, Disponible en

https://topicflower.com/blog/analisis-de-sentimiento-machine-learning-para-tu-marca/

-DAU, A., (2017) *Twitter presentó las API Premium*. Tecnogamming. Disponible en

https://tecnogaming.com/twitter-presento-las-api-premium/

-HARARY, F. (1969) *Graph Theory*. Addison-Wesley Publishing Company, New York.

-*Información sobre las API de Twitter*. (s.f.) Centro de ayuda, Twitter. Recuperado el 06 /06/ 2022 de https://help.twitter.com/es/rules-and-policies/twitter-api

-We are Social y Hootsuite (2018) *Q2 Global Digital statshot: essential insights into internet, social media, and e-commerce use around the world*, Recuperado en

https://www.slideshare.net/wearesocialsg/2018-q2-global-digital-statshot-94084375

-DÁVILA HERNÁNDEZ, F. y SÁNCHEZ CORALES, Y. (2022). Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas. Vol.4, nº2, Ciudad de la Habana, Cuba

-Fernando Sancho Camparrini, (2020). *"Algoritmos de clustering".* Disponible en: http://www.cs.us.es/~fsancho/?e=230

-J. Kleinberg, (2022) "*An impossibility theorem for clustering*" Proceedings of the 15th International Conference on Neural Information Processing Systems, NIPS'02, pp. 463–470, MIT Press. https://dl.acm.org/doi/10.5555/2968618.2968676

-PUNITHAVALLI M., PUNITHA S. C., NATHIYA G. (2020) *"An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm"*, International Journal of Computer Science and Information Security, vol. 7, nº 3, Disponible en: https://arxiv.org/pdf/1004.1743.pdf

-ARGINTZONA, J. (2020) *Que es la reputación de marca y por qué es importante*. Disponible en https://blog.digimind.com/es/insight-driven-marketing/que-es-la-reputacion-de-marca-y-por-que-es-importante

-IAB SPAIN y ELOGIA (2019) "Relación entre las Redes Sociales y e-commerce" *Estudio Anual Redes Sociales*. Disponible en https://iabspain.es/wp-content/uploads/2019/06/estudio-anual-redes-sociales-iab-spain-2019_vreducida.pdf

ANEXO:

Extracción de datos:

```
# Librerías
library(twitteR)
library(rtweet)
library(ggplot2)
library(dplyr)
library(tidytext)
library(rtweet)
library(tables)
library(scales)
library(RColorBrewer)
library(wordcloud2)
library(stringr)
```

```
# Direcciones centrales de la API

reqURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
options(httr_oauth_cache=T)

# Credenciales TWITTER

consumer_key <- "o7pZ6SxYBnVk42CSFuY8DPepJ"
consumer_secret <-
"7x0XgtP8UnlZT3fRsS94VwTfqnz0BW5EyWYaQOPYz1OO8vQoys"
access_token <- "76430855-0dZrg1o8EAVWnE0CsFjEXVkh9jyHwKO8Ndos0FoZF"
access_secret <- "FrzmzlsWSfFxe4dBgLlUel16FeKCn3KeWaC8SoePxqSYr"

options(httr_oauth_cache=TRUE)
setup_twitter_oauth(consumer_key,
                    consumer_secret,
                    access_token,
                    access_secret)

# Consulta a la API
data1 <- searchTwitter("Universidad Pontificia de Comillas
                       -filter:retweets",
                       n=3000,
                       lang = "es")

# Consulta a la API
data2 <- searchTwitter("UCOMILLAS
                       -filter:retweets",
                      n=3000,
                      lang = "es")
```

```
# Consulta a la API
data3 <- searchTwitter("ICADE
                       -filter:retweets",
                       n=3000,
                       lang = "es")

# Consulta a la API
data4 <- searchTwitter("Universidad de Comillas
                       -filter:retweets",
                       n=3000,
                       lang = "es")

# Consulta a la API
data5 <- searchTwitter("@ucomillas
                       -filter:retweets",
                       n=3000,
                       lang = "es")

# Consulta a la API
data6 <- searchTwitter("@ICADE_Derecho
                       -filter:retweets",
                       n=3000,
                       lang = "es")

# Consulta a la API
data7 <- searchTwitter("#CIDICADE
                       -filter:retweets",
                       n=3000,
                       lang = "es")

# Tweets a data fragme
df1 = twListToDF(data1)
df2 = twListToDF(data2)
df3 = twListToDF(data3)
```

```
df4 = twListToDF(data4)
df5 = twListToDF(data5)
df6 = twListToDF(data6)
df7 = twListToDF(data7)

# Unir tablas de datos:

data <- rbind(df1, df2, df3, df4, df5, df6, df7)
```

AGRUPAMIENTO

```
# Importar data:

data_api <- read.csv("data_comillas_API.csv",
                     comment.char="#")

data <- read.csv("Tweets_Comillas.csv",
                 comment.char="#")

data_api <- dplyr::select(data_api,
                          text,
                          created,
                          favoriteCount,
                          retweetCount,
                          id,
                          screenName)
names(data)
data <- dplyr::select(data,
                      text,
                      created_at,
                      like_count,
                      retweet_count,
                      tweet_id,
                      user_username)
```

```
names(data)[2] <- "created"
names(data)[3] <- "favoriteCount"
names(data)[4] <- "retweetCount"
names(data)[5] <- "id"
names(data)[6] <- "screenName"

tuits <- rbind(data_api,data)

data <- data %>%
  group_by(text) %>%
  slice(1)

# Sumar los favoritos con RT y almacenarlos en la variable interaccion
data$favoriteCount <- as.numeric(data$favoriteCount)
data$retweetCount <- as.numeric(data$retweetCount)
data$interaccion = rowSums (data[ , 3:4])

hist(data$interaccion)

data <- data[order(data$interaccion), ]
data <- head(data, 20)

data$text = gsub("( RT | via ) ( (?:\\b\\W*@\\w+)+)",
                 " ",
                 data$text)

# Eliminar direcciones web
data$text = gsub("http\\S*", " ", data$text)

# quitar links
data$text = gsub("http\\w+", " ", data$text)

# quitar @OTRASCUENTAS
```

```
data$text = gsub("@\\w+", " ", data$text)

# quitar simbolos de puntuación
data$text = gsub("[[:punct:]]", " ", data$text)

# quitar números
data$text = gsub("[[:digit:]]", " ", data$text)

# quitar saltos de línea y tabulaciones
data$text = gsub("[[:cntrl:]]", "",data$text)

data$text <- tolower(data$text)
data$text <- removeNumbers(data$text)
data$text <- removePunctuation(data$text)
data$text <- removeWords(data$text, words = stopwords("spanish"))
data$text <- removeWords(data$text,
                         stopwords::stopwords("es", source = "snowball"))
data$text <- removeWords(data$text,c("rt"))
data$text <- stripWhitespace(data$text)

data$text <- str_replace_all(data$text,
                             "[^[:alpha:][:space:]]*",
                             "")

lista_stopwords1 <- stopwords::stopwords("es", source = "nltk")
lista_stopwords2 <- stopwords::stopwords("es", source = "snowball")

lista_stopwords3 <- c("u","a","f","d",
                      "h","c","i"," ","u b")

# Se filtran las stopwords
data <- data %>%
  filter(!(text %in% lista_stopwords1))
```

```
data <- data %>%
  filter(!(text %in% lista_stopwords2))

data <- data %>%
  filter(!(text %in% lista_stopwords3))

docs <- data$text

CORPUS <- Corpus(VectorSource(docs))
# Utilizando la función Corpus(),
# indicamos la fuente de nuestro texto

# Verificar que el corpus se haya creado correctamente
VCorpus(VectorSource(CORPUS))

# Se contruye la matriz term-document,
# la cual es una tabla que contiene la
# frecuencia de las palabras.
# Esta matriz es el insumo principal para la construcción
# de la nube de palabras.

mtd <- DocumentTermMatrix(CORPUS)
mtd

mt2 <- as.matrix(mtd)
mt3 <- mt2

df <- as.data.frame(mt3)
mt3<-t(df)



# Calcular el numero optimo de cluster
```

```
library(factoextra)
datos <- scale(mt4)
fviz_nbclust(datos,
             FUNcluster = kmeans,
             method = "wss",
             k.max = 8) +
  labs(title = "Número óptimo de clusters")

library("cluster")
silk4<- silhouette(k2$cluster,dist=c)

plot(silk4,border=blues9)
View(MT)
p<-cbind(mt4,k2$cluster)

p1 <- as.data.frame(p)

uno <- filter(p1, V2=="1")
dos <- filter(p1, V2=="2")
names(p1)
fviz_cluster(k2, p,a.rm = FALSE,
             palette = c("#00AFBB","#2E9FDF", "#E7B800", "#FC4E07"),
             ggtheme = theme_minimal(),
             geom = c("point"),
             pointsize = 1.5,
             textsize=5,
             labelsize = 10,
             main = "Parcela de clusterización"
)

library("igraph")
library("tidygraph")
library("ggraph")
```

```
fviz_cluster(k2, c , ellipse.type = "convex",
             show.clust.cent = TRUE,
             ellipse = TRUE,
             ellipse.alpha = 0.2,
             repel = TRUE) +
  theme_bw() +
  labs(title = "Análisis de conglomerados") +
  theme(legend.position = "none")

library(Ckmeans.1d.dp)
x <- rnorm(mt4)
k<-2
result <- Ckmeans.1d.dp(x,k)
plot(result)
#k <- max(result$cluster)

result <- Ckmeans.1d.dp(x,k,mt4)
plot(result, main = "Agrupamiento ponderado (n términos)",
ylab="Frecuencias/término",xlab="n", sub = "A=red , B=black , C=")

plot(x, col=result$cluster, pch=result$cluster, cex=1.5,
     main="Agrupación Univariante con k-medias estimada",
     sub=paste("Número de clusters estimado", k))
abline(h=result$centers, col=1:k, lty="dashed", lwd=2)
legend("topright", paste("Cluster", 1:k), col=1:k, pch=1:k, cex=1.5, bty="n")

res <- Ckmeans.1d.dp(x, k=2,  result$cluster)
plot(res)
TemasPC16b <- cbind(cluster =res$cluster,mt4)

g1<-TemasPC16b[res$cluster==1, ]
g2<-TemasPC16b[res$cluster==2, ]
g3<-TemasPC16b[res$cluster==3, ]
#g4<-TemasPC16b[res$cluster==4, ]
```

```
t1<- as.data.frame(g1)
t2<- as.data.frame(g2)
t3<- as.data.frame(g3)
#t4<- as.data.frame(g4)

t1<- select(t1,mt4)
t1<- as.data.frame(t1)

t2<- select(t2,mt4)
t2<- as.data.frame(t2)

t3<- select(t3,mt4)
t3<- as.data.frame(t3)

t11<-t(t1)
t22<-t(t2)
t33<-t(t3)


t111<-data.frame(Palabra=names(t1),frecuencia=g1)
t222<-data.frame(Palabra=names(t2),frecuencia=g2)
t333<-data.frame(Palabra=names(t3),frecuencia=g3)

d1 <- select(t111, -frecuencia.cluster,-Palabra)
write.csv(d1, file="d1.csv")
d1 <- read.csv("~/R/R/TUITS/d1.csv")

d2 <- select(t222, -frecuencia.cluster,-Palabra)
write.csv(d2, file="d2.csv")
d2 <- read.csv("~/R/R/TUITS/d2.csv")

d3 <- select(t333, -frecuencia.cluster,-Palabra)
write.csv(d3, file="d3.csv")
```

```
d3 <- read.csv("~/R/R/TUITS/d3.csv")
#d4 <- select(t4, -frecuencia.cluster)

d11 <- d1[1:23, ]
d22 <- d2[1:30, ]
d33 <- d3[1:40, ]

library(wordcloud2)

wordcloud2(d11, size = 0.5, shape = "cloud", color="random-dark", ellipticity = 1)
wordcloud2(d22, size = 0.5, shape = "cloud", color="random-dark", ellipticity = 1)
wordcloud2(d33, size = 0.5, shape = "cloud", color="random-dark", ellipticity = 1)
```

Análisis de sentimiento

```
# SELECCIONAR VARIABLES DE TRABAJO
#tuits <- select(tuits,id,created,etiqueta,text)

afinn <- read.csv("lexico_afinn.csv",
                  stringsAsFactors = F,
                  fileEncoding = "latin1") %>%
  tbl_df()

afinn

# Fechas
tuits <-
  tuits %>%
  separate(created, into = c("Fecha", "Hora"), sep = " ") %>%
  separate(Fecha, into = c("Periodo", "Mes", "Dia"), sep = "-",
           remove = FALSE) %>%
  mutate(Fecha = dmy(Fecha),
         Semana = week(Fecha) %>% as.factor(),
         text = tolower(text))
```

```
# ANALISIS DE SENTIMIENTO / PUNTUACION POR TUIT
tuits_afinn <-
  tuits %>%
  unnest_tokens(input = "text", output = "Palabra") %>%
  inner_join(afinn, ., by = "Palabra") %>%
  mutate(Tipo = ifelse(Puntuacion > 0, "Positiva", "Negativa")) %>%
  rename("Categoria" = Categoria)

table(tuits_afinn_neutra$Tipo)

tuits <-
  tuits_afinn %>%
  group_by(id) %>%
  summarise(Puntuacion_tuit = mean(Puntuacion)) %>%
  left_join(tuits, ., by = "id") %>%
  mutate(Puntuacion_tuit = ifelse(is.na(Puntuacion_tuit), 0, Puntuacion_tuit)) %>%
  rename("Categoria" = Categoria)

# Explorando los datos, (CANTIDADES DE TERMINOS CON SENTIMIENTO
POSITIVO Y NEGATIVO ENCONTRADO EN CADA UNO DE LOS
DOCUMENTOS Y ETIQUETADOS POR SUS TENDENCIAS)

# Total
tuits_afinn %>%
  count(Categoria)

# Únicas
tuits_afinn %>%
  group_by(Categoria) %>%
  distinct(Palabra) %>%
  count()
```

```
# palabras positivas y negativas más usadas por cada tendencia

map(c("Positiva", "Negativa"), function(sentimiento) {
  tuits_afinn %>%
    filter(Tipo ==  sentimiento) %>%
    group_by(Categoria) %>%
    count(Palabra, sort = T) %>%
    top_n(n = 10, wt = n) %>%
    ggplot() +
    aes(Palabra, n, fill = Categoria) +
    geom_col() +
    facet_wrap("Categoria", scales = "free") +
    scale_y_continuous(expand = c(0, 0)) +
    coord_flip() +
    labs(title = sentimiento)
})

conteo_palabras <- tuits_afinn %>%
  group_by(Tipo) %>%
  count(Palabra)

conteo_palabras[1:20, ]

conteo_palabras %>%
  group_by(Tipo) %>%
  top_n(20) %>%
  ggplot(aes(reorder(Palabra, n), n, fill = Tipo)) +
  geom_bar(alpha = 0.8, stat = "identity", show.legend = FALSE) +
  facet_wrap(~Tipo, scales = "free_y") +
  labs(y = "Cantidad por palabra", x = "Palabras o terminos") +
  coord_flip()

mapeo <- tuits_afinn %>%
  select(Mes,Tipo,Categoria,Puntuacion)
```

```
# Eliminar términos que no cumplan con las condiciones del modelo:

tuits_afinn <-
  tuits_afinn %>%
  filter(Palabra != "home")%>%
  filter(Palabra != "latex")

#Graficamos nuevamente

map(c("Positiva", "Negativa"), function(sentimiento) {
  tuits_afinn %>%
    filter(Tipo ==  sentimiento) %>%
    group_by(Categoria) %>%
    count(Palabra, sort = T) %>%
    top_n(n = 10, wt = n) %>%
    ggplot() +
    aes(Palabra, n, fill = Categoria) +
    geom_col() +
    facet_wrap("Categoria", scales = "free") +
    scale_y_continuous(expand = c(0, 0)) +
    coord_flip() +
    labs(title = sentimiento) +
    tema_graf
})

tuits_afinn_fecha <-
  tuits_afinn %>%
  group_by(id) %>%
  mutate(Suma = mean(Puntuacion)) %>%
  group_by(Categoria, Dia) %>%
  summarise(Media = mean(Puntuacion))
tuits_afinn_fecha
```

```
# Gráficamos nuevamente excluyendo las palabras y terminos "maduro" y "lucha"

conteo_palabras <- tuits_afinn %>%
  group_by(Tipo) %>%
  count(Palabra)

conteo_palabras[1:20, ]

conteo_palabras %>%
  group_by(Tipo) %>%
  top_n(20) %>%
  ggplot(aes(reorder(Palabra, n), n, fill = Tipo)) +
  geom_bar(alpha = 0.8, stat = "identity", show.legend = FALSE) +
  facet_wrap(~Tipo, scales = "free_y") +
  labs(y = "Cantidad por palabra", x = "Palabras o terminos") +
  coord_flip()


#Comparando sentimientos positivos y negativos

tuits_afinn %>%
  count(Categoria, Tipo) %>%
  group_by(Categoria) %>%
  mutate(Proporcion = n / sum(n)) %>%
  ggplot() +
  aes(Categoria, Proporcion, fill = Tipo) +
  geom_col() +
  scale_y_continuous(labels = percent_format()) +
  tema_graf +
  theme(legend.position = "top")

# Bloxplots (diagrama caja y bigotes)

tuits %>%
```

```
  ggplot() +
  aes(Categoria, Puntuacion_tuit, fill = Categoria) +
  geom_boxplot() +
  ylim(-0.5,0.5) +
  tema_graf


# Se pueden analizar las Categoria de sentimientos usando las funciones de densidad de las puntuaciones

tuits %>%
  ggplot() +
  aes(Puntuacion_tuit,
      color = Categoria) +
  geom_density(size = 1) +
  facet_wrap(~Categoria) +
  tema_graf

# Categoria a través del tiempo
tuits %>%
  ggplot() +
  aes(Puntuacion_tuit, color = Categoria) +
  geom_density() +
  facet_grid(Categoria~Dia) +
  tema_graf
```

Cuentas relevantes:

```
library("tidyverse");
library("tidytext");
library("tm");
library("lubridate");
library("zoo");
library("scales");
```

```
library("dplyr");
library("tibble")

data_api <- read.csv("data_comillas_API.csv",
                     comment.char="#")

data <- read.csv("Tweets_Comillas.csv",
                 comment.char="#")

data_api <- dplyr::select(data_api,
                          text,
                          created,
                          favoriteCount,
                          retweetCount,
                          id,
                          screenName)

names(data)

data <- dplyr::select(data,
                      text,
                      created_at,
                      like_count,
                      retweet_count,
                      tweet_id,
                      user_username)

names(data)[2] <- "created"
names(data)[3] <- "favoriteCount"
names(data)[4] <- "retweetCount"
names(data)[5] <- "id"
names(data)[6] <- "screenName"

tuits <- rbind(data_api,data)
```

```
# Tema para los grÃ¡ficos
tema_graf <-
  theme_minimal() +
  theme(text = element_text(family = "serif"),
        panel.grid.minor = element_blank(),
        strip.background = element_rect(fill = "#EBEBEB",
                                        colour = NA),
        legend.position = "none",
        legend.box.background = element_rect(fill = "#EBEBEB",
                                             colour = NA))

tuits <- tuits %>%
  mutate(Categoria = case_when(created < "2018-12-31" ~ '2018',
                               created < "2019-12-31" ~ '2019',
                               created < "2020-12-31" ~ '2020',
                               created < "2021-12-31" ~ '2021',
                               created < "2022-12-31" ~ '2022'
  ))

tuits <- dplyr::select(tuits,
                       text,
                       favoriteCount,
                       retweetCount,
                       id,
                       screenName,
                       created,
                       Categoria)

tuits$favoriteCount <- as.numeric(tuits$favoriteCount)
tuits$retweetCount <- as.numeric(tuits$retweetCount)



# Sumar los favoritos con RT y  almacenarlos en la variable interaccion
```

```
tuits$interaccion = rowSums (tuits[ , 2:3])
tuits <- tuits %>%
  group_by(text) %>%
  slice(1)

# segmentar por años:

uno <- filter(tuits, Categoria=="2018")
uno <- uno[order(-uno$interaccion), ]
uno <- head(uno,20)
ggplot(uno, aes(x =
                  reorder(screenName, -interaccion, mean),
                y = interaccion,
                fill=uno$screenName)) +
  geom_bar(stat = "identity") +
  theme_ipsum(base_family ="TT Arial") +
  scale_fill_viridis_d(labels = uno$screenName,
                       legend_title <- "Nombre de usuario") +
  labs(title = "Usuarios más relevantes - 2018",
       x = " ") +
  theme(axis.text.x = element_text(angle = 45,
                                   size = 12,
                                   hjust = 1))



dos <- filter(tuits, Categoria=="2019")
dos <- dos[order(-dos$interaccion), ]
dos <- head(dos,20)
ggplot(dos, aes(x =
                  reorder(screenName, -interaccion, mean),
                y = interaccion,
                fill=dos$screenName)) +
  geom_bar(stat = "identity") +
```

```
  theme_ipsum(base_family ="TT Arial") +
  scale_fill_viridis_d(labels = dos$screenName,
                       legend_title <- "Nombre de usuario") +
  labs(title = "Usuarios más relevantes - 2019",
       x = " ") +
  theme(axis.text.x = element_text(angle = 45,
                                   size = 12,
                                   hjust = 1))

tres <- filter(tuits, Categoria=="2020")
tres <- tres[order(-tres$interaccion), ]
tres <- head(tres,20)
ggplot(tres, aes(x =
                   reorder(screenName, -interaccion, mean),
                 y = interaccion,
                 fill=tres$screenName)) +
  geom_bar(stat = "identity") +
  theme_ipsum(base_family ="TT Arial") +
  scale_fill_viridis_d(labels = tres$screenName,
                       legend_title <- "Nombre de usuario") +
  labs(title = "Usuarios más relevantes - 2020",
       x = " ") +
  theme(axis.text.x = element_text(angle = 45,
                                   size = 12,
                                   hjust = 1))

cuatro <- filter(tuits, Categoria=="2021")
cuatro <- cuatro[order(-cuatro$interaccion), ]
cuatro <- head(cuatro,20)
ggplot(cuatro, aes(x =
                     reorder(screenName, -interaccion, mean),
                   y = interaccion,
                   fill=cuatro$screenName)) +
  geom_bar(stat = "identity") +
```

```
  theme_ipsum(base_family ="TT Arial") +
  scale_fill_viridis_d(labels = cuatro$screenName,
                       legend_title <- "Nombre de usuario") +
  labs(title = "Usuarios más relevantes - 2021",
       x = " ") +
  theme(axis.text.x = element_text(angle = 45,
                                   size = 12,
                                   hjust = 1))

cinco <- filter(tuits, Categoria=="2022")
cinco <- cinco[order(-cinco$interaccion), ]
cinco <- head(cinco,20)
ggplot(cinco, aes(x =
                    reorder(screenName, -interaccion, mean),
                  y = interaccion,
                  fill=cinco$screenName)) +
  geom_bar(stat = "identity") +
  theme_ipsum(base_family ="TT Arial") +
  scale_fill_viridis_d(labels = cinco$screenName,
                       legend_title <- "Nombre de usuario") +
  labs(title = "Usuarios más relevantes - 2022",
       x = " ") +
  theme(axis.text.x = element_text(angle = 45,
                                   size = 12,
                                   hjust = 1))

# Total
tuits <- tuits[order(-tuits$interaccion), ]
tuits <- head(tuits,20)
names(tuits)
ggplot(tuits, aes(x =
                    reorder(screenName, -interaccion, mean),
            y = interaccion,
            fill=tuits$screenName)) +
```

```
  geom_bar(stat = "identity") +
  theme_ipsum(base_family ="TT Arial") +
  scale_fill_viridis_d(labels = tuits$screenName,
                       legend_title <- "Nombre de usuario") +
  labs(title = "Usuarios más relevantes",
       x = "Interacciones totales") +
  theme(axis.text.x = element_text(angle = 45,
                                   size = 12,
                                   hjust = 1))
str(tuits)
```