UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Clasificador de textos del Servicio de Administración Tributaria para análisis de mensajes en Twitter mediante algoritmos de Machine Learning Naive-Bayes

PROYECTO DE TRABAJO PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I O

P R E S E N T A

CÉSAR ANTONIO CORTÉS HERNÁNDEZ

Tutor

M. en F. Jorge Luis Reyes García

Ciudad de México, 2021

1. Datos del alumno

Nombre: César Antonio Cortés Hernández

Número de cuenta: 310013330

2. Datos del Tutor

M. en F. Jorge Luis Reyes García

3.- Sinodal 1

M. en C. Jaime Vázquez Alamilla

4. Sinodal 2

Act. Eduardo Selim Martínez Mayorga

5. Sinodal 3

Dra. Ruth Selene Fuentes García

6.-Sinodal 4

Act. Yadira Rivas Godoy

Agradecimientos

Gracias a la UNAM que me dio amigos, grandes profesores, diversiones, educación, crecimiento personal, desarrollo intelectual y las bases para lograr el éxito profesional.

Gracias a mi tutor Jorge, porque es un excelente profesor, un hombre con vocación, y guía para concluir esta meta, que representa uno de los grandes logros de la vida de una persona.

Gracias a cada uno de los sinodales Ruth, Yadira, Eduardo y Jaime, por cada una de las aportaciones realizadas para la elaboración y conclusión de la presente investigación, ya que gracias a ellas he concluido satisfactoriamente una de las etapas más importantes de mi vida profesional.

Gracias a mis papás y familia por su gran paciencia, apoyo y motivación para concluir de forma exitosa esta etapa de mi vida.

Gracias a Lupita por su apoyo y consuelo incondicional que recibí para este importante logro.

Gracias a cada uno de mis amigos, porque siempre me dieron lo mejor de ellos.

Índice

Índice de tablas

Índice de ilustraciones

Índice de gráficos

Introducción

Machine learning es una herramienta que tiene un gran alcance en distintas áreas, ya que ayuda a realizar tareas de forma eficiente y en un menor tiempo a comparación de un ser humano.

Dado el auge digital al que nos enfrentamos hoy en día, las redes sociales se han vuelto de interés con un universo extenso de información, de esta forma es primordial el análisis para conocer la interacción que existe entre los usuarios.

Para cualquier servicio que brinde algún tipo de atención mediante una red social, es indispensable realizar un análisis de la información para poder conocer el mercado meta y así analizar dudas, quejas o comentarios sobre el servicio, por lo que la clasificación de textos tiene gran relevancia para poder realizar análisis y crear un beneficio de manera bilateral.

En este estudio, en el título se menciona la institución, Servicio de Administración Tributaria, lugar donde se utilizaron ciertos algoritmos, sin embargo, por cuestiones de privacidad de datos personales, se decidió aplicar a una base sobre el Servicio de Transporte Colectivo Metro de la CDMX para visualizar las problemáticas existentes en este servicio.

En el primer capítulo se hará una introducción a los temas principales del trabajo, la historia de la clasificación de textos para conocer el propósito y evolución que ha tenido a lo largo del tiempo esta rama de *machine learning*. Relatar sobre características y usos en Twitter y finalmente se hará un resumen sobre la historia de la Red de Transporte Metro de la Ciudad de México.

El segundo capítulo plasmará distintos algoritmos que se ocupan para la clasificación de textos, realizando una indagación más detallada en el algoritmo Naive-Bayes mismo que se aplicará para fines de esta investigación.

En el tercer capítulo se mostrará el análisis de la explotación de la base de datos del metro de la CDMX en el periodo de junio y julio del 2019, se realizará un análisis descriptivo de la información para finalmente explicar, aplicar a través del algoritmo la clasificación; y mostrar los resultados que se obtienen de la misma.

Finalmente, el cuarto capítulo contendrá las conclusiones sobre la aplicación del algoritmo, así como las obtenidas en el análisis en cuestión.

Capítulo 1. Clasificación de Textos

1.1 Historia y motivación de estudiar la clasificación de textos

En el año de 1439, Lorenzo Valla cercano a la edad de 30 años tenía un amplio conocimiento y un análisis detallado sobre el leguaje latino, observó que tenía suficiente soporte para desmentir que el decreto imperial fue escrito por Constantino I. Su búsqueda se centró en los términos anacrónicos contenidos en el texto, donde concluyó que el estilo de éste era más cercano al latín del medievo que a la época del emperador. Sin intención, al realizar la investigación del texto utilizó métodos como análisis de frecuencias, Valla quedó en un punto muy cercano a lo que ahora es conocido como minería de textos.

Otro acercamiento que se cuenta en la historia es la del Padre Roberto Busa, sacerdote jesuita italiano, el cual con el propósito de concluir su tesis en 1946 decidió explorar el trabajo de Tomás de Aquino para poder crear un índice de todas sus obras, con el objetivo de poder encontrar de manera sencilla cada una de las palabras contenidas, y asociarlas con las obras en las que se encontraba cada palabra. Realizar esta tarea tomaría mucho tiempo, sin embargo en 1949 viajó a Estados Unidos buscando reemplazar un total de 10,000 fichas escritas a mano, llegó con Thomas J. Watson fundador de IBM, quien a pesar de tener pocas probabilidades de brindar su ayuda, decidió apoyar el trabajo de Busa.

Años más tarde, con el documento *The Federalist Papers*, artículos donde se buscaba ratificar la Constitución de los Estados Unidos, fueron publicados por Alexander Hamilton, James Madison y John Jay, quienes compartieron la autoría bajo el nombre de “Publius” sin revelar quien había escrito cada uno de los artículos. Fue un gran objeto de estudio para el análisis de textos, bajo la premisa de conocer a los verdaderos autores de cada uno de los artículos. La investigación quedó al mando de Frederick Mosteller, Frederick Williams y David Wallace, el estudio mostró los inicios en la aplicación de la inferencia Bayesiana. Crearon una base sobre los potenciales autores donde se incluyeron sentencias escritas por Hamilton y Madison, mismas que fueron ocupadas para apoyar la atribución de sus trabajos. El resultado fue publicado en 1964 consiguiendo un gran interés generando críticas y estudios relacionados, éste se considera uno de los primeros documentos que ocuparon la probabilidad en el análisis de textos.

La minería de textos es una disciplina con base en la minería de datos, los orígenes de la minería de datos es a finales de la década de 1980 e inicios de 1990, tuvo un desarrollo por el aprendizaje de máquina. Cubre una serie de herramientas y metodologías que están diseñadas para identificar patrones interesantes de las bases de datos

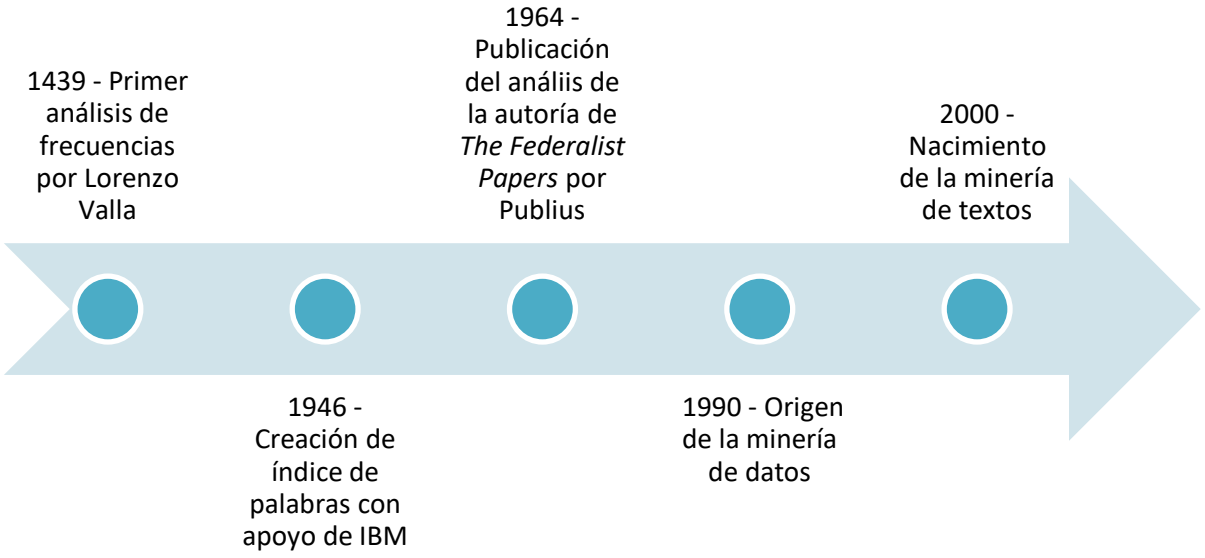
La minería de datos es separada de las prácticas generales que involucra el análisis de la información, ya que el objetivo principal es la identificación de conocimientos que previamente no se conocían.

La minería de datos se sustenta en gran medida del campo de la estadística. Las aplicaciones pueden ser la unión de bases de datos, predecir tendencias, hacer uso de información en tiempo real y apoyo en la toma de decisiones mediante el desarrollo de modelos; mismas que son capaces de confirmar o refutar prejuicios sobre la información, así como la identificación de nuevas conjeturas. La aplicación de la minería de datos es útil para conjuntos de información suficientemente grandes y complejos, de tal forma que su análisis de forma manual sea complicado.

Cualquier tipo de información puede ser analizada, en muchas ocasiones el origen más accesible de la información está disponible en texto no estructurado, esta es la razón por la cual la minería de textos ha tenido un gran auge en el campo de la minería de datos.

El área de la minería de textos es muy reciente, el nacimiento se da a principios del año 2000, en inicio se había planteado darle el nombre de “minería de datos de texto” por ser una variación del campo de la minería de datos, concluyéndose como minería de textos.

Ilustración 1 - Línea del tiempo minería de textos



La minería de textos es definida por Tuffery (2011) como, el conjunto de técnicas y métodos utilizados para el proceso de la información del lenguaje natural de texto disponible en gran medida por documentos de computadoras, con el objetivo de extraer y estructurar su contenido y temas, con el propósito de un rápido análisis, encontrar información oculta o para la toma de decisiones. Es un estudio distinto al análisis de los estilos de textos que buscan identificar al autor o la fecha del escrito, teniendo en común el análisis de la estadística lingüística. La idea se puede plantear de la siguiente forma:

$$Minería\ de\ textos = Análisis\ de\ vocabulario + minería\ de\ datos$$

La minería de textos se popularizó por el gran volumen de información en forma de texto en la sociedad, y por supuesto del almacenamiento de los mismos en computadoras.

En la minería de textos existen al menos 2 métodos:

- El método descriptivo puede ser ocupado para encontrar temas dentro de un conjunto de documentos, incluso sin tener un conocimiento previo sobre la información.
- El método predictivo, en este método se encuentran reglas, las cuáles de forma automática pueden realizar la asignación de los documento en temas predefinidos.

Existe una relación entre la minería de textos y un grupo de documentos de lingüística computacional, un área que hace uso de metodologías computacionales para identificar patrones y similitudes en el uso del lenguaje

Así mismo, tiene una relación con el procesamiento de lenguajes naturales, que es un área amplia que se utiliza para entender y manipular la información en una parte del texto para poder extraer resultados útiles.

Hay una gran cercanía entre la minería de textos, la extracción de información y la administración del conocimiento; ambas herramientas están diseñadas para aplicar la minería de textos, que incluye filtros que son diseñados para modificar un texto complejo en uno simple, esto permite extraer la información que es de nuestro interés, así como combinarlo con otra para que nos pueda brindar un conocimiento mayor sobre temas en los que se está involucrado.

- La extracción de información, uno de sus objetivos es obtener datos estructurados de textos que no son estructurados.
- Administración de conocimiento, explora el uso de sistemas de información para poder construir, compartir y aplicar conocimiento; normalmente dentro de un tema en específico.

Para una buena administración del conocimiento dentro de cierto contexto, es indispensable que el origen de la extracción de la información sea heterogénea.

De la misma forma, comparte un vínculo con la recuperación de información, ambos tienen el inconveniente de la interpretación y caracterización del texto, y buscan filtrar la información, cada una con un propósito:

- En la recuperación de la información, el objetivo es agrupar las características de cada texto para poder indexarlo y que esté disponible en la búsqueda que los usuarios requieran para ese tipo de texto.
- Para la minería de textos, se busca la identificación sobre los temas que el texto contiene.

El aprendizaje de máquina también comparte similitudes con la minería de textos, y tiene el propósito de brindar a las computadoras la habilidad de poder generar un conocimiento sobre la información sin tener que programarlo, se aprende sobre el dominio de la aplicación a través de la evaluación de la información. Éste generalmente se basa en la estadística y en la evidencia de la información para poder clasificarla. Es bastante útil en varios campos de la informática que trabaja con la evaluación de la información que el usuario ingresa. Los métodos son buenos encontrando formas para aproximarse al juicio humano sobre problemas utilizando ejemplos disponibles para la creación de reglas, ya que los algoritmos pueden probar conjuntos de reglas a gran velocidad y de forma automática, además son capaces de construir de forma iterativa mejores reglas que las que un humano pueda crear.

Otro campo con similitud a la minería de textos es la interacción humano-máquina, como lo es la computación afectiva, que es la habilidad de una computadora para prestar atención a las emociones humanas, el sentimiento causado por un evento, situación o comunicación. La investigación se basa en la premisa de que el afecto puede depender de señales fisiológicas, lenguaje hablado o señales no verbales para identificar el actual estado emocional del usuario. Aplicaciones de la minería de textos están contenidas en la computación afectiva que comprende algunas disponibles para el investigador que esté buscando caracterizar, entender, sentir y responder a una respuesta emocional; los textos también pueden ser analizados para buscar indicios sobre la personalidad de autores.

La minería de textos es un proceso interdisciplinario, que tiene una colaboración entre los individuos y las especialidades, que van desde lo técnico hasta las humanidades. Existen barreras legales y éticas que reduce su aplicación a fines escolares ya que no siempre es posible obtener derechos para aplicar la minería de textos en información, en especial

información privada o publicaciones oficiales, por lo que no se ha podido ocupar en varios campos en los cuáles sería de bastante utilidad. En la práctica, es necesario contar con un experto sobre el dominio de la aplicación, y por lo regular son pocos los recursos o herramientas disponibles en áreas específicas para poder identificar enfoques y proporcionar información para apoyar en el desarrollo y mejora del rendimiento de las herramientas a ocupar.

Ciertos proyectos contienen un entendimiento detallado sobre el lenguaje natural, éstos requieren un mayor apoyo de lingüistas, esto no significa que siempre se deba tener a este tipo de especialista.

La minería de textos no es una disciplina individual, más bien uno de los puntos finales para especialistas en información, informáticos, ingenieros y especialistas en la materia; por lo que no se puede dar una descripción única sobre el perfil que un experto en minería de textos debe tener. Así mismo, no existe una carrera o especialización que prepare de forma óptima a las personas interesadas en esta disciplina.

El análisis de texto automatizado es un área muy significativa para ser ignorada ya que ha sido parte en muchos otros campos, particularmente en la investigación.

Algunas de las herramientas ocupadas en la minería de textos inicialmente fueron discutidas en conferencias sobre inteligencia artificial y aprendizaje de máquina

Actualmente, con las redes sociales existe un gran volumen de información que se comparte a cada momento alrededor del mundo, muchos de los espacios creados funcionan con la finalidad de mantener una comunicación más cercana y personal con su red de usuarios, por lo que es importante poder captar dicha interacción para conocer los comentarios que los usuarios manifiestan. Una plataforma que en su mayoría está compuesta de información en forma de texto es Twitter.

Para fines de esta investigación, el algoritmo que se aplicará será similar en el procesamiento, aunque la fuente de datos es distinta.

1.2 Origen e historia de Twitter

Twitter es un servicio de microblogging, los usuarios siguen a otros o son seguidos, a diferencia de la mayoría de las redes sociales, la relación de seguir y ser seguido no es de forma recíproca. Un usuario seguidor significa que recibirá todos los mensajes, o tweets, de los usuarios que sigue, siempre y cuando no hayas bloqueado las notificaciones. La práctica común para poder crear un tweet contiene las siguientes características:

- “@” seguido por el identificador del usuario.
- “#” seguido por una palabra que representa una etiqueta.
- Mensaje con una restricción a 280 caracteres incluyendo los dos puntos anteriores.
- RT se entiende como *retweet*, funcionalidad ocupada para compartir la información de un tweet respetando la autoría del usuario.

Inicialmente existía una restricción a 140 caracteres por tweet, sin embargo ante la diferencia en la forma de escribir de los distintos idiomas, en septiembre de 2017 se extendió la restricción a 280 caracteres.

La aplicación de estas características define un ambiente cómodo para poder expresarse de forma breve.

Evan Williams, uno de los fundadores de la compañía Twitter, fundó Pyra Labs una *startup*, desde la que se creó Blogger.com que es una página web que funciona para la creación y gestión de blogs. Williams es quien inventó el término “blogger” y en el año 2003 recibió un reconocimiento por el *Technology Review* del MIT formando parte del Top 100 de las personas más innovadoras menores de 35 años del mundo, mismo año que Google compró Pyra Labs.

Al siguiente año Williams se deslindó de Google para poder fundar Odeo, una compañía dedicada a los podcast, Williams decidió llamar a su compañero Biz Stone para que formara parte de la misma, que también trabajó en Google. Esta empresa tuvo su oficina en el departamento de Williams donde contrató a Jack Dorsey y a Blaine Cook, programador e ingeniero respectivamente. Casi de forma simultánea Apple anunció que iTunes iba a incluir una plataforma de podcast, y Apple al tener una estimación de venta de 200 millones de iPods, haría que Odeo se diera cuenta que no podría competir en ese negocio.

Trabajando en distintos proyectos, Dorsey le comentó a Williams una idea que tenía en mente sobre un servicio basado en mensajes de estatus, donde un usuario pudiera enviar un mensaje mediante SMS con el fin de comunicarlo a un grupo de personas. Noah Glass, quien estuvo a cargo del proyecto, decidió llamarlo “twttr”. Al querer completar la idea, buscaron el significado de la palabra “twitter” que contenía 2 significados:

- Uno significaba “un corto estallido de información inconsecuente”;
- Y el otro significaba “pitido de pájaros”.

Por lo que se dieron cuenta que esta palabra reflejaba de manera más precisa lo que era su producto.

En marzo de 2006, ya tenían en funcionamiento un prototipo de forma interna, el primer tweet fue publicado por Dorsey el 21 de marzo a las 12:50 PM, el cuál decía “inviting coworkers” (invitando a compañeros de trabajo), decidieron abrirlo al público cuatro meses después. Ese mismo año, aconteció un pequeño terremoto en la zona y en Twitter, con los usuarios que ya contaba, se extendió la noticia de forma inmediata causando un gran asombro sobre la rapidez de servicio.

A finales del 2006, Williams, Dorsey y Stone fundaron Obvious Corp, con la finalidad de adquirir las propiedades de Odeo, donde estaba incluida Twitter, y en abril del siguiente año, ésta se convertiría en una compañía independiente. Williams presentaba una carta a los inversores de Odeo exponiendo que no veía futuro para la compañía y no se sentía cómodo con esa situación, proponiendo la compra de sus acciones para que ellos no tuvieran pérdidas, así mismo expuso en la carta que Twitter era una pieza de valor que veía en la compañía por lo que seguiría invirtiendo en ésta. La venta fue realizada al estar de acuerdo los inversores con los puntos mencionados en la carta.

En abril de 2007, Jack Dorsey quedó como director ejecutivo de Twitter convirtiéndose en una compañía propia. En este momento, no se tenía una gran percepción de si Twitter pudiera funcionar, pero llegó un festival, *South by SouthWest (SXSW)*, en Austin, Texas, en el que fue el protagonista llevándose el premio de *Southwest Web Award*; en este festival colocaron salas con pantallas gigantes con transmisiones sobre Twitter y la popularidad de la plataforma despegó, fue un negocio redondo, mientras los ponentes y panelistas mencionaban la plataforma de manera repetida, los bloggers la promocionaban, se logró que la cantidad de tweets enviados a lo largo del evento subiera de 20,000 a 60,000.

Uno de los primeros casos en los que Twitter empezaba a tener efectos directos y prácticos en la vida real de las personas fue el 16 de marzo del 2008, cuando el estudiante egipcio de periodismo, James Karl Buck fue detenido en este país después de publicar fotos de la manifestación que acontecía en esos momentos, y acto seguido tuiteó un mensaje diciendo “arrestado” con el que alertó a amigos y autoridades sobre su encarcelación.

El día 23 de abril de 2008 dio un gran paso ya que la plataforma llegó a Japón con su interfaz completamente traducida al idioma de este territorio añadiendo la presencia de publicidad.

Posteriormente, en el mes de mayo, la compañía recibió la primera inversión fuerte, una cantidad, que no fue confirmada pero, se estima que fue entre uno y cinco millones de dólares; en el siguiente mes, Jeff Bezos, fundador de Amazon y Bijan Sabet de Spark Capital entran al equipo inversor con 22 millones. Para julio, Twitter dio un gran paso al comprar la empresa

Summize, sobre la cual construyó un sistema de búsquedas completo. En agosto, Chris Messina crea la etiqueta acompañada del símbolo hash (#) seguida de una palabra o varias de forma concatenada mejor conocida como *hashtag*, funcionalidad que Twitter adoptaría de forma oficial, la inclusión fue de gran importancia ya que permitió a los usuarios categorizar y poder dar seguimiento a temas de su interés. Twitter sufrió el primer gran rediseño en septiembre. Para octubre de 2008, hubo un cambio en la dirección ejecutiva, Dorsey prefirió quedarse como presidente del consejo de administración, mientras que Williams pasó a ser el nuevo director ejecutivo; Dorsey en estos momentos, ante un gran crecimiento en la cantidad de usuarios prefirió buscar una estabilidad en el funcionamiento del servicio, dejando de lado la búsqueda en la obtención de más ingresos.

A finales de año en el mes de noviembre, el uso de la plataforma era de forma masiva y alcanzan los mil millones de tweets publicados; de los 400 mil tweets enviados por trimestre en 2007, este año cerró con 100 millones de tweets por trimestre.

En enero del 2009, hubo otro acontecimiento que se hizo viral, así como el estudiante de periodismo preso en 2007. Hubo un accidente de avión en el Río Hudson y la primera imagen, que le dio la vuelta al mundo, se publicó en Twitter rectificando su eficiencia al informar en tiempo real.

En abril, una de las celebridades más influyentes en Estados Unidos, Oprah Winfrey, abrió en su programa en vivo su cuenta en la plataforma siendo de las primeras celebridades que hacían esto; el acercamiento con los famosos a través de Twitter fue parte de una estrategia para ser un servicio más masivo. En este mismo mes, se lanzó la funcionalidad de los Trending Topics (TT) o temas del momento, para poder estar enterados de lo que se más se compartía en tiempo real.

En mayo, hubo otra detención por un tweet, un joven guatemalteco animó para que todos sacaran su dinero del banco Banrural y al poco tiempo las autoridades lo detuvieron acusándolo de generar pánico financiero, acto seguido se realizó una campaña para pedir su liberación, esto demostró el peso que tenía Twitter en la sociedad. El siguiente mes, acontecieron protestas una contra el régimen iraní y casi de forma simultánea el golpe de estado en Honduras, en ambos casos, Twitter se convirtió en la principal fuente de información, esto con la ayuda de *hashtags*.

En septiembre, hasta ese momento recibieron la mayor inversión por un monto de 100 millones de dólares por parte de Insight Venture Partners y T. Rowe Price. Para octubre, Twitter superó la marca de 5 mil millones de tweets. Bing, motor de búsqueda, en sus resultados empezó a

mostrar tweets en tiempo real, así mismo Google firmó un acuerdo para la misma funcionalidad. En los 2 últimos meses del año, se lanzó la versión en español, añaden la funcionalidad de actualización automática en el *timeline*, interfaz donde se muestran las menciones de las cuentas que sigues, y oficializan el *retweet*.

A inicios del 2010, Haití fue golpeado por un terremoto y Twitter volvió a ser protagonista al mantener una comunicación sobre la información en tiempo real y en apoyo a promover la ayuda a los damnificados. En los días posteriores se presentó la funcionalidad de Local TT, en donde la diferencia entre el ya existente es la localidad en la que el algoritmo muestra la información. En febrero Yahoo!, siguió la línea de Google y Bing para incluir tweets en sus resultados de búsqueda y mostrarlos en tiempo real. En este mismo mes, Chile fue golpeado por un terremoto con una magnitud de 8.8, y como venía ocurriendo, Twitter se volvió a convertir en uno de los principales canales de comunicación ya que las líneas telefónicas estuvieron cortadas y saturadas.

Para abril de 2010, se abrían 300,000 cuentas de Twitter por día, y al mes recibía 180 millones de usuarios por mes. Twitter anunció los “Promoted Tweets”, un servicio de pago para las compañías que quisieran aparecer en los resultados de búsqueda en la plataforma. Compañías como Sony Pictures, Red Bull, Best Buy y Starbucks fueron algunas de las primeras en contratar este servicio. En junio, se lanzó Twitter Places, un servicio de geolocalización para que los usuarios que te sigan puedan tener conocimiento en qué lugar te encuentras.

En este año, en agosto recibió la mayor inversión en la historia, un monto de 800 millones de dólares por parte de la empresa Digital Sky Technology, dejando a la compañía valuada en 8.4 mil millones de dólares. Para septiembre se presentó el rediseño más completo con la opción de ver videos, fotos y otros vínculos sin tener que salir de la web principal, así mismo se llegó a 145 millones de usuarios.

A pesar de estas nuevas estrategias y que sus ingresos anuales fueron de 45 millones de dólares, la mayor parte de este año Twitter estuvo operando con pérdidas, y en ese momento, la estimación para el 2011 mostraban ganancias entre los 100 y 110 millones de dólares.

En el mes de octubre, Williams anunciaba que dejaría el cargo de director para ser parte del consejo de administración y dedicarse a crear una nueva *startup*, dejando a Dick Costolo, el director general de operaciones en su lugar.

En enero del 2011, el pueblo egipcio se levanta contra sus dirigentes y Twitter tomó parte del movimiento como pieza importante para la organización de las revueltas, dichas

comunicaciones llegaron a tal grado que Egipto decidió bloquear la plataforma. Twitter fue señalado como un elemento fundamental en la difusión y comunicación sobre cambios políticos y sociales, siendo la plataforma de divulgación sobre los informes diplomáticos por parte de WikiLeaks, por lo que se mostró como una herramienta insustituible para conectar ciudadanos tanto de forma local como global.

Jack Dorsey volvió en marzo como director ejecutivo encargado del desarrollo del producto, dentro de Twitter vieron este movimiento como una necesidad ante la idea de volver a los orígenes de la visión del creador. Este mismo mes Twitter volvió a tener fuerza de comunicación por un terremoto en Japón.

Para mayo, en España empezó una protesta popular que se convirtió en el movimiento ciudadano 15M donde Twitter volvió a ser la principal herramienta informativa. La muerte de Bin Laden fue una de las noticias más retuiteadas en la historia y apareció en primera ocasión a través de Twitter.

En junio se realizó una integración con Firefox, se mejoró el buscador, se creó un sistema propio para compartir imágenes, así como reducción de enlaces propios. Se informó que se habían alcanzado los 200 millones de tweets al día, y al siguientes mes, mostraron el dato de la existencia de un millón de aplicaciones relacionadas con Twitter.

El 25 enero de 2011 se realizaron manifestaciones como protesta sobre la brutalidad policiaca, altas tasas de desempleo, corrupción, falta de aumento del salario mínimo, carencia de vivienda y alimentación, por el gran impacto que se tuvo en esta plataforma, el gobierno bloqueó el servicio para disminuir el impacto y conocimiento al exterior sobre este tema.

En septiembre de 2011 anunciaron que contaban con 100 millones de usuarios activos, así como un sistema de estadísticas propio y de pago.

En el 2012, se realizó un análisis de las 383 millones de cuentas creadas antes de este año para conocer el mayor número de usuarios por país: en Estados Unidos (107.7 millones), Brasil (33.3 millones), Japón (29.9 millones), Reino Unido (23 millones), Indonesia (19 millones), India (12 millones), México (10.5 millones), Filipinas (8 millones), España (7.9 millones) y Canadá (7.5 millones).

En el mes de septiembre del 2013, Twitter anunció que había iniciado los trámites para su salida de la bolsa.

A mediados del 2014, se añade la posibilidad de ingresar en los mensajes los emojis, así mismo, para junio se celebró el mundial de futbol en Brasil donde se habilitaron diseños

especiales de emojis, llamados “hasflags” que cumplen con la función de añadir una imagen alusiva a un evento teniendo la función del *hashtag*.

Para el 2014, Twitter cerró el año con un total de 300 millones de usuarios activos.

En marzo de 2015, se lanza la aplicación Periscope con la funcionalidad de poder emitir video en tiempo real. En abril, Turquía bloqueó Twitter por revelar fotografías del fiscal de Estambul amenazado con una pistola por partidarios extremistas. En junio, se anunció la eliminación de la restricción sobre los caracteres en los mensajes privados, quedando en un total de 10,000 caracteres.

En junio de 2016, se implementó la función de añadir *stickers*, así como los momentos, que consisten en una pestaña en la aplicación que permite ver los tweets con base en una relevancia personalizada.

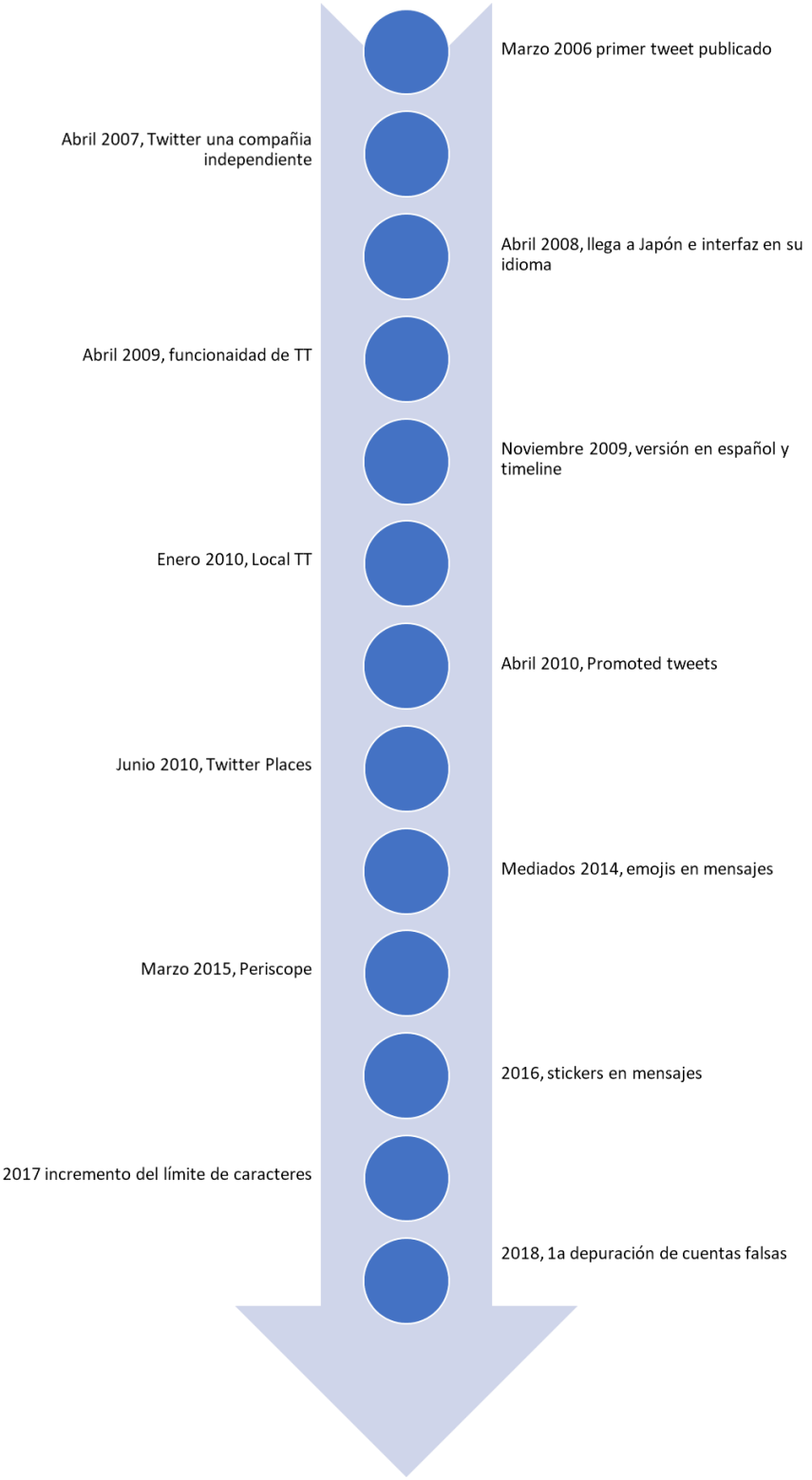
En octubre de 2017, los servicios de inteligencia estadounidense acusaron a medios de comunicación rusos de haber influido en las elecciones presidenciales, por lo que prohibió la publicidad a la televisora RT y Sputnik una agencia de noticias rusas. En noviembre, se incrementó el límite de caracteres pasando de 140 a 280.

En julio de 2018, se realizó la primera limpieza de cuentas falsas con el fin de restablecer la transparencia de la red para evitar que el uso fuera de *bots*.

En la primera semana de mayo de 2019, la plataforma fue parte de un impacto en los mercados bursátiles, la pérdida fue de 13 billones de dólares tras una serie de tweets de la cuenta del presidente Donald Trump creando nerviosismo en el mercado ante una guerra comercial con China; lo que rectifica la relevancia que tiene la plataforma en épocas actuales. Twitter tiene una capitalización aproximada de 28 mil millones de dólares.

Dentro de este universo, además de la existencia de cuentas personales o de celebridades, existen cuentas sobre marcas, noticieros y servicios brindados; con el fin de mantener a sus seguidores informados, es esencial mantener una interacción continua y de interés para que el funcionamiento de la misma sea óptimo y cumpla con su objetivo.

Ilustración 2 - Línea del tiempo Twitter



1.3 Historia del metro de la Ciudad de México

El Sistema de Transporte Colectivo Metro (STC) se creó el 19 de abril de 1967, con el decreto de su creación en el Diario Oficial de la Federación (DOF), el STC asume la responsabilidad para ofrecer al público el servicio de transporte masivo en la capital del país, así como de las funciones de imagen, diseño, planeación, construcción, operación y mantenimiento de la infraestructura de transporte. El STC nace con la tarea inmediata, la construcción de las tres primeras líneas del metro.

Para la creación de éstas líneas, se tomaron en cuenta cuatro componentes técnicos para poder brindar viabilidad en las mismas:

1. Obra civil: dado el hundimiento general del Valle de México, un estudio sobre las condiciones del subsuelo.
2. Factibilidad económica y financiera: debido al alto costo de las nuevas tecnologías.
3. Trazo de las líneas: en función de la demanda de los futuros usuarios.
4. Trenes: cantidad de unidades que deberán estar en servicio para atender el volumen de viajes proyectados.

A cada una de estas variables, por la previsión de la empresa ICA, desde 1958 ya estaba haciendo la entrega del anteproyecto, lo que le valió que STC, como organismo público descentralizado, los contrató en 1967 para hacerse cargo de la construcción del nuevo sistema de transporte.

ICA desarrolló el proyecto conceptual asesorado por empresas privadas de Francia y por el metro de París. Posteriormente llevó a cabo el proyecto básico que consistía en realizar estudios de mayor profundidad, como pruebas especializadas de geología y mecánica de suelos, necesarios para poder elegir la tecnología idónea en la construcción de la obra. ICA propuso el esquema de financiamiento que fue aceptado por el gobierno federal, que por parte de Francia fueron ofrecidas líneas de crédito.

Para la construcción de la primera etapa del metro se integraron equipos de trabajo multidisciplinarios, en los que participaron ingenieros geólogos, de mecánica de suelos, civiles, químicos, hidráulicos y sanitarios; mecánicos, electricistas, arqueólogos, biólogos, arquitectos, especialistas en ventilación, estadística, computación, en tráfico y tránsito, contadores, economistas, abogados y obreros.

En el caso de las líneas 1, 2 y 3, el gobierno federal asignó de forma directa a ICA la realización de los proyectos ejecutivos de la obra civil y electromecánica, así como su construcción. De la misma forma se asignó que empresas francesas serían los proveedores de los trenes.

Se generó una sinergia para la integración de equipos de trabajo multidisciplinarios que llegaron a conjuntar a cuatro mil técnicos y tres mil administradores, para que en conjunto con los 48 mil obreros se pudiera construir un kilómetro por mes.

Para la segunda etapa de la construcción de la red del metro, hubo una modificación sobre la relación entre ICA y el STC, la cual había permitido el mandato de construir, operar y explotar la infraestructura y el servicio del sistema de trenes; a partir de 1977 el STC dejaría de ser responsable de la construcción. Para el 15 de enero de 1978, la Comisión de Vialidad y Transporte Urbano (COVITUR), fue creada quedando como responsable de planear y construir, así como de la adquisición de trenes. Gran parte de la gente especializada de la primera etapa se conservó ya que los altos funcionarios de COVITUR provenían de ICA. COVITUR se hizo cargo de la preparación del Plan Rector de Transporte en el Distrito Federal para planear los subsistemas de infraestructura con una perspectiva de todos los modos de transporte así como de desarrollo urbano y cuidado del medio ambiente.

En esta etapa se pueden identificar dos fases:

- La primera corresponde a la ampliación de la línea 3 hacía el norte y el sur.
- La segunda fase, se inició con la construcción de las líneas 4 y 5.

Las obras estuvieron a cargo de la empresa Ingeniería de Sistemas de Transporte Metropolitano del consorcio ICA.

La relevancia de las obras fueron necesarias para brindar el servicio de metro a la población del Distrito Federal que le fijó a COVITUR la función de planeación y diseño, esta atribución es acompañada a una asignación en gran proporción del presupuesto que fue destinado a obras públicas en el Distrito Federal, en 1978 se destinó el 20 por ciento, mientras que en 1982 fue del 60 por ciento.

El desarrollo del sistema de líneas de metro constituyó la columna vertebral de la planeación del transporte del Distrito Federal. En el caso de esta etapa, la empresa del gobierno federal Constructora Nacional de Carros de Ferrocarril (CONCARRIL), realizó las primeras entregas de vagones para el metro en diciembre de 1975, y para abril de 1985 ya había hecho la entrega de 1,061 vagones.

A partir de esta etapa, se institucionaliza una división de tareas para realizar nuevas líneas y ampliar las existentes.

Las líneas 4 y 5, construidas en la segunda etapa han adolecido desde el inicio de su funcionamiento una deficiente contribución a la captación de viajes, resultando en que la capacidad del sistema se vea reducido, es decir, se va ampliando el número de líneas y al mismo tiempo el sistema como un único servicio, tiende a captar cada vez menos viajes en proporción a su capacidad total. El problema detrás de estas deficiencias, son derivadas de fallas en el diseño del Plan Maestro del Metro en los supuestos en que se sustenta el trazo de las líneas. El supuesto de la construcción de las líneas 4 y 5, resuelve áreas de viviendas de la clase obrera urbana creciente con la industrialización de la Ciudad de México. La demanda masiva de viajes surgió desde la ubicación de la fuerza de trabajo de origen campesino que necesitaba trasladarse hacía los grandes centros de servicio de la ciudad. La problemática recae en las líneas de convergencia de viajes en la zona oriente-norponiente, lo que ha generado que las líneas sean subutilizadas.

La línea 4 se construyó como viaducto elevado por el bajo número de construcciones altas en la zona y consta de 10 estaciones, ocho elevadas y dos en superficie.

La línea 5 se construyó en 3 tramos, la edificación de esta línea se realizó en dos formas, a nivel de superficie y subterránea.

La tercera etapa constó de la ampliación de las líneas 1, 2 y 3; así mismo se inició la construcción de dos líneas nuevas, la 6 y 7:

- La línea 1 se amplía una estación al oriente.
- La línea 2 al poniente quedando al límite con el Estado de México.
- La línea 3 se prolonga hacia el sur.

La ampliación de estas líneas concluye con las estaciones que a la fecha están en funcionamiento.

A la línea 6 se le dio una construcción combinada, subterránea y superficial, esta línea corre de oriente a poniente en la zona norte de la ciudad.

La línea 7 rodea el Valle de México por el poniente, la construcción que se ocupó fue de túnel profundo.

La cuarta etapa estuvo permeada por una reforma administrativa aprobada a nivel federal que inició con la presentación del Plan Nacional de Desarrollo (PND) en 1983, simultáneamente

en el caso del Distrito Federal se reformó la Ley Orgánica del Departamento del Distrito Federal, vigente desde 1970.

Se creó la Coordinación General de Transporte (CGT), donde fueron agrupados los organismos como las funciones sectoriales correspondientes para llevar a cabo un planeación del sistema de transporte y vialidad en el Distrito Federal; su misión nominal fue el apoyo en la evaluación del desarrollo de las entidades agrupadas en el subsector correspondiente al programa sectorial y los demás programas pertinentes.

La CGT adquirió las facultades del STC Metro de:

- Coordinar los proyectos y programas de construcción de las obras de ampliación del Sistema de Transporte Colectivo.
- Participar en la elaboración de los programas institucionales de las entidades paraestatales.

Esta nueva reforma a la administración pública realizó la trasferencia de la responsabilidad y los recursos del suministro de los trenes de COVITUR a STC, con la intención de generar una mayor participación entre el proveedor y el organismo que opera las unidades. La CGT asumió las funciones de planeación y coordinación, que anteriormente eran competencia de COVITUR, y el metro recupera la atribución de conseguir los trenes.

La CGT se transformó, a diez años de su creación, en la Secretaría de Transporte y Vialidad (SETRAVI) en 1994.

Se realizaron ampliaciones de la línea 6 y 7, y se dio inicio a la construcción de la línea 9. La ampliación de la línea 6 fue inaugurada el 8 de julio de 1988, mientras que la línea 7 fue el 29 de noviembre de 1988.

La construcción de la línea 9 se realizó en 2 etapas, la primera concluida el 26 de agosto de 1987, y la segunda fue inaugurada un año más tarde. Esta línea tiene un trazo paralelo a la línea 1 con el propósito de descongestionarla en las horas con mayor afluencia. La construcción fue subterránea.

En la quinta etapa, se dio la primera extensión de la red del metro al Estado de México y fue con la construcción de la línea A, en ésta se optó cambiar de trenes de ruedas férreas en lugar de neumáticos con el motivo de reducción de costos de construcción y mantenimiento. Se edificó un puesto de control y talleres exclusivos para esta línea que se inauguraron el 12 de agosto de 1991.

El trazo inicial de la línea 8 también fue modificado, por el cruce en la zona del Centro Histórico, se pondrían en peligro las estructuras de las edificaciones coloniales y daños en la ciudad prehispánica. Con esta modificación, la línea 8 se inauguró el 20 de julio de 1994

La sexta etapa dio origen a la línea B, de 20 km de recorrido mediante 21 estaciones, inició su construcción en octubre de 1994 durante el sexenio del presidente Carlos Salinas, cuya inauguración estaba programada para 1997, cabe aclarar que, esta línea tendría 2 inauguraciones, la primera en diciembre de 1999 a finales del sexenio del presidente Ernesto Zedillo donde pusieron en funcionamiento 13 estaciones, y la segunda el 30 de noviembre de 2000 poco antes del término del periodo de gobierno de la jefa de Gobierno del Distrito Federal, Rosario Robles quien inauguró las 8 estaciones restantes. El diseño, construcción y equipamiento de trenes tardó siete años, el lapso de la entrega y recepción que la Secretaría de Obras hizo de toda la infraestructura al STC Metro hasta 2002, se realizó en un lapso de dos años más; por lo que la construcción de la línea B fue de 9 años.

El cambio de régimen político produce un ajuste legal y organizativo con la Ley Orgánica de la Administración Pública del Distrito Federal de 1998, misma que afecta con una reestructuración del sector de transporte. En este proceso, la CGT dejó de abordar los asuntos del transporte y pasaron a ser materia de la SETRAVI en 1994. La COVITUR se transformaría en la Dirección General de Construcción de Obras del Sistema de Transporte Colectivo, adscrita a la Secretaría de Obras.

En el caso de la línea B, la elección de los trenes que se hacían en función de la empresa paraestatal CONCARRIL fue eliminada, por lo que el responsable para determinar sus especificaciones, licitarlos y comprarlos sería el STC Metro. La Secretaría de la Contraloría y Desarrollo Administrativo (SECODAM) aparece para regular la licitación internacional con la que el STC Metro pretendía adquirir 252 vagones y 28 trenes, para la línea B. En el mismo proceso de implementación surgió el riesgo de incumplimiento con la fecha de entrega de la línea B, esto se debió a una serie de contingencias que el metro no pudo controlar relacionadas con la inversión de 400 millones de dólares para la compra y mantenimiento de los 28 trenes.

El Estado de México condicionará la decisión de la construcción de la línea B ya que le corresponde acompañar la edificación de la obra para que puedan contrarrestar el impacto urbano, ambiental y de movilidad en el territorio en el que la obra ocuparía el Estado de México.

Los problemas del segundo tramo de la línea B se complican por el nuevo marco normativo; en la etapa de decisión de política pública, la Cámara de Diputados como actor federal que tiene la facultad de fijar el endeudamiento, vital para sacar adelante el financiamiento de la

parte faltante de la línea B, solamente se aprobó 1,700 millones de pesos de los 7,500 solicitados, razón principal del retraso existente. En cuanto a la compra de los trenes, a través del consejo de administración, se resuelve hacer frente a la invalidación de la licitación que SECODAM realizó y canceló la compra de los trenes, provisionalmente operaría con trenes rehabilitados. Esto provoca que la línea B empiece a funcionar de forma deficiente, los trenes rehabilitados no cuentan con la tecnología moderna para poder programar y ajustar los tiempos de recorridos, por lo que las quejas por parte de los usuarios se vuelven recurrentes.

ICA, que en 1967 hizo posible la construcción de las líneas 1, 2 y 3 mediante un sistema de flotación de los túneles en un terreno fangoso, se había convertido en una empresa sin capacidad de cumplir con un contrato parcial, la causa fue que en 1994 se encontraba al borde de la bancarrota, sin maquinaria ni cartera de proyectos. En 2003, regresó a cotizar a la bolsa con un nuevo modelo de negocios.

En diciembre de 2006, se anunció la construcción de una nueva línea del metro con el fin de atender la demanda al sur de la Ciudad de México, mediante una encuesta a la ciudadanía se propusieron 2 opciones de posibles rutas, Iztapalapa-Acoxpa e Iztapalapa-Tláhuac, y el 8 de agosto de 2007 se presentó el proyecto oficial de Iztapalapa-Tláhuac con el nombre de Línea 12.

El objetivo de la construcción de esta línea es la de poder brindar un servicio de transporte masivo de forma rápida, económica a los habitantes de siete delegaciones, así como poder proporcionar conectividad con cuatro líneas del metro en la zona sur.

Para esta línea se propone el color oro como identidad por la celebración del Bicentenario de la Independencia de México y el Centenario de la Revolución Mexicana, en esta línea se usaría trenes de rieles, fue inaugurada el 30 de octubre del 2012 en el sexenio del presidente Felipe Calderón Hinojosa.

Sin embargo, al ser la línea con la construcción más reciente, por motivos de malos cálculos en la construcción de las vías y las dimensiones de las ruedas del metro, no se previó un desgaste prematuro en los rieles ni problemas en 26 curvas, lo que derivó en fractura de durmientes de concreto, fractura de fijaciones nabla de rieles, así como problemas en contra-rieles, aparatos de cambio de vía y aparatos de dilatación. Incluso antes de su inauguración, señalaron que los daños eran evidentes y tuvieron que sustituir, semanas antes, rieles entre las estaciones Zapotitlán y Nopalera.

Tabla 1 - Línea del tiempo de estaciones del metro

Línea	Tramo	Fecha inauguración	Estaciones	Kilómetros
1	Zaragoza-Chapultepec	04/09/1969	16	12.660
1	Chapultepec-Juanacatlán	11/04/1970	1	1.046
1	Juanacatlán-Tacubaya	20/11/1970	1	1.140
1	Tacubaya-Observatorio	10/06/1972	1	1.705
1	Zaragoza-Pantitlán	22/08/1984	1	2.277
2	Taxqueña-Pino Suárez	01/08/1970	11	11.321
2	Pino Suárez-Tacuba	14/09/1970	11	8.101
2	Tacuba-Cuatro Caminos	22/08/1984	2	4.009
3	Tlatelolco-Hospital General	20/11/1970	7	5.441
3	Tlatelolco-La Raza	25/08/1978	1	1.389
3	La Raza-Indios Verdes	01/12/1979	3	4.901
3	Hospital General-Centro Médico	07/06/1980	1	0.823
3	Centro Médico-Zapata	25/08/1980	4	4.504
3	Zapata-Universidad	30/08/1983	5	6.551
4	Martín Carrera-Candelaria	29/08/1981	7	7.499
4	Candelaria-Santa Anita	26/05/1982	3	3.248
5	Consulado-Pantitlán	19/12/1981	7	9.154
5	La Raza-Consulado	01/07/1982	3	3.088
5	La Raza-Politécnico	30/08/1982	3	3.433
6	El Rosario-Instituto del Petróleo	21/12/1983	7	9.264
6	Instituto del Petróleo-Martín Carrera	08/07/1986	4	4.683
7	Tacuba-Auditorio	20/12/1984	4	5.424
7	Auditorio-Tacubaya	22/08/1985	2	2.730
7	Tacubaya-Barranca del Muerto	19/12/1985	2	5.040
7	Tacuba-El Rosario	29/11/1988	4	5.590
8	Garibaldi-Constitución de 1917	20/07/1994	19	20.078
9	Pantitlán-Centro Médico	26/08/1987	9	11.669
9	Centro Médico-Tacubaya	29/08/1988	3	3.706
A	Pantitlán-La Paz	12/08/1991	10	17.192
B	Villa de Aragón-Buenavista	15/12/1999	13	12.139
B	Ciudad Azteca-Nezahualcóyotl	30/11/2000	8	11.583
12	Mixcoac-Tláhuac	30/10/2012	20	23.722

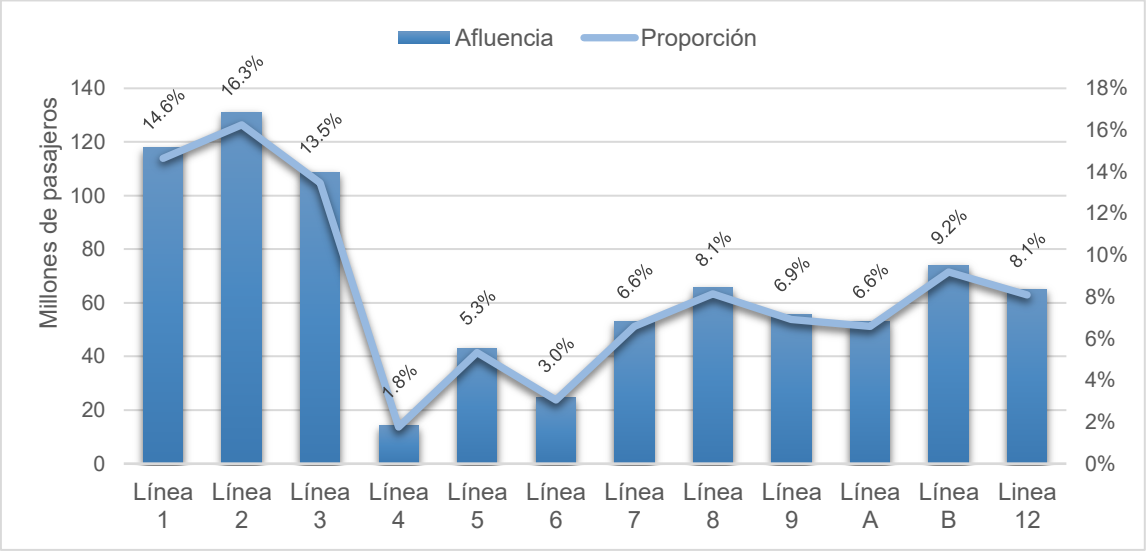
El STC Metro consta de un total de 225.11 km con un total de 195 (163 únicas por conexiones) estaciones distribuidas en la siguiente tabla:

Tabla 2 - Distancia líneas del metro

Línea	Estaciones	Kilómetros
1	20	18.828
2	24	23.431
3	21	23.609
4	10	10.747
5	13	15.675
6	11	13.947
7	14	18.784
8	19	20.078
9	12	15.375
A	10	17.192
B	21	23.722
12	20	23.722
Total	**195**	**225.11**

La afluencia en el periodo de enero a junio del 2019, fue de 805.3 millones de pasajeros, el 44.36% de los viajes se realizan entre las líneas 1, 2 y 3. Cabe señalar que esta información contempla los ingresos en la línea a la que ingresan sin incluir a los que realizan un transbordo.

Gráfico 1 – Afluencia de las estaciones del metro



El Sistema de Transporte Colectivo Metro a partir de mayo del 2010, tiene bajo su responsabilidad una cuenta de Twitter “@MetroCDMX” que a septiembre del 2019 cuenta con

un total de 1.8 millones de seguidores. Esta cuenta tiene como objetivo mantener informados a los usuarios de este transporte, el estatus de los tiempos de espera, así como la afluencia presentada e incidentes generados en toda su red; dicha información se presenta mediante una infografía que en su mayoría se genera cada hora del día. Asimismo, cuando sucede un incidente, desde retiro de trenes, accidentes en vías, lluvias que afectan el avance, fallas mecánicas o de energía eléctrica, o se tiene un evento cultural; la cuenta sube un tweet mencionando lo sucedido, informando a la ciudadanía la razón por la que el servicio no está siendo el adecuado.

La cuenta también mantiene una comunicación personal para aquellos seguidores que tengan un comentario o queja sobre el servicio. La ciudadanía al querer siempre obtener un servicio eficiente, tiene quejas la gran acumulación de gente en las estaciones, así como la falta de trenes y tiempo en que tarda en pasar una unidad y el avance de los trenes, la venta de productos diversos dentro de los vagones, lo cual es una actividad ilegal, la limpieza y ventilación de las unidades.

Este tipo de comentarios se pueden dar a conocer mediante Twitter, si se quiere tener una idea más completa sobre el volumen de comentarios recibidos de las distintas quejas, es posible realizar una clasificación de las menciones para poder extraer información valiosa y generar un direccionamiento eficiente hacia los focos rojos de las deficiencias que existen en el servicio.

Capítulo 2. Algoritmos de Clasificación de Textos

2.1 Conocimiento y preparación del texto

Antes de poder aplicar cualquier algoritmo de minería de textos es imprescindible realizar un conjunto de actividades llamadas normalización de texto, en donde las expresiones cotidianas juegan un lugar importante. La normalización de textos se puede definir como la conversión del texto a una forma más estandarizada y conveniente para su explotación.

Por lo regular, al querer individualizar las palabras, en un texto están separadas por un espacio en blanco, pero no siempre esto resulta funcional, hay palabras que son compuestas, por ejemplo *Ciudad de México* o *Nuevo León*, por lo cual no se pueden manipular de la misma forma, se tienen que juntar y tomar como una única palabra.

Otro proceso dentro de la normalización del texto es la de comparar palabras que tienen la misma raíz con el objetivo de reducir la cantidad de términos que existen en el texto, como ejemplo, se puede mencionar a cualquier verbo, suponiendo que el texto contiene las siguientes palabras: *avance*, *avanza* y *avanzas*; éstas tienen su raíz en la palabra *avanzar*, por lo que en vez de contar tres palabras distintas se reducirá a una palabra, con lo cual se podrá hacer un análisis más eficiente.

Dentro de los procesos de normalización, existe una versión más sencilla para encontrar las raíces de las palabras, ésta solo busca truncar, con el propósito de poder eliminar las distintas terminaciones de una palabra.

Uno de los éxitos, sobre la estandarización, menos atribuidos en la informática ha sido el uso de expresiones regulares (ER), un lenguaje para especificar cadenas de búsqueda de texto; este lenguaje es ocupado en cada lenguaje de computadora, así como procesadores de textos y palabras. Es útil para la búsqueda de textos cuando existen patrones, al realizar una, se regresará a todos los documentos que coincidan con el patrón.

La forma más simple de una ER es la secuencia común de caracteres, en donde realizará la búsqueda exactamente de esa cadena diferenciando entre mayúsculas y minúsculas, así como acentos; por lo que se puede mencionar que las ER son sensibles. Siendo recomendable homogeneizar el texto y realizar la eliminación de acentos, así como transformar todo a minúsculas.

Este tipo de búsquedas servirán para encontrar ciertos patrones en las palabras y reducir la cantidad de términos, manipularlos para poder utilizar sinónimos y homogeneizar las palabras

del texto que buscas analizar; así como encontrar y eliminar palabras que no agregan un valor al análisis deseado.

Teniendo esta herramienta, antes de procesar palabras es necesario decidir, dentro de nuestro universo de documentos, cuáles contarán como tal, y dependerá de la aplicación de cada estudio, pero por lo regular están los artículos determinados, indeterminados, preposiciones, conjunciones que son parte importante en la estructura de una oración pero al ser ocupadas en su mayoría, el análisis de éstas no son de importancia a menos que el estudio sea sobre estas figuras de la ortografía.

Hay palabras que una es derivada de otra, ejemplo de esto es la situación sobre singular y plural, la raíz es la misma, aunque son palabras distintas y se tiene que decidir cómo manejarlas para efectos del estudio, por lo regular se reducen a su forma singular.

Las palabras no aparecen de la nada, cada pieza particular de texto estudiada es producida por un recitador o escritor, en un dialecto e idioma específico y un tiempo, lugar y función en particular; es posible que el lenguaje sea el mejor ejemplo de variación. El procesamiento de lenguaje natural (NLP) por sus siglas en inglés, son los algoritmos más útiles. Las herramientas de NLP por lo general son desarrolladas para los idiomas con mayor industria, como el español, inglés, chino, árabe, etc. La mayoría de los idiomas tienen múltiples variedades, como son los dialectos hablados en distintas regiones o diferentes grupos sociales. También es común que se ocupen múltiples idiomas en un único acto comunicativo, este uso es llamado cambio de código; es ocupado de gran manera alrededor del mundo.

Continuando con las variaciones, también está el género, el texto que nuestro algoritmo tiene que procesar puede venir de un entorno noticioso, libros de ficción o no ficción, artículos científicos o religiosos. Incluso la fuente de información puede venir de distintos medios, como conversaciones telefónicas, reuniones de negocios, cámaras ocupadas por policías, entrevistas médicas, redes sociales o transcripciones de programas de televisión o películas, también pueden recibirse de notas médicas, textos legales, así como reformas de los gobiernos.

El texto también puede reflejar características demográficas del escritor o hablante, por ejemplo la edad, género, raza o clase socio-económica; estas características pueden influenciar en las propiedades lingüísticas del texto que se está procesando.

El tiempo también es importante, el lenguaje cambia a través del tiempo y para ciertos idiomas se tiene un gran número de documentos de distintos periodos históricos, al desarrollar un

modelo computacional para procesar el lenguaje es necesario conocer el contexto, el propósito y estar seguros de que el modelo se ajusta a la información.

Antes de querer aplicar cualquier algoritmo de NLP, el texto tiene que ser normalizado. Al menos 3 tareas son aplicadas como parte del proceso de normalización:

- Segmentación e individualización de las palabras contenidas en los documentos.
- Normalización de los formatos de las palabras.
- Segmentación de oraciones contenidas en los documentos.

Una forma de normalización de texto para tareas como reconocimiento y recuperación de información, es transformar todas las palabras a minúscula. Para análisis de sentimientos y tareas de clasificación de textos, extracción de información es útil aplicar esta herramienta.

En la práctica, como la individualización es necesaria antes de aplicar cualquier algoritmo, es indispensable que sea rápida. El método estándar de normalización es ocupar algoritmos determinísticos basados en expresiones regulares compilados en un eficiente código automatizado.

Al reducir las palabras a la raíz de las mismas, tiene un beneficio, este proceso de normalización ayuda a enfrentar el problema de que aparezcan palabras desconocidas al ingresar nueva información. Las palabras desconocidas son relevantes para los sistemas de aprendizaje de máquina.

Otra cosa que posiblemente quisiéramos conocer es el idioma en el que está escrito el texto. Por ejemplo, lo que se escribe en redes sociales puede estar expresado en distintos idiomas, para lo cual será necesario aplicar diferentes tipos de procesos. La tarea de identificación del idioma es el primer paso necesario de aplicar. Tareas relacionadas como determinar el autor o características como género, edad o idioma nativo del mismo en un texto, son aplicaciones de la clasificación de textos también relevantes a las humanidades digitales, ciencias sociales y lingüística forense.

Posterior a la aplicación de estas estrategias, se tendrá un mejor manejo y conocimiento de la base, por lo que se podrán aplicar de mejor forma los algoritmos de clasificación.

2.2 Bases en los algoritmos de clasificación

La clasificación es una operación que asigna a cada elemento, conforme a sus propias características llamadas variables independientes, dentro de las clases específica bajo estudio

en un universo. Un elemento es asignado a una clase con base en sus características usando una fórmula, algoritmo o grupo de reglas, que forman un modelo.

Existen clasificadores probabilísticos que construyen un modelo que cuantifica la relación que existe entre ciertas variables características y una clase o grupo, y esto lo expresa como una probabilidad.

Hay modelos que asignan probabilidades a secuencias y son llamados modelos lingüísticos o LM. El modelo más simple, por la estructura de su construcción, son los n-gram, que es una secuencia de *N* palabras. Un 2-gram es un bigrama, una secuencia de dos palabras, por ejemplo “tiempo esperando”, “sin avanzar”; un 3-gram es un trígrama, una secuencia de 3 palabras, “mucho tiempo esperando”, “minutos sin avanzar”.

Tratando de explicar esto mediante probabilidades, iniciemos encontrando $P(w|h)$, la probabilidad de la palabra w dado h; suponiendo que la historia de h es “su agua es tan transparente que”, y deseamos conocer la probabilidad de que la siguiente palabra w sea “el”, expresándolo de la siguiente manera:

$$P(el|su\ agua\ es\ \tan\ transparente\ que)$$

Una manera de estimar esta probabilidad es mediante el conteo de frecuencia, se tomaría una gran cantidad de documentos, para contar las veces que aparezca “su agua es tan transparente que” e identificar las veces que posteriormente lleva la palabra “el”, por lo que con esto se estaría resolviendo lo siguiente:

$$P(el|su\ agua\ es\ \tan\ tranparente\ que) = \frac{C(su\ agua\ es\ \tan\ tranparente\ que\ el)}{C(su\ agua\ es\ \tan\ tranparente\ que)}$$

Dependiendo del contexto del estudio, se pueden tomar de internet documentos para calcular este número y estimar la probabilidad. Mientras que este método es eficiente en muchos casos, se puede verificar que incluso los documentos en internet no son suficientes para poder encontrar una buena estimación, esto se debe a que el lenguaje es creativo, nuevas oraciones son creadas todo el tiempo y no siempre será posible contar todas.

La clasificación se basa tanto en la inteligencia humana como el de la máquina. Decidir qué letra, palabra o imagen ha sido presentado a nuestros sentidos, reconociendo rostros o voces, ordenando correos electrónicos, asignar calificaciones a las tareas; todos estos son ejemplos de asignación de una categoría a una entrada de información.

Muchos lenguajes con tareas de procesamiento involucra a la clasificación, por ejemplo el análisis de sentimientos, la extracción del sentimiento, la positiva o negativa orientación que un escritor expresa hacia un objeto. La crítica a una película, un libro o un producto en venta

en internet, el consumidor expresa su sentimiento hacia el producto, mientras que una editorial o texto político expresa el sentimiento hacia un candidato o acción política. La extracción de un consumidor o de la gente es relevante desde los campos de mercadotecnia hasta la política.

La versión más simple del análisis de sentimientos es una tarea binaria de clasificación con nivel positivo y negativo, las palabras de las reseñas proporcionan información de relevancia.

La detección de spam es otra importante aplicación comercial, la clasificación binaria de asignar a un correo a uno de los 2 casos, spam o no-spam. Muchas palabras pueden ser ocupadas para realizar esta clasificación. Por ejemplo, uno sospecharía de un correo que contiene frases como “sin ningún costo” o “has sido elegido ganador”.

Con estos sencillos ejemplos es posible observar que para poder realizar una clasificación, hay necesidad de contar con una información previa para agrupar los documentos, y así obtener una clasificación, a este mecanismo entre una base a entrenar y una que clasifica, se le conoce como el aprendizaje supervisado.

2.3 Clasificación mediante el aprendizaje supervisado

Algunos problemas de la minería de datos están dirigidos hacia un objetivo específico, y están representados por una característica en particular llamada etiqueta. De esta forma, estos problemas se vuelven supervisados, la relación entre las características de la información y la etiqueta es aprendida.

Bajo la premisa anterior, la clasificación se vincula al aprendizaje supervisado, una base de datos es ocupada para aprender la estructura de los grupos o etiquetas. Mientras los grupos son aprendidos por un modelo de clasificación, que generalmente tiene similitud en la estructura de las características de las variables. En la clasificación, es primordial que en la base de entrenamiento se provean ejemplos claros de cómo están definidos los grupos; dados los documentos que se ingresen, y con el sustento de esta base, el modelo tratará de reflejar la estructura de los grupos disponibles hacia la base de evaluación.

La mayoría de los algoritmos de clasificación consta de dos fases:

1) Fase de entrenamiento: En esta fase un modelo de entrenamiento es construido. Intuitivamente, esto se puede entender como un resumen de las etiquetas de los grupos.
2) Fase de prueba: En esta fase, el modelo de entrenamiento es usado para determinar la clase, etiqueta o grupo.

Cuando la base de entrenamiento es pequeña, el desempeño de los modelos de clasificación puede ser pobre, en tal caso el modelo puede describir de forma aleatoria las características de la base de entrenamiento, pero puede tener un bajo desempeño a nueva información.

Una de las tareas más antiguas en la clasificación de textos es la asignación de una biblioteca de categorías o temas a un texto. La clasificación es esencial para tareas de un nivel más detallado del documento. El objetivo de la clasificación es tomar una única observación y extraer características útiles para que se clasifiquen las observaciones en un grupo de clases.

Las reglas para hacer una clasificación pueden ser imprecisas, y el cambio continuo de la información, así como que los humanos no son necesariamente buenos para aplicar reglas en la mayoría de los casos, la clasificación del lenguaje es realizado mediante el aprendizaje de máquina supervisado. En el aprendizaje supervisado se tiene un conjunto de observaciones, cada una asociada con la correcta salida de la información; por lo que el objetivo del algoritmo será aprender cómo asociar nuevas observaciones a un correcto resultado.

La selección de las características para cada grupo es el primer punto en cualquier proceso de clasificación. La información puede contener características de relevancia que varía para la predicción de las etiquetas de las clases. Las características irrelevantes pueden dañar la precisión del modelo de clasificación, el objetivo de la selección de éstas es identificar las características más informativas. Existen tres métodos ocupados para la selección de características en la clasificación:

1) Filtro: Este criterio es ocupado para excluir las características irrelevantes.
2) Envoltura: Se asume que un algoritmo de clasificación puede evaluar qué tan bien se desempeña el algoritmo con un conjunto específico de características, por lo que la búsqueda se queda envuelta dentro del conjunto, para determinar el grupo de características.
3) Incorporación: La solución a un modelo de clasificación regularmente contiene pistas útiles sobre las características más relevantes, las cuales están aisladas y el clasificador es reentrenado.

En el modelo de filtros, una característica o un conjunto de características es evaluado con el uso de criterio de discriminación sensitiva. La ventaja de evaluar a un grupo de características al mismo tiempo es que las redundancias son bien contabilizadas. Considerando que se tengan dos características que están relacionadas una con otra, es posible que una se pueda predecir con la otra. En la práctica, los métodos de selección de las características evalúan de forma independiente cada uno y seleccionan las más discriminativas.

En los modelos de envoltura, la característica es definir de forma iterativa el conjunto de características, que se lleva a cabo en dos pasos:

1) Crea un conjunto aumentado de características, añadiendo una o más al conjunto actual.
2) Usa un algoritmo de clasificación para evaluar la precisión de las características y poder aceptar o rechazar el conjunto aumentado.

La precisión ocupada en el segundo paso funcionará para decidir si se acepta o revierte el grupo de características. Este proceso se continúa aplicando hasta que no haya mejoramiento en el conjunto de características.

En los modelos de incorporación, la idea principal es que muchos cálculos de clasificación proveen pistas importantes sobre las características más relevantes, el aprendizaje sobre éstas son insertadas a la solución sobre la clasificación.

2.4 Metodologías de clasificación

Existen diferentes metodologías para realizar una clasificación, entre ellas está la clasificación mediante árboles de decisión, procesos modelados con el uso de un conjunto de decisiones jerárquicas sobre las características y están arregladas en estructura de árbol. La decisión en un nodo en particular, referida como un punto de criterio de división, es una condición de una o más características en la base de entrenamiento, este criterio divide la base en dos o más partes. El objetivo es identificar un criterio de división para que el nivel de generalidad de las variables, en cada una de las ramas, sean reducidas lo más posible. Cada nodo en el árbol de decisión representa un subconjunto de información, definida por la combinación del criterio de división de los nodos anteriores.

Dentro de esta metodología, existe el árbol de decisión con algoritmo de inducción que contiene dos tipos de nodos, nodos internos (ni) y nodos de hoja (nh), cada *nh* es etiquetado con la clase dominante de ese nodo, un *ni* es el nodo raíz que corresponde al universo de características. La decisión de un árbol inductivo empieza con la base de entrenamiento en el nodo raíz y recursivamente hace particiones de la información en niveles más pequeños divididos por cada uno de los criterios. Eventualmente, el algoritmo de decisiones se detendrá dependiendo del criterio para concluirlo.

Hay árboles de decisión con criterio de separación y tiene como objetivo maximizar la separación de las diferentes clases a través de los nodos. El diseño del criterio de separación depende de la naturaleza de los atributos:

1) Binario: solo un tipo de separación es posible, cada rama corresponde a uno de los valores binarios.
2) Categóricos: si el atributo contiene diferentes valores, hay varias maneras de separación, este tipo no es recomendable cuando los valores son muy grandes.
3) Numérico: si contiene un valor pequeño de valores, es posible crear una separación con ese número de valores, pero si el valor es grande, se ocupa la separación binaria.

Adicional a los árboles de decisión, existen los clasificadores basados en reglas que usan la fórmula "si-entonces" para poder enlazar el antecedente con el consecuente. La regla se expresa de la siguiente forma:

$$SI\ Condición\ ENTONCES\ Conclusión$$

La condición en la parte izquierda de la regla, antecedente, contiene una variedad de operadores lógicos como $<$, $\leq$, $>$, $\geq$, $=$ o $\in$, que son aplicados a las características de las variables. El lado derecho de la regla se refiere a la consecuencia y contiene la clase de la variable; las reglas son generadas de la base de entrenamiento. En general, el antecedente puede ser cualquier condición arbitraria, posteriormente estas reglas son ocupadas para la clasificación. La regla se dice que cubrió la instancia de entrenamiento cuando la condición en el antecedente encaja de forma correcta para cumplir la condición.

La fase de entrenamiento de un algoritmo basado en reglas crea un conjunto de reglas, en la fase de clasificación se descubren todas las reglas que son activadas en la fase de entrenamiento, una regla se dice que es activada cuando la condición lógica en el antecedente se satisface.

Las reglas deben satisfacer una o más de las siguientes propiedades:

1) Reglas mutuamente exclusivas: cada regla cubre una parte de la separación de la información, las reglas generadas de un árbol de decisión satisface esta propiedad.
2) Reglas exhaustivas: la base es cubierta por al menos una regla, y para cada nodo de decisión se aplica mínimo una regla, así mismo los árboles de decisión satisfacen esta propiedad.

En los casos cuando el conjunto de reglas no sean mutuamente exclusivas, existen problemas en la aplicación de las reglas y pueden ser resueltas con alguno de los siguientes puntos:

1) Ordenamiento de reglas: las reglas son ordenadas por prioridad, la etiqueta de la clase consecuente de la activación superior es considerada como la más relevante.

2) Reglas sin orden: no existe prioridad en la ordenación de las reglas, la etiqueta de la clase dominante por encima de todas las reglas aplicadas, tal acercamiento puede ser más robusto porque no es sensible a la elección de una única regla.

Dentro de los clasificadores basados en reglas, existen los clasificadores asociativos, basados en patrones de asociación y existen muchos algoritmos eficientes, por ejemplo el algoritmo de a priori. La característica principal de las reglas de asociación basada en clases es que son extraídas de la misma forma que las reglas de asociación, excepto las que tengan una única clase en la parte consecuente. La estrategia básica para un clasificador asociativo sería:

1) Extrae todas las reglas de asociación basadas, con un nivel de soporte y confianza mínimo.
2) Para el momento de una prueba, usar las reglas extraídas para la clasificación.

Existe una variedad de opciones para la implementación de ambos pasos, una forma simple de implementar el primer paso sería extraer todas las reglas de asociación, y después filtrar sólo las reglas en donde la parte consecuente corresponda a una única clase, no obstante este enfoque no es eficiente porque genera reglas sin obtener un resultado en la parte consecuente.

La clasificación basada en asociaciones usa una modificación del método a priori, algoritmo que ocupa la propiedad de cierre hacía atrás para acortar el espacio de búsqueda de la parte consecuente. El primer paso es generar una regla por elementos, estos son creados y corresponden a la combinación de elementos y atributos de las clases, posteriormente los elementos de las reglas son extendidos usando un procesamiento del estilo tradicional a priori. Otra modificación es cuando los patrones son generados de reglas con un 100% de confianza, estas reglas no se extienden en orden para conservar una mayor generalidad en el conjunto de reglas, este amplio enfoque puede ser ocupado en conjunto con casi cualquier algoritmo de árbol.

El segundo paso usa el conjunto de reglas generado para realizar predicciones de las pruebas no vistas, se pueden ocupar estrategias ordenadas y desordenadas. Las ordenadas priorizan las reglas dedicadas a la base de precisión. Después de que las reglas son ordenadas, el top de las reglas que coinciden son determinadas y la etiqueta de la clase dominante es referida como la regla relevante. Las estrategias no ordenadas es donde se determina la etiqueta de la clase dominante de todas las reglas aplicadas.

Así mismo existen metodologías que no se basan en reglas de decisión, una metodología sin esta característica son las máquinas de soporte vectorial (MSV), naturalmente están definidas

para una clasificación binaria de información numérica. Las variables con características categóricas se pueden transformar de atributos categóricos a binarios. Se asume que las etiquetas de las clases tienen dos posibles resultados $\{-1, 1\}$. Así como los modelos lineales, las *MSV* ocupan hiperplanos para separar los límites de decisión entre las dos clases. En el caso de las *MSV*, el problema de optimización al determinar estos hiperplanos está basado en la noción de márgenes; intuitivamente un hiperplano de margen máximo es aquel que separa, de forma contundente las dos clases, y para la cual un gran margen existe en cada lado de los límites de la información que aún no se entrena. Construir un hiperplano que separe de forma clara dos clases es inusual porque la información rara vez tiene una separación marcada, y al menos unos pocos datos como los no etiquetados o atípicos pueden impedir esta separación lineal.

Cuando la información es separada linealmente, hay un número infinito de maneras de construir un hiperplano entre las clases. La razón en la variación del desempeño de dos clasificadores es que al ingresar la información es colocada en una región de límite incierta entre las dos clases, que no es fácilmente generalizable con la información disponible de entrenamiento. Existen pocos datos de la base de entrenamiento en esta región incierta que son parecidos a la información de prueba, en estos casos un hiperplano de separación, cuya distancia perpendicular mínima a los datos de entrenamiento, de ambas clases, es la más lejana posible, de esta forma se puede hablar de una robusta y correcta clasificación. Esta distancia puede ser cuantificada usando el margen del hiperplano.

El margen del hiperplano es definido como la suma de las distancias al punto más cercano de los puntos de entrenamiento en cada una de las clases. Con respecto al hiperplano que separa, es posible construir otros de forma paralela que toquen la información de las clases opuestas en cada lado y no tengan puntos entre ellos. La información de entrenamiento en estos hiperplanos son conocidos como los vectores de soporte, y a la distancia entre los hiperplanos también se le llama margen. El hiperplano de separación o límite de decisión, está en medio de los vectores de soporte para poder lograr una clasificación precisa.

En varios casos, las soluciones lineales no son apropiadas para problemas donde el límite de decisión no es lineal. En general, es posible aproximar cualquier límite de decisión polinomial añadiendo un conjunto extra de dimensiones por cada exponente del polinomio. Polinomios con grados altos son significativos en la aproximación a muchas funciones no lineales. Este tipo de transformación puede ser efectiva en casos donde no se sabe si el límite de decisión es o no lineal.

Otra metodología que no es basada en reglas de decisión son los modelos de redes neuronales, que simulan el sistema nervioso humano. El sistema nervioso está compuesto por neuronas, las neuronas se conectan entre ellas mediante la sinapsis, el aprendizaje es realizado cambiando la fuerza de las conexiones sinápticas, normalmente esta fuerza en las conexiones varían en respuesta de un estímulo externo. Las redes neuronales se pueden considerar como una simulación de este proceso biológico.

Los nodos individuales en una red neuronal son referidos como las neuronas, éstas son unidades de cómputo que reciben y envían información a las demás neuronas. Las funciones en las neuronas son definidas por los pesos de la información recibida, este peso se puede visualizar como la fuerza en la conexión sináptica. Cambiando estos pesos de forma apropiada, la función puede ser aprendida. El "estímulo externo" para el aprendizaje es proporcionada por la información de entrenamiento. La idea es modificar progresivamente los pesos cuando las predicciones incorrectas son realizadas con el conjunto de pesos actual.

La llave para la efectividad en las redes neuronales, es la arquitectura ocupada para arreglar las conexiones a través de los nodos. Existe una amplia variedad de arquitecturas, desde la más simple, capa única, hasta una compleja como la red multicapa.

La estructura más básica de una red neuronal es llamada perceptrón, contiene dos capas de nodos, uno corresponde a los nodos donde se ingresa la información y un único con la información de salida. El número de los nodos de la información que ingresa es igual a la dimensión de la información subyacente. Cada nodo de ingreso recibe y transmite un único atributo numérico al nodo de salida. Los nodos de ingreso solo transmitirán valores de ingreso y no realizará ningún cómputo de estos valores, por lo que el nodo de salida es el que aplica funciones matemáticas. Las características individuales en la base de entrenamiento se asumen que son numéricas, variables categóricas son tratadas creando una variable binaria para cada valor de las variables categóricas. Por lo que para los problemas de clasificación se asume que contiene dos valores posibles para la etiqueta de la clase.

Las conexiones entre los nodos tienen distintos pesos que definen una función de los valores que se transmiten mediante los nodos de ingreso a los valores binarios; este valor se puede interpretar como la predicción perceptrón de la clase de la variable alimentada por los nodos de ingreso. El aprendizaje es realizado modificando los pesos en las uniones cuando las etiquetas predichas no corresponden a la etiqueta correcta. A la función aprendida por este modelo se le conoce como función de activación que es una función signo lineal. El proceso de entrenamiento es relativamente sencillo porque el nodo de la salida esperada es igual a la

etiqueta del valor de entrenamiento. La precisión conocida es ocupada para crear un problema de optimización con mínimos cuadrados con la función de actualizar los pesos; ya que el nodo de salida es el único con peso en una red de una capa, en este caso es un proceso sencillo de implementar.

Las redes neuronales multicapa contienen capas ocultas adicionales a las de ingreso y salida. Los nodos en las capas ocultas, principalmente pueden ser conectados con diferentes tipos de topologías, y es nombrado como red multicapa de prealimentación. Los nodos en una capa también se suponen que son conectados en su totalidad a los nodos de la siguiente capa. El problema en el proceso de entrenamiento es que la precisión de la salida de los nodos ocultos no es conocida porque no existen etiquetas de entrenamiento asociadas. Por lo que cuando existe un error en la clasificación, una acción de retroalimentación es requerida de los nodos en las capas siguientes a los nodos de las capas anteriores sobre las salidas esperadas. Esto es posible lograrlo con el uso del algoritmo retropropagación, que consiste en 2 fases principales, que son aplicados en el proceso de actualización de los pesos:

1) Fase hacia adelante: la información para la fase de entrenamiento es ingresada a la red neuronal. Los resultados de los cálculos se realizan a través de las capas usando el conjunto de los pesos. La salida final predicha puede ser comparada con la etiqueta de la clase y verificar si se clasificó de forma correcta o incorrecta.
2) Fase hacia atrás: El objetivo principal de esta fase es aprender los pesos en una dirección hacia atrás, proporcionando un error estimado de la salida de un nodo en las capas anteriores. El error estimado de un nodo en las capas ocultas es calculado como una función de los errores estimados y pesos de los nodos en las capas posteriores.

La red neuronal multicapa tiene la habilidad no solo obtener los límites de decisión en las diversas formas, sino también obtiene las distribuciones de las clases con los distintos límites de decisión en diferentes regiones de la información. Los diferentes nodos en las capas ocultas pueden capturar los diferentes límites de decisión en distintas regiones de la información, y el nodo de la capa de salida puede combinar los resultados de estos límites de decisión.

Otro tipo de clasificadores son los probabilísticos, que construyen un modelo que cuantifica la relación existente entre las características de las variables y el objetivo o clase de la variable en forma de una probabilidad.

Unos ejemplos de algoritmos de aprendizaje de máquina son ocupados para construir clasificadores.

- Regresión logística, un clasificador discriminativo, el cual aprende las características más útiles de la información que se ingresa para discriminar entre las diferentes posibles clases.
- Naive-Bayes, un clasificador generativo, el cual construye un modelo de cómo una clase se puede generar, es decir, dada una observación, regresa la etiqueta que más se parece que ha generado la observación.

Por parte de los clasificadores discriminativos, la regresión logística es la base de los algoritmos de aprendizaje de máquina supervisado, se puede ocupar para clasificar una observación en una, dos o incluso más clases.

La mayor diferencia entre Naive-Bayes y la regresión logística es la función discriminatoria. Un ejemplo para conocer cómo funciona un modelo generativo y uno discriminatorio, sería considerar que intentamos clasificar imágenes entre perros y gatos. Un modelo generativo tendrá el objetivo de entender cómo se ven los perros y cómo los gatos, por lo que se le pediría al modelo que genere a un gato, una vez que genere este resultado, tendrás que revisar el sistema y verificar qué modelo se acerca más a la realidad, posteriormente le dará la etiqueta correspondiente.

Un modelo discriminativo, por el contrario, intenta aprender a distinguir entre las clases. Para el ejemplo propuesto, suponiendo que todos los perros visten collar y que los gatos no, si una de estas características separa a las clases, el modelo está completo, por lo que como conclusión, si se le preguntara al modelo qué sabe sobre los gatos dirá que no usan collares.

El modelo discriminativo intenta calcular de forma directa $P(c|d)$, asigna pesos a las características de los documentos que mejorarán su habilidad para discriminar entre las posibles clases.

En la regresión logística existen dos fases:

- Entrenamiento: Se entrenará el sistema utilizando el descenso de gradiente estocástico y la pérdida de entropía cruzada.
- Prueba: Se calcula $P(y|x)$ y regresa la mayor probabilidad de la etiqueta $y = 1$ o $y = 0$

El objetivo de una regresión logística binaria es entrenar un clasificador que pueda realizar una decisión sobre la clase de la nueva información que se ingresa.

Considerando un documento x, representado por un vector de características $[x_1, x_2, \dots, x_n]$; el clasificador y puede resultar 1, significando que el documento sí pertenece a la clase o 0 si no pertenece, por lo que se busca conocer es la probabilidad $P(y = 1|x)$. La decisión buscada es

positivo y negativo, las características representan las cuentas de las palabras en un documento por lo que $P(y = 1|x)$ es la probabilidad de que el documento es positivo, mientras que $P(y = 0|x)$ es que es negativo.

Esta tarea se resuelve mediante el aprendizaje de una base de entrenamiento, un vector de pesos y un término de sesgo. Cada peso w_i es un número real y está asociado a una de las características x_i. Cada peso representa la importancia que la característica de ingreso representa para la decisión de clasificación, el cuál puede ser positivo si está asociado a la clase o negativo si no lo está. Con esto se puede esperar que la palabra "increíble" tenga un peso positivo alto mientras que "pésimo" un gran peso negativo; el término de sesgo, también llamado la interceptación, es otro número real que se añade a los pesos.

Para tomar una decisión, el clasificador realiza la suma del producto $w_i x_i$ y suma el término b de sesgo.

$$z = \left(\sum_{i=1}^{n} w_i x_i\right) + b$$

Los pesos al ser números reales hacen que la imagen tenga un rango $(-\infty, \infty)$, por lo que para crear una probabilidad se pasará z a una función logística teniendo la siguiente ecuación:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Las variables pueden tomar valores reales y su rango quedará en [0,1], lo necesario para una probabilidad, y se tienen las siguientes ecuaciones:

$$P(y = 1) = \sigma(w \cdot x + b) = \frac{1}{1 + e^{-(w \cdot x + b)}}$$
$$P(y = 0) = 1 - \sigma(w \cdot x + b) = 1 - \frac{1}{1 + e^{-(w \cdot x + b)}}$$

Por lo que para tomar una decisión, se debe cumplir que $P(y = 1|x) > 0.5$. Si se toman otros puntos de corte, uno mayor estaría clasificando de una forma más estricta y se podría dejar fuera documentos que si tengan información de interés; por otro lado, poner un corte menor haría que nuestra etiqueta de interés pueda tener información más genérica, alejándolo de nuestro objetivo de clasificar de la mejor forma los documentos.

En la regresión logística se conoce la correcta etiqueta para cada observación, por lo que se requiere conocer los parámetros w y b para hacer que cada observación se acerque a su correcta etiqueta. Para esto se requieren dos componentes:

- Primero es conocer qué tan cerca se tiene la estimación con la etiqueta real que será nombrado como la función de pérdida de entropía cruzada.

- Segundo, es necesario un algoritmo recursivo de optimización que actualice los pesos que se realiza con el descenso del gradiente estocástico.

Se requiere una función que exprese de una observación x, qué tan cerca está del resultado del clasificador, se ocupa la función de pérdida que prefiera la correcta etiqueta de la base de entrenamiento que se parezca más, esta función es la estimación condicional de máximo verosimilitud, se eligen los parámetros w, b que maximicen el logaritmo de la probabilidad de la etiqueta y en la base de entrenamiento dada la observación x. La función de pérdida resultante es la pérdida del logaritmo negativo de verosimilitud, llamada función de pérdida de entropía cruzada.

Cuando es necesario hacer una clasificación de 3 o más clases, en este caso se realiza la regresión logística multinomial; el clasificador logístico usa una forma generalizada de la función logística, donde $z = [z_1, z_2, \dots, z_k] \in (0,1]$ es un vector con k valores arbitrarios; donde la suma de la siguiente ecuación debe resultar en 1.

$$softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad 1 \leq i \leq k$$

En la siguiente sección se abordará el clasificador generativo Naive-Bayes, mismo que será aplicado para efectos de este estudio.

2.5 Clasificador Probabilístico Naive-Bayes

Como anteriormente se ha mencionado, la tarea de un clasificador supervisado es ingresar la información x y un grupo de clases de resultados $Y = y_1, y_2, \dots, y_m$ para mostrar la predicción $y \in Y$. Bajo esta premisa, para la clasificación de textos, se usará la letra c por clase en lugar de y y d por documento en lugar de x. Para un algoritmo supervisado, se tiene un conjunto de información para entrenamiento de N documentos y cada uno tiene una etiqueta con la clase $(d_1, c_1), \dots, (d_N, c_N)$. El objetivo es entrenar un clasificador que sea capaz de predecir un nuevo documento d a su correcta clase $c \in C$.

Naive-Bayes es un clasificador probabilístico, es decir para un documento d, entre todas las clases $c \in C$ el clasificador regresa a la clase $\hat{c}$ que tiene la mayor probabilidad dado el documento ingresado.

$$c = P(c|d)$$

Esta idea de la inferencia Bayesiana ha sido conocida desde el trabajo de Bayes (1763); la intuición de la clasificación Bayesiana es para usar el teorema de Bayes de probabilidades condicionales. El teorema cuantifica la probabilidad condicional de una variable aleatoria

(clases), dada una observación conocida sobre el valor de otro conjunto de variables aleatorias (variables características).

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Sustituyendo las variables para la clasificación de textos, se puede visualizar de la siguiente forma.

$$c = P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

En este caso, se puede simplificar el denominador $P(d)$. Esto es posible ya que la formula se calculará para cada clase posible; pero $P(d)$ no cambia en cada clase ya que siempre se estará calculando esa probabilidad para el mismo documento, donde se está realizando el cálculo, por lo que podemos elegir la clase que maximice la siguiente formula.

$$\hat{c} = \underset{c \in C}{argmax}\, P(d|c)P(c)$$

Sin pérdida de generalidad, se puede representar al documento *d* como un conjunto de características $f_1, f_2, \dots, f_n$, de tal forma que:

$$\hat{c} = \underset{c \in C}{argmax}\, P(f_1, f_2, \dots, f_n|c)P(c)$$

El clasificador Naive-Bayes realiza 2 grandes suposiciones.

- La primera: el conjunto de palabras existentes en el documento; asume que la posición en la que se encuentra la palabra no importa.
- La segunda: comúnmente llamada la suposición Naive-Bayes, es la suposición de independencia condicional de que las probabilidades $P(f_i|c)$, es decir que son independientes para cada clase.

Por lo que se tiene la siguiente igualdad.

$$P(f_1, f_2, \dots, f_n|c) = P(f_1|c) \cdot P(f_2|c) \cdot \dots \cdot P(f_n|c)$$

La ecuación final para elegir una clase mediante el clasificador Naive-Bayes es:

$$\hat{c} = \underset{c \in C}{argmax}\, P(c) \prod_{f \in F} P(f|c)$$

Considerando el estimador de máxima verosimilitud, que ocupará las frecuencias de las palabras de los documentos. Para la probabilidad previa $P(c)$ buscamos el porcentaje de los documentos, de la base de entrenamiento, que están en cada clase c. Si N_c es el número de documentos en nuestra base de entrenamiento con la clase c y N_{doc} el número total de documentos; entonces:

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

Para conocer la probabilidad $P(f_i|c)$, se asume como característica la existencia de la palabra en el conjunto de palabras, es decir se quiere conocer $P(w_i|c)$, que se calcula como las veces que la palabra w_i aparece entre todas las palabras y todos los documentos de la clase c. Se concatenará todos los documentos de la categoría c. Posteriormente usamos la frecuencia de w_i en este documento concatenado para estimar la probabilidad de máxima verosimilitud.

$$\hat{P}(w_i|c) = \frac{frec(w_i\,, c)}{\sum_{w \in V} frec(w, c)}$$

Donde el vocabulario V consiste en la unión de todas las palabras en todas las clases. Existe un problema, si se está queriendo estimar la probabilidad de la palabra "genial", dada la clase *positivo*, pero supongamos que en las bases de entrenamiento no existe la palabra en ambas, y posiblemente la palabra se esté ocupando de forma sarcástica en la clase *negativo*. En este caso, la probabilidad resultará cero:

$$\hat{P}(genial|positivo) = \frac{frec(genial\,, positivo)}{\sum_{w \in V} frec(w, positivo)} = 0$$

Ante la independencia que asume Naive-Bayes, realiza el producto de todas las características, por lo que cualquier cero en las probabilidades hará que la probabilidad de la clase sea cero sin importar la demás evidencia.

La solución más simple es añadir un nivel de suavizamiento de Laplace, a diferencia del estimador original, esta acción tiene el propósito de dar la oportunidad que todas las palabras del vocabulario aparezcan al menos una vez en los documentos, esta solución puede ser efectiva si las etiquetas de clasificación contienen temas en conjunto; se expresa de la siguiente forma:

$$\hat{P}(w_i|c) = \frac{frec(w_i\,, c) + 1}{\sum_{w \in V}(frec(w, c) + 1)} = \frac{frec\,(w_i, c) + 1}{(\sum_{w \in V} frec(w, c)) + |V|}$$

Por lo regular, cuando en nuestra base de prueba aparecen palabras desconocidas, es evidente que no se encuentra en nuestra base de entrenamiento, es decir no se encuentra en el vocabulario, la mejor solución es ignorar estas adiciones para que no se incluyan en el cálculo de las probabilidades.

Ciertos sistemas eligen ignorar otro tipo de palabras, *stop words*, palabras de gran frecuencia que no tienen un significado concreto, como son las conjunciones, proposiciones, entre otras. Estas palabras son removidas tanto de la base de entrenamiento como de prueba. En la mayoría de las aplicaciones de clasificación de textos, utilizar estas palabras no mejora el rendimiento del análisis.

Para la clasificación multi-clase, lo más común en el procesamiento de lenguaje es la clasificación multinomial, en el cual cada clase es mutuamente exclusiva y cada documento aparece en una sola clase.

El procedimiento de prueba y error para la clasificación de textos, se ocupa para entrenar el modelo, se ocupa una base de desarrollo para poder ajustar ciertos parámetros y poder generar el modelo más eficiente. Esta base de desarrollo puede ser un subconjunto de la base de prueba, ya que es necesario conocer la clase a la que pertenece cada documento. Mientras que la base de desarrollo evita sobre ajustar la base de prueba, tener una base arreglada puede que la base de desarrollo no sea lo suficientemente grande para que sea representativa, por lo que lo mejor sería ocupar toda nuestra información tanto para la base de entrenamiento como de prueba, sin embargo esto no siempre es posible por el volumen de información que se busca clasificar. Para verificar esto se realiza una validación cruzada, de forma aleatoria se selecciona un conjunto de la base de prueba, y se calcula el porcentaje de error en la base de prueba.

En el siguiente capítulo se aplicará el algoritmo de Naive-Bayes a una base de datos de la red social Twitter, enfocado a un análisis de la cuenta oficial del Servicio de Transporte Metro de la Ciudad de México para poder conocer los comentarios que los usuarios manifiestan mediante este medio.

Capítulo 3. Aplicación del algoritmo

3.1 Descarga de la información

La descarga de información se realizará directo desde el programa R, sin embargo para poder obtener de forma legítima y en gran cantidad la base de datos para efectos de este proyecto, es necesario ingresar a Twitter Developer, apartado de la misma red social donde, posterior a un pequeño cuestionario sobre las intenciones de obtener los permisos y nombrar una app, recibirás los 4 códigos necesarios de acceso para que poder ingresar a Twitter desde el programa R. Es importante señalar que estos accesos son únicos y personales, por lo que no es recomendable compartirlos ya que está vinculado a tu cuenta de Twitter.

Para que en R se pueda realizar la descarga de la información, es necesario el uso de la librería “rtweet”. Para este caso en particular se ocuparan 2 comandos, uno para el acceso y otro para la obtención de la información.

El primer comando será “create_token” que constará de 5 argumentos, 1 será el nombre de la app que diste de alta, adicional a los 4 códigos que se reciben de Twitter Developer, de esta forma podrás realizar de forma satisfactoria tu acceso a Twitter.

Posteriormente se ocupará el comando “search_tweets2” que podrá descargar los 280 caracteres de cada tweet, y para efectos de este estudio se ocuparán 5 argumentos con las siguientes características:

1) Q = de tipo texto, donde se colocará la palabra o vector de palabras que deseas que se busque para la descarga de información, donde no hace distinción entre mayúsculas y minúsculas.
2) N = de tipo numérico, solicitando la cantidad de tweets que quieres que extraiga, con un máximo de 18,000.
3) Since = de tipo fecha, que indica la fecha mínima de la descarga de información.
4) Until = de tipo fecha, que indica la fecha máxima de la descarga de información.
5) Include_rts = de tipo lógico, para descargar o no *retweets*.

A pesar de tener 2 argumentos de fecha para poder colocar el tiempo que quieres obtener, es importante señalar que la misma app no te permite acceder a información con una fecha anterior a 9 días, esto para cuidar la información y privacidad tanto de los usuarios como de la plataforma, por lo que la información de los meses de junio y julio del 2019 se estuvo descargando de manera semanal, por la razón anterior, la información se fue guardando en archivos de hojas de cálculo en Excel.

Para poder obtener la información del STC Metro, se colocó en el primer argumento el texto “metrocdmx”, caracteres contenidos en la cuenta oficial de Twitter (@MetroCDMX); no se colocó el signo de arroba para obtener los tweets que pudieran contener un *hashtag* con esa cadena de texto y pudieran contener información valiosa sin la necesidad de mencionar la cuenta oficial. Adicional, se colocó la opción de no *retweets* para tener una base más limpia por posibles noticias que pudieran ser de gran impacto, lo cual incrementaría de manera considerable la base en un solo tema.

La implementación de los códigos mencionados con anterioridad, se verían de la siguiente forma:

library(rtweet)

token = create_token(app = "Test_Data_App",consumer_key = "código 1",

consumer_secret = "código 2",access_token = "código 3", access_secret = "código 4")

base=search_tweets2(c("metrocdmx"), since = ”aaaa-mm-dd”, until = ”aaaa-mm-dd”, include_rts=FALSE)

La base obtenida contiene un total de 92 variables, entre las que se destacan:

- El texto del tweet.
- La palabra de búsqueda asociada al tweet.
- La fecha de creación.
- El usuario del tweet.
- La cantidad de caracteres del tweet.

Como información adicional, es recomendable verificar la diferencia en husos horarios con la que se descarga la información, esto para tener los tiempos correctos en que fueron creadas las menciones.

En la descarga entre los meses de junio y julio del 2019, 61 días, se tuvo un total de 86,157 tweets distribuidos de la siguiente forma.

Gráfico 2 - Tweets por día

De forma general, es posible observar valles en el Gráfico 2 con cierta periodicidad, dicho evento se debe a que son los días sábados y domingos, días donde la afluencia es menor.

3.2 Desarrollo de la base de datos

Una vez teniendo la información, conforme lo comentado en el capítulo anterior, es necesario conocer la base, para ello hay que hacer una lectura de unos cuantos tweets. La razón principal para esta actividad es conocer ciertas frases que se ocupan o *hashtag* que pueden ser ocupados para poder realizar de forma más sencilla la clasificación de los mismos.

Para esto, es importante resaltar que al ser una red social, aunque se tiene permitido un máximo de 280 caracteres, la escritura es libre por lo que hay que tener cuidado de los errores más comunes como el uso de las “s”, “c”, “z”, “b”, “v” y “h”, así como el uso de los acentos. Nosotros si somos capaces de darle un sentido y saber que se está hablando de lo mismo, sin embargo, al ser distintas, una máquina simplemente las reconocerá como si tuvieran significados distintos, por lo que se requiere realizar una homologación de términos para contar con menos palabras en la base.

Para poder iniciar con una depuración se deben tener las siguientes consideraciones sobre el manejo de la base de tweets.

- Existen emojis, los cuales se presentan con cierta codificación al ingresarlos a R.
- El idioma español cuenta con acentos, así como distintas virgulillas.
- Hay palabras escritas con letras mayúsculas y minúsculas.
- Se pueden colocar números.

- Los usuarios pueden subir tanto imágenes como videos, por lo que aparecen como si fuera una liga URL.

Una vez que se tiene un conocimiento más detallado de la base, se prosigue a realizar la depuración y homologación del vocabulario, los números y los signos de puntuación al no tener gran relevancia en la información que proporcionan, es viable eliminarlos desde un inicio, de cualquier manera uno puede notar que la forma de referirse a las diferentes líneas del metro son distintas, por el nombre oficial, por colores, con abreviaciones y con números, por lo tanto no se eliminarán símbolos ni números en un primer paso.

Para este estudio, el primer paso de la depuración será eliminar los emojis, donde se ocupará el siguiente comando

Iconv(base, from = "latin1", to = "ASCII", sub = "")

De esta forma se reemplazará el texto codificado de los emojis en vacío.

Posteriormente, se deberán hacer las correcciones pertinentes para poder contar con una base lo más limpia posible, para esto se ocupará la función “gsub” con las variantes incluidas en la Tabla 3.

Tabla 3 - Comandos de depuración

#	**Comando**	**Acción**
1	gsub("á", "a", base)	Se colocará cada letra con el acento correspondiente
2	gsub("palabra\\w+", "", base) gsub("\\w+palabra", "", base)	Con el texto “\\w+” reemplazará toda la cadena de texto que se encuentre antes o después, según la posición, de la palabra que se busca modificar.
3	tolower(base)	Cambia todas las letras a minúsculas
4	gsub("[[:punct:]]", "", base)	El comando escrito eliminará todos los signos de puntuación en la base.
5	gsub("[[:digit:]]", "", base)	El comando escrito eliminará todos los números en la base.

Con el comando 1 se podrán sustituir los acentos y virgulillas que existan en las letras, de la misma forma se podrá colocar una cadena de texto específica y reemplazarla por lo que deseé.

Con el comando 2, es posible realizar la mayoría de los reemplazos requeridos para poder dejar las palabras con la raíz que deseé, sin embargo, es importante mencionar que los reemplazos son por caracteres, por lo que si se busca reemplazar una palabra, se debe tener cuidado de que otras pueden contener esa cadena de texto y realizar el reemplazo.

Con estos primeros dos comandos es posible terminar con el reemplazo de las distintas formas de referirse a la líneas del metro.

La siguiente acción será cambiar todas las letras a minúsculas con el comando número 3 y aplicar los comandos 4 y 5 para tener un manejo más sencillo, y posteriormente ocupar el comando 2 para poder dejar un vocabulario más compacto, dentro de estas acciones, se concluirá con la homologación de las distintas formas al referirse a las estaciones del metro.

Y una vez que se tienen todos los cambios requeridos, es necesario hacer una última depuración. La base se ingresará a un “corpus”, este tipo de dato se encuentra en la librería “tm”, dentro de la misma, contiene una colección de palabras “stopwords” que tiene palabras como artículos, conjunciones y adverbios, las cuáles se deberán eliminar de la base ya que son palabras demasiado comunes que no servirán para el estudio; para esto, también existe la función “removeWords” para realizar esta depuración.

corpus = Corpus(VectorSource(base))

tm_map (corpus,removeWords, stopwords("spanish"))

Una vez que se tiene toda la base depurada, se creará una matriz documento-término (dtm) ingresando en el comando el corpus ocupado en el paso anterior, donde cada fila o documento, en este caso, contiene cada tweet, y las columnas contienen a cada uno de los términos que existen en la base.

dtm = DocumentTermMatrix(corpus)

Esta matriz se puede expresar de la siguiente forma:

$$\mathcal{M}_{d\times t} \qquad donde: \qquad \begin{matrix} d\ es\ la\ cantidad\ de\ tweets\ de\ la\ base \\ t\ son\ los\ términos\ en\ la\ base, vocabulario \end{matrix}$$

De esta manera, ya tenemos la base preparada para poder aplicar el algoritmo de Naive-Bayes y realizar la clasificación.

3.3 Aplicación de algoritmo Naive-Bayes

De acuerdo a lo mencionado en el capítulo anterior, el algoritmo Naive-Bayes realiza la clasificación conforme la probabilidad de que las palabras estén en un grupo o en otro. Existe independencia ya que cada tweet es distinto a otro, así como la omisión de *retweets*. Por lo que es indispensable determinar los grupos que se ocuparán en la clasificación, para conocer las problemáticas del STC Metro, se decidieron crear 5 etiquetas con las siguientes características.

Tabla 4 - Grupos de clasificación

Grupo	**Descripción**
Servicio	Comentarios sobre las deficiencias que existen en el servicio, como es la compra de boletos en ventanilla, fallas en las escaleras eléctricas, en ventiladores y puertas; así como la lentitud en el avance de los trenes.
Comercio	Comentarios acerca del comercio informal que existe en las inmediaciones y las instalaciones del metro, así como en los vagones.
Seguridad	Comentarios relacionados con los delitos, robos e inseguridad existente en las inmediaciones e instalaciones, así como en los vagones.
Fallas	Mensajes que envía la cuenta del STC Metro por fallas mecánicas, eléctricas, entre otras, de los trenes mientras se encuentran en servicio.
Fuera	Mensajes con falta de información, que no pueden ser catalogados en ninguno de los grupos anteriores.

Una vez que se han determinado, es importante reconocer *hashtags*, oraciones y palabras que se puedan relacionar con cada uno de los grupos, es posible que existan palabras entre diferentes grupos, por lo que se debe tener cuidado en seleccionarlas y poner que se encuentran en ambos grupo. Este ejercicio no se realiza una única vez, se tiene que estar probando el modelo para quitar o anexar vocabulario, y de esta forma obtener el mejor resultado posible.

Ya que en la depuración se eliminaron acentos, en la siguiente tabla se visualizarán de esa forma algunas de las palabras seleccionadas para cada grupo.

Tabla 5 - Palabras de grupos

Comercio	Seguridad	Servicio	Falla	Fuera
ambulantaje	arma	basura	avisometro	adopta
bocina	arresto	boleto	balatas	aniversario
comercio	asalto	espera	corriente	anticipatusalida
compra	delincuencia	lento	desalojo	elmetroestuyocuidalo
discos	denuncia	limpieza	electrica	iniciamoselservicio
informal	inseguridad	minuto	falla	mayorculturamejormovilidad
musica	narco	parar	humo	museo
vagonero	robo	taquilla	mecanismo	planeatuviaje
vendedores	vigilancia	tarde	movilidadcdmx	selfie
venta	violencia	ventilacion	sistema	taller

Como apoyo para poder reconocer estas palabras, se realizó un diagrama de correlaciones que se puede observar en la Ilustración 3, con la finalidad de tener una forma visual de la agrupación de dichas palabras. Se debe señalar que el diagrama está conformado con las palabras más frecuentes de la base, razón por la cual no aparecen palabras de todos los grupos. Las líneas del metro no fueron colocadas en ningún grupo ya que a este nivel de la clasificación nos interesa conocer el panorama general del servicio.

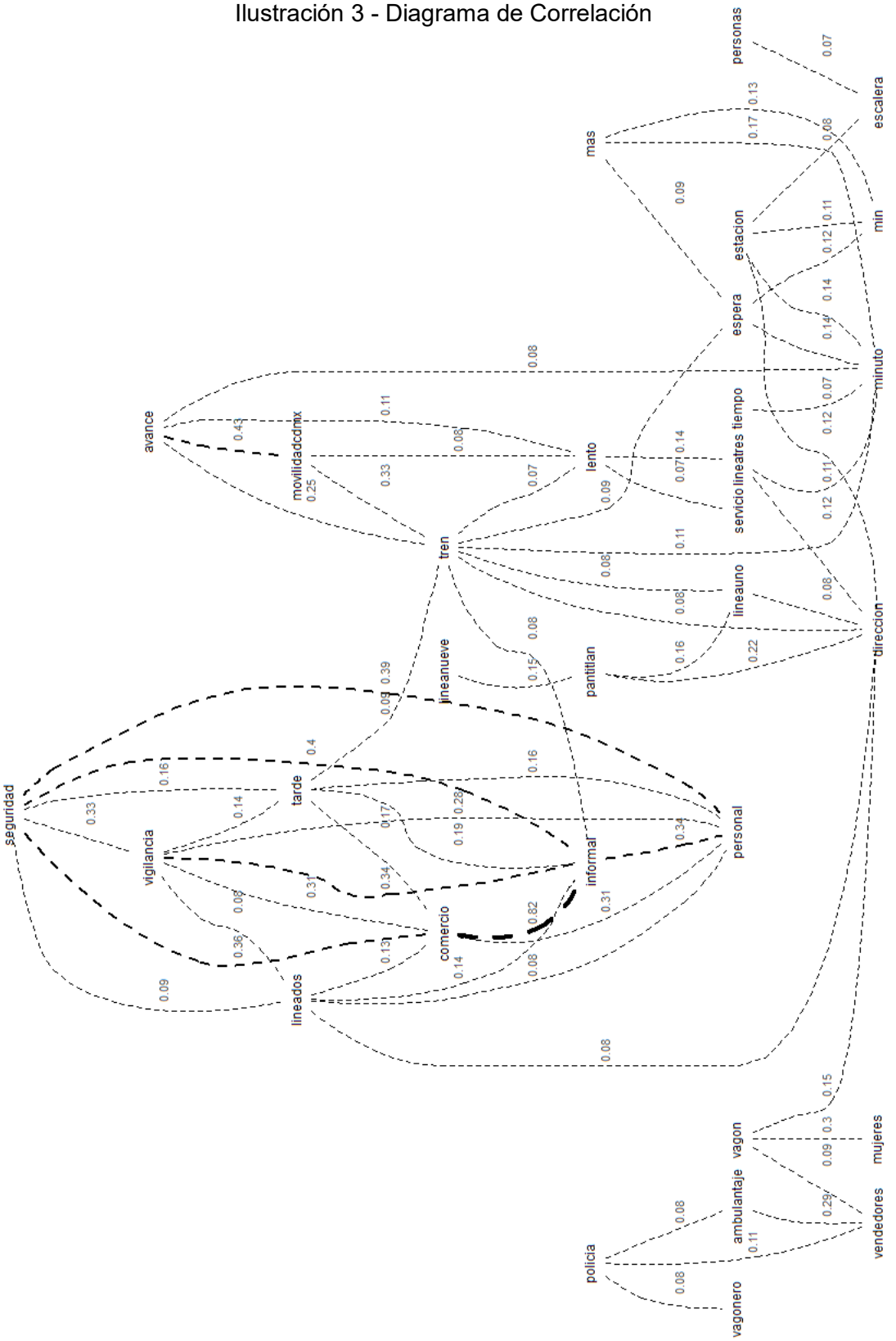
En esta ilustración es posible visualizar en la agrupación con mayor número de palabras al centro, que el grupo de comercio y seguridad están relacionados, así mismo, en la parte derecha se observa el grupo de servicio; la agrupación con más palabras parece tener la información sobre el servicio referente al avance del transporte y la pequeña agrupación más hacía el servicio en las estaciones.

Una vez identificadas las frases o palabras, conforme la clasificación lo requiera, se deberán depositar en una base donde tendrá dos variables:

1) Una contendrá la frase exclusiva.
2) El grupo al que pertenece la frase.

Esta base servirá para clasificar la base de tweets y poder realizar la agrupación correspondiente, también deberá pasar por su proceso de depuración y finalmente se deberá colocar en una variable de atributo “dtm”. Es importante mencionar que esta base no quedará al primer intento, posteriormente al realizar la clasificación, se tendrá que retroalimentar la base para poder conseguir la precisión deseada.

Ilustración 3 - Diagrama de Correlación



El siguiente paso para realizar la clasificación será extraer las etiquetas (grupos) de la base de clasificación y transformarlas en una variable de factor para ocuparse como categorías, en este caso, la base que fue ocupada contiene 440 registros, así mismo se coloca la base a clasificar bajo la “dtm”, que son 86,157 tweets.

```
dtm_entrenadora = dtm_grupos[1:440, ]

dtm_etiquetas = as.factor(base_grupos[1:440,]$Grupo)

dtm_clasificar = dtm[1:86157, ]
```

Se extraerá la frecuencia de las palabras de la base entrenadora para poder colocarlas, tanto en la dtm de la base que queremos clasificar como en la que entrena. Para este paso es importante señalar que las palabras de la frecuencia deben existir en la base que queremos clasificar, de lo contrario, el programa nos arrojará un error. Ya que estas palabras salieron de la misma base, no se tendrá este inconveniente, sin embargo en caso de querer clasificar nueva información es importante cumplir con lo anteriormente señalado.

```
frecuencia = findFreqTerms(dtm_entrenadora,1)

dtm_freq_entrenadora = dtm_entrenadora[ ,frecuencia]

dtm_freq_clasificar = dtm_clasificar[ ,frecuencia]
```

Para poder realizar los conteos de las palabras en cada uno de los tweets se creará una función para que realice esta operación, colocando un “Sí” cuando la palabra aparezca en el tweet y un “No” en caso contrario; esta función se aplicará a las 2 bases anteriores creadas con la frecuencia de la base entrenadora, el comando “MARGIN=2”, significa que la función se aplicará por columnas.

```
conteo = function(x){x<-ifelse(x>0,"Si","No")}

tw_entrenadora = apply(dtm_freq_entrenadora, MARGIN = 2,conteo)

tw_clasificar = apply(dtm_freq_clasificar, MARGIN = 2,conteo)
```

Una vez teniendo estas bases con los conteos, se aplicará el modelo de Bayes para poder calcular las probabilidades de cada una de las filas en la base entrenadora, este cálculo se realiza mediante la función “naiveBayes” contenida en la librería “e1071”, donde ingresa en los argumentos el conteo de la base entrenadora y las etiquetas de los grupos, esta función calculará la probabilidad y finalmente, estas se deberán ingresar en una función “predict” junto

con la base que se quiere clasificar para poder obtener la clasificación y las etiquetas de los grupos.

tw_clasificador = naiveBayes(tw_entrenadora, dtm_etiquetas)

clasificación = predict(tw_clasificador, tw_clasificar)

El resultado de esta función *predict* dará como resultado un vector que corresponde al grupo al que fue clasificada la mención.

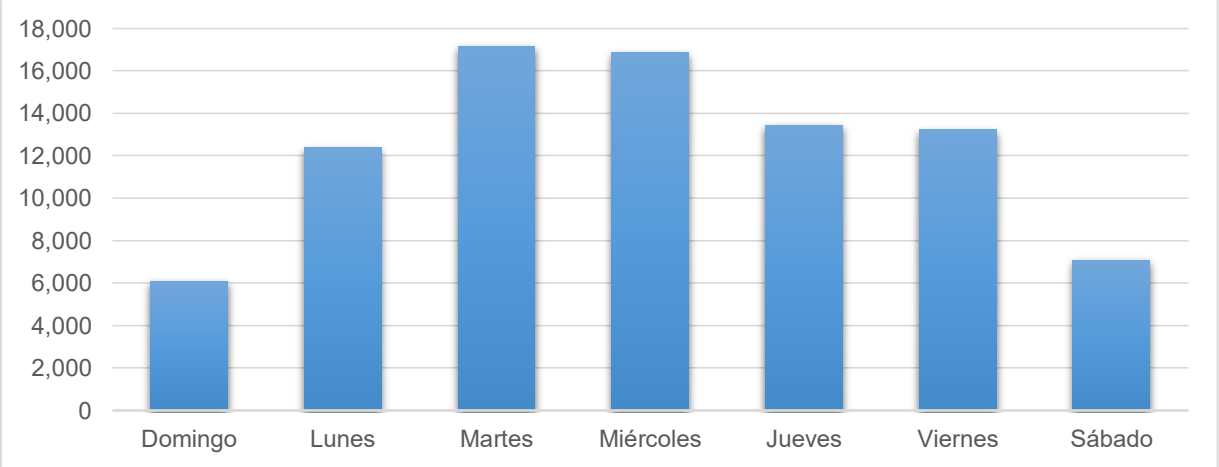
De esta forma se concluye con el proceso de clasificación de los tweets, en la siguiente sección se colocarán los resultados obtenidos y se mostrará un análisis descriptivo de los 4 grupos de interés.

Este proceso se aplicó se manera similar para poder vincular los tweets con las distintas líneas del metro, con el fin de poder encontrar qué línea es la que tiene mayor número de interacciones. La forma en que se realizó esta clasificación, fue crear una base con todas las estaciones del metro y asociarlas con su línea correspondiente, en caso de que una misma estación estuviera en más de una línea, este dato se repetiría pero con su línea correspondiente, adicionalmente, se debe crear una asociación a un grupo “Sin línea”, donde se colocaron palabras genéricas que aparecen en la base para que cuando en el tweet no se mencione ninguna estación, ésta caiga en este grupo.

3.4 Resultados del modelo

Como había mencionado, una de las variables que se consiguen en la base es la fecha y hora en que se realizó el tweet, por lo que es posible extraer los días de la semana en que hay una mayor interacción; como un primer acercamiento, es de interés conocer en qué momentos del

Gráfico 3 - Tweets por día de la semana

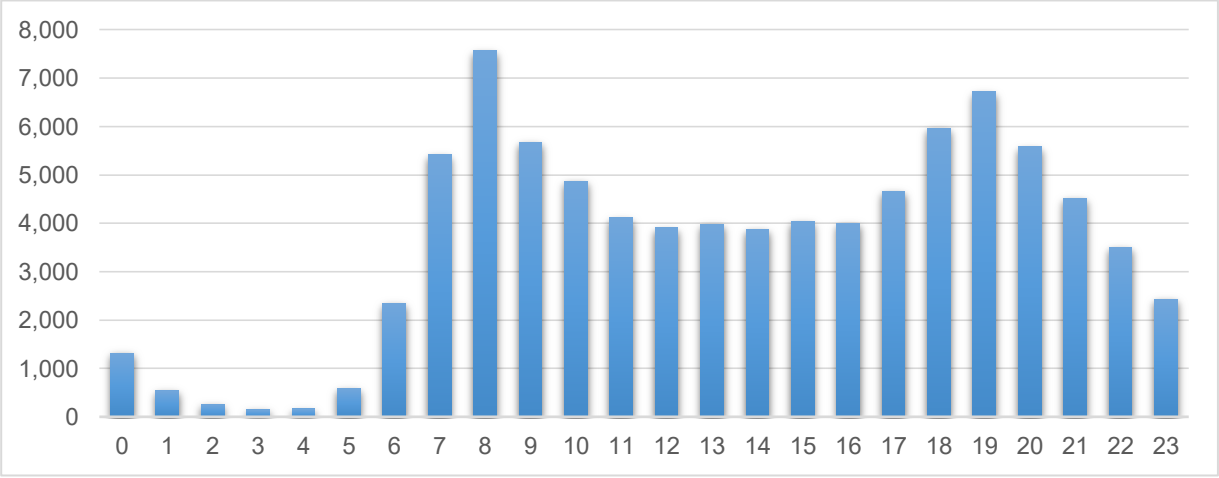


día y de la semana es mayor la cantidad de interacciones que involucran a este servicio de transporte.

Es posible observar que los días entre semana son prácticamente iguales, a excepción de los días martes y miércoles, con 17,135 y 16,851 respectivamente, mientras que los fines de semana de forma considerable es menor la interacción, ya que hay menos de la mitad de interacciones con 7,065 y 6,083 para sábado y domingo.

Conforme al Gráfico 4, es posible verificar que los puntos donde hay mayor cantidad interacción son las horas alrededor de la entrada y salida de la jornada laboral, la hora máxima es a las 8 de la mañana con 7,569 tweets.

Gráfico 4 - Tweets por hora



Pasando al tema de la clasificación, se realizó un muestreo aleatorio simple para poder evaluar la precisión del modelo, la muestra fue de 382 tweets. Se realizó una tabla de validación cruzada, donde se comparan las etiquetas entre las correctas y las que el modelo predijo, misma donde la diagonal representa la cantidad y proporción de tweets clasificados mediante el modelo de manera correcta.

Crosstable = CrossTable(clasificacion, dtm_muestreo_etiqueta, prop.chisq = F, prop.t = F, prop.r = T, prop.c = F, dnn = c("Predichos","Correctos"))

En la clasificación de las problemáticas del metro, 350 resultaron con una buena clasificación concluyendo con una precisión del 91.62%. Por otro lado, en la clasificación de las líneas del

metro, 356 resultaron con una clasificación correcta, es decir, con una precisión del 93.19%. Es importante recalcar que estos resultados de precisión no fueron obtenidos en primera instancia, para poder llegar a dichos valores se realizaron varias pruebas, anexando o quitando palabras o frases a la base entrenadora para poder tener el mejor resultado posible; la razón por la que algunos tweets obtuvieron una etiqueta errónea, es porque comparten palabras de otros grupos, cabe destacar que el sentido de la mención no hace referencia al mismo.

En la Tabla 6, se muestra la validación cruzada sobre los grupos, como se mencionó, la diagonal señala la clasificación correcta. Se puede observar que se comparten etiquetas con los demás grupos; “Comercio” junto con “Fuera” son los que tienen mayor dispersión, la razón de esta situación es que el vocabulario de distintas etiquetas en la base de entrenamiento están incluidas dentro de un mismo tweet, y hace que sea clasificado en otro grupo.

Tabla 6 - Validación cruzada grupos

	Correctos ↓					
Predichos →	**Comercio**	**Falla**	**Seguridad**	**Servicio**	**Fuera**	**Total Fila**
Comercio	30	0	3	1	3	**37**
	0.811	0	0.081	0.027	0.081	**0.097**
Falla	0	5	0	1	0	**6**
	0	0.833	0	0.167	0	**0.016**
Seguridad	2	0	14	0	1	**17**
	0.015	0	0.824	0	0.059	**0.045**
Servicio	2	0	0	125	4	**131**
	0.015	0	0	0.954	0.031	**0.343**
Fuera	2	0	3	10	176	**191**
	0.015	0	0.081	0.052	0.921	**0.500**
Total Columna	**36**	**5**	**20**	**137**	**184**	**382**

En la Tabla 7, se puede observar una clasificación más limpia, esto no es coincidencia ya que para esta clasificación solo comparten el vocabulario de las estaciones en donde se encuentran los transbordos. Si se observan las filas, la línea 1 con la 3 coinciden con la estación Balderas, y la 9 con 1 donde comparten las estaciones Tacubaya y Pantitlán. Existen etiquetas en el grupo “sin línea” que pueden corresponder a una línea en específico, esto sucede por el vocabulario que se ocupa para esta clasificación, es decir, el tweet a pesar de contener palabras propias de las líneas o estaciones, contiene una mayor cantidad para el grupo “sin línea”.

Tabla 7 - Validación cruzada estaciones

	Correctos ↓													
Predichos →	Línea 1	Línea 2	Línea 3	Línea 4	Línea 5	Línea 6	Línea 7	Línea 8	Línea 9	Línea A	Línea B	Línea 12	sinlinea	Total Fila
Línea 1	15	0	1	0	0	0	0	0	0	0	0	0	0	**16**
	0.938	0	0.062	0	0	0	0	0	0	0	0	0	0	**0.042**
Línea 2	0	17	0	0	0	0	0	0	0	0	0	0	0	**17**
	0	1.000	0	0	0	0	0	0	0	0	0	0	0	**0.045**
Línea 3	0	0	19	0	0	0	0	0	0	0	0	0	0	**19**
	0	0	1.000	0	0	0	0	0	0	0	0	0	0	**0.050**
Línea 4	0	0	0	1	0	0	0	0	0	0	0	0	0	**1**
	0	0	0	1.000	0	0	0	0	0	0	0	0	0	**0.003**
Línea 5	0	0	0	0	4	0	0	0	0	0	0	0	0	**4**
	0	0	0	0	1.000	0	0	0	0	0	0	0	0	**0.010**
Línea 6	0	0	0	0	0	1	0	0	0	0	0	0	0	**1**
	0	0	0	0	0	1.000	0	0	0	0	0	0	0	**0.003**
Línea 7	0	0	0	0	0	0	13	0	0	0	0	0	0	**13**
	0	0	0	0	0	0	1.000	0	0	0	0	0	0	**0.034**
Línea 8	0	0	0	0	0	0	0	7	0	0	0	0	1	**8**
	0	0	0	0	0	0	0	0.875	0	0	0	0	0.125	**0.021**
Línea 9	0	1	0	0	0	0	0	0	9	0	0	0	0	**10**
	0	0.100	0	0	0	0	0	0	0.900	0.000	0	0	0	**0.026**
Línea A	0	0	0	0	0	0	0	0	0	5	0	0	1	**6**
	0	0	0	0	0	0	0	0	0	0.833	0	0	0.167	**0.016**
Línea B	0	0	0	0	0	0	0	0	0	0	4	0	0	**4**
	0	0	0	0	0	0	0	0	0	0	1.000	0	0	**0.010**
Línea 12	0	0	0	0	0	0	0	0	0	0	0	6	0	**6**
	0	0	0	0	0	0	0	0	0	0	0	1.000	0	**0.016**
sinlinea	1	4	5	1	0	0	3	2	2	1	2	1	255	**277**
	0.004	0.014	0.028	0.004	0	0	0.011	0.007	0.007	0.004	0.007	0.004	0.921	**0.725**
Total Columna	16	22	25	2	4	1	16	9	11	6	6	7	257	382

En la tabla 8 y 9 se muestra la distribución de cómo quedaron conformados cada una de las etiquetas para ambas clasificaciones, así mismo se realizó el producto entre los tweets clasificados con el porcentaje obtenido de la diagonal de la validación cruzada.

Tabla 8 - Clasificación de grupos

Grupo	Conteo	Proporción	% Clasificación correcta	Clasificados correcto
Comercio	8,120	9.42%	81.10%	6,585
Falla	1,260	1.46%	83.30%	1,050
Seguridad	6,992	8.12%	82.40%	5,761
Servicio	28,615	33.21%	95.40%	27,299
Sin clasificar	41,170	47.78%	92.10%	37,918
Total	**86,157**			**78,613**

Para la clasificación de grupo se esperan 78,613 tweets clasificados de manera correcta, lo que equivale a un 91.24% de la base.

Tabla 9 - Clasificación de líneas

Línea	Conteo	Proporción	% Clasificación correcta	Clasificados correcto
Línea 1	3,151	3.66%	93.80%	2,956
Línea 2	2,905	3.37%	100.00%	2,905
Línea 3	5,123	5.95%	100.00%	5,123
Línea 4	402	0.47%	100.00%	402
Línea 5	615	0.71%	100.00%	615
Línea 6	511	0.59%	100.00%	511
Línea 7	3,198	3.71%	100.00%	3,198
Línea 8	1,401	1.63%	87.50%	1,226
Línea 9	2,519	2.92%	90.00%	2,267
Línea A	1,825	2.12%	83.30%	1,520
Línea B	1,503	1.74%	100.00%	1,503
Línea 12	1,823	2.12%	100.00%	1,823
Sin clasificar	61,181	71.01%	92.10%	56,348
Total	**86,157**			**80,397**

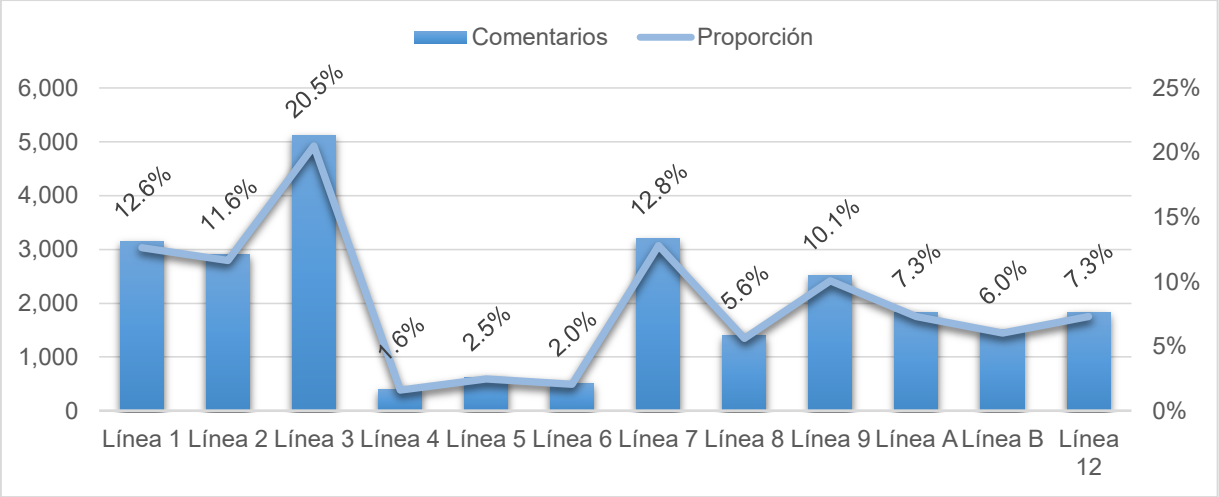
Para la clasificación de las estaciones se esperan 80,397 tweets clasificados de manera correcta, lo que equivale a un 93.31% de la base.

En ambas tablas, se señala que los grupos sin clasificación tienen la mayor proporción, en este caso al conocer la base de trabajo, es de esperarse esta distribución, pues la precisión de los tweets no es la adecuada para poder concluir en algún otro grupo ya que la idea que expresan es muy ambigua, por lo que no es posible conocer el contexto de dicho comentario, así mismo hay tweets que el mismo metro publica de manera recurrente que no proporciona información de interés para este estudio.

Omitiendo el grupo “Sin clasificar” en la tabla 7 de las líneas del metro se observa que la línea con mayor cantidad de menciones es la línea 3 con 5,097, seguida de la línea 7 con 3,191; posteriormente aparece la línea 1 y 2 con 3,141 y 2,898 respectivamente

Después de este resultado es posible tomar la proporción de la afluencia en el Gráfico 1 y la de comentarios del Gráfico 5, de esta forma es posible realizar una comparación entre la afluencia y las menciones de la base estudiada.

Gráfico 5 - Tweets por línea del metro



Conforme la afluencia que existe en este periodo de tiempo, lo esperado sería tener una distribución similar entre estos 2 datos, aunque hay 4 casos en los que la proporción de menciones es mayor a la de afluencia, en el Gráfico 6 se observa que la mayor proporción en comentarios es la línea 3 con un 20.48%; el segundo puesto en menciones tiene un 12.82%, pero este dato no corresponde a ninguna de las líneas con mayor afluencia si no a la línea 7, que en comparación de su afluencia tiene una diferencia en porcentaje de casi el doble, lo que puede indicar que es un foco rojo y se debe prestar mayor atención a las situaciones que suceden en esta línea. Posteriormente se encuentra la línea 9 con 10.11%, y finalmente la Línea A con 7.30% que no es mucho mayor al 6.59% de su afluencia.

Gráfico 6 - Afluencia vs Menciones

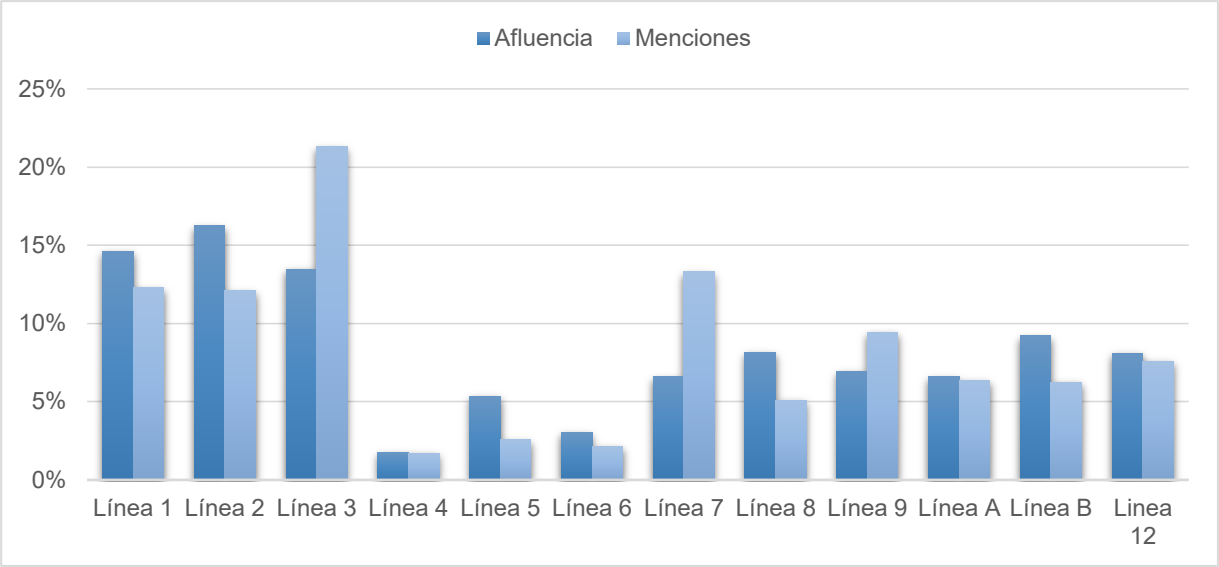


Tabla 10 - Afluencia vs Menciones

Línea	Afluencia	Menciones
Línea 1	14.63%	12.29%
Línea 2	16.25%	12.08%
Línea 3	13.48%	21.30%
Línea 4	1.76%	1.67%
Línea 5	5.33%	2.56%
Línea 6	3.05%	2.12%
Línea 7	6.58%	13.30%
Línea 8	8.14%	5.10%
Línea 9	6.92%	9.43%
Línea A	6.59%	6.32%
Línea B	9.20%	6.25%
Linea 12	8.08%	7.58%

Para conocer un poco más sobre lo contenido en la clasificación de los grupos, se extrajeron las frecuencias de las palabras con el fin de conocer de qué es lo que se habla, de este ejercicio se omitió el grupo “Sin clasificar”. Adicional al top 10 de las palabras, se mostrará una gráfica del tipo *word cloud*, el principio de la misma es que a mayor frecuencia, mayor es el tamaño que ocupa en la ilustración, de esta forma se puede tener un mayor panorama de las palabras

contenidas en cada uno de los grupos sin tener que colocar todas en una tabla, es importante mencionar que no todo el vocabulario involucrado en cada grupo se muestra en la gráfica.

freq = sort(colSums(as.matrix(dtm_servicio)), decreasing=T)

wfwc = data.frame(Palabra=names(freq), Frecuencia=freq)

wfwc = subset.data.frame(wfwc,wfwc$Frecuencia>=300)

wc_servicio = wordcloud2(head(wfwc,200),size=.4,shape='circle',color='random-dark', backgroundColor = 'wine')

Iniciando por el mayor grupo “Servicio”, se muestra el top 10 mediante la siguiente tabla.

Tabla 11 - Frecuencia grupo Servicio

Palabra	Frecuencia
minuto	5,490
mas	4,238
tren	4,063
estacion	3,925
servicio	3,895
lento	2,996
lineatres	2,881
avance	2,502
min	2,456
tarde	2,266

Ilustración 4 - Palabras grupo Servicio

Con base en la frecuencia de las palabras, se puede suponer que la línea que tiene un mayor problema sobre el servicio, es la línea 3, ya que aparece en este top, así mismo 5 de las 10 palabras hacen alusión a la tardanza en el servicio.

Como había mencionado no se muestra el vocabulario del grupo, en esta ilustración se colocó una frecuencia mayor a 300. En la Ilustración 4, es relevante señalar que además de la línea 3, aparece con un tamaño considerable la línea 1, 2, 7 y 9; poniendo mayor atención es posible visualizar estaciones específicas, como son Pantitlán, Tacubaya e Indios Verdes, dos de estas estaciones son conexiones con otras líneas y de hecho la que tiene un mayor número es Pantitlán con 4 líneas (1, 5, 9 y A) y Tacubaya con 3 líneas (1, 7 y 9), por otra parte Indios Verdes es una estación terminal de la línea 3.

Dentro de esta misma ilustración adicional al tema del avance del mismo transporte, aunque no es a gran tamaño, es posible observar palabras como escalera, limpieza y personal, palabras relacionadas con el servicio de los servidores públicos y de la infraestructura de las estaciones.

Pasando al grupo de “Comercio”, las frecuencias son las siguientes

Tabla 12 - Frecuencia grupo Comercio

Palabra	Frecuencia
ambulantaje	2,105
vendedores	1,900
policia	1,566
comercio	1,507
informal	1,350
seguridad	1,178
vagonero	1,126
vigilancia	1,117
vagon	1,024
lineados	932

Ilustración 5 - Palabras grupo Comercio

Para este grupo, al ser un poco más objetivo, se puede determinar que existe un problema importante de comercio informal, siendo este en mayor proporción en la línea 2 ya que aparece esta línea dentro del top 10.

La Ilustración 5 fue construida mostrando una frecuencia mínima de 100, en este grupo existe una particularidad con el grupo que habla sobre seguridad, aparecen palabras como policía, vigilancia, seguridad, palabras más relacionadas con seguridad que con comercio, las menciones contienen ambas palabras y he aquí la importancia de tener un buen conocimiento de la base para poder colocar el vocabulario específico en la base clasificadora, pudiendo realizar una clasificación de forma exitosa; estas palabras están en este grupo ya que los cuerpos de seguridad, al visualizar algún tipo de comercio no regulado dentro de las instalaciones deberían actuar para solicitar que esta gente se retire o presentarlos ante las autoridades correspondientes

En esta ilustración, además de que aparece la línea 2 con un buen tamaño, la línea 1, 3, 8 y 9 es posible visualizarlas; algunas de las palabras con mayor tamaño son ambulantaje, vagonero y vagón, lo que nos puede dar una idea de que el mayor problema que existe en el comercio informal se presenta dentro de los vagones, sin embargo con menor tamaño se encuentran palabras como zona y escalera, que puede referirse a este comercio pero en las inmediaciones de las estaciones.

Para el grupo de “Seguridad”, se obtuvo la siguiente información.

Tabla 13 - Frecuencia grupo Seguridad

Palabra	Frecuencia
seguridad	1,263
mas	1,127
policia	1,044
vagon	661
dinero	507
denuncia	498
limpieza	492
vigilancia	465
estacion	459
delincuencia	454

En este grupo, no aparece una referencia geográfica en el top 10. Retomando lo que se señala en la situación del grupo anterior, éste también contiene palabras como seguridad, policía y vigilancia, por lo que las demás palabras en este top las comparte con dinero, denuncia y

delincuencia, que son mayormente relacionadas con el tema de inseguridad que se vive en la Ciudad de México.

La frecuencia mínima con la que está construida la Ilustración 6 es con un mínimo de 90, en ésta es posible visualizar que aparece con un tamaño pequeño la línea 2 y 3.

Ilustración 6 - Palabras grupo Seguridad

Es notable el tamaño de 2 palabras, mas y seguridad, una de ellas muestra la necesidad que los usuarios expresan al solicitar de manera prioritaria contar con mayor seguridad, adicionalmente, es posible observar palabras como vendedores o vagonero, más enfocadas al comercio pero el hecho de que estas palabras aparezcan da información de que posiblemente ellos se identificaban como vendedores y al final delinquían.

Como se mencionó en el capítulo anterior, una de las situaciones con las que la clasificación de texto se enfrentaba era el de dar sentido a las palabras que se ocupan para clasificar, en este grupo aparece la palabra limpieza que está incluida en el top 10, ésta misma se encuentra en el grupo de servicio, pero en éste el enfoque es sobre la suciedad que existe en las unidades, por el otro lado, en este grupo de seguridad se menciona de forma figurativa, donde se pide que haya una “limpieza”, es decir que se inhiba la delincuencia.

Por último se mostrará el grupo de “Falla”, cabe recordar que este grupo es sobre fallas reportadas en la cuenta oficial del metro y que son atribuidas al servicio que brinda el STC Metro.

Tabla 14 - Frecuencia grupo Falla

Palabra	Frecuencia
tren	483
marcha	406
movilidadcdmx	380
avisometro	378
continua	279
minuto	266
toma	259
servicio	257
vias	234
mecanismo	231

Ilustración 7 - Palabras grupo Falla

Para este último grupo, se pueden señalar los *hashtag* que ocupa la cuenta oficial para hacer estos anuncios, #avisometro y #movilidadcdmx, mismos que aparecen en el top 10.

Para la Ilustración 7, se ocupó una frecuencia mínima de 20 palabras, las líneas visibles son la línea 2 y 3, así como la estación zapata, que es la conexión entre la línea 3 y 12.

Este grupo posiblemente no dé gran cantidad de información para este estudio, siendo un muy buen ejemplo sobre la importancia de los *hashtags*; en la ilustración es posible observar palabras como lento, puerta, minuto, seguridad, servicio, éstas son propias de los grupos

anteriormente descritos, cabe destacar que al llevar todas el acompañamiento de los *hashtag* #avisometro y #movilidadcdmx da como resultado una manera más sencilla para clasificar.

Para concluir este capítulo, se puede mencionar que es posible encontrar el foco de atención entre las distintas líneas, así como el tema que debería tener mayor prioridad de atención en el STC Metro. A pesar de que una gran proporción de menciones se fueron a un grupo que no se estudió, es importante señalar que hay veces que la gente simplemente se queja sin poder concretar una idea específica sobre cuál fue su problemática.

Capítulo 4. Conclusiones

4.1 Conclusiones del modelo

Este modelo es muy versátil y al estar basado en el cálculo de probabilidades se tiene la certeza de tener un buen funcionamiento. Para aplicarlo, basta tener un conjunto de documentos, ya sean tweets, libros, frases o cualquier archivo que contenga información en el formato de texto, el principal objetivo es contar con una clasificación, además de tener un conocimiento previo del contexto de la información, y de esta forma poder obtener información importante sin la necesidad de haber leído cada uno de estos archivos.

Posteriormente al realizar el desarrollo del modelo en este estudio, hay ciertos puntos que se deben tener en cuenta para poder elegir la aplicación de este:

1) Conocer el objetivo del estudio que se va a realizar.
2) Temporalidad con la que se estará realizando el estudio.
3) Definir de forma muy estricta los grupos de la clasificación.

El punto de convergencia de lo anterior es el tiempo que se dedica a poder encontrar el vocabulario necesario para realizar una buena separación entre cada uno de los grupos, por lo que en caso de que se requiera el estudio de forma constante, es un buen modelo. Adicionalmente, al ser un modelo supervisado, se podrá tener la clasificación enfocada a las necesidades del estudio.

Por otro lado, si se siguen de manera estricta los puntos mencionados, al ingresar nueva información al modelo, se tendrá la certeza que éste trabajará de forma correcta y no demorará demasiado en obtener la clasificación.

Para efectos de este estudio se realizaron 2 clasificaciones, una para conocer los temas que los usuarios expresan y la otra para ubicar en donde existen dichas quejas, cabe aclarar, que es posible llegar a niveles más profundos de clasificación, esto depende del detalle que se requiera, y podrá aplicar siempre y cuando se tenga el vocabulario necesario para realizar la clasificación correspondiente.

4.2 Conclusiones generales del trabajo

De acuerdo al estudio se puede concluir que las quejas de los usuarios conforme a las frecuencias de los distintos grupos, se encuentran en mayor medida dentro del grupo de servicio, específicamente a la lentitud en el avance de los trenes. Esta lentitud, aunque es atribuible al STC Metro por no brindar de forma correcta el servicio, este mismo no da ninguna

explicación del motivo por el que se presenta esta lentitud. Adicional a esto es posible ver que las líneas en donde se localizan estas insuficiencias son dentro de las líneas 1, 2, 3, 7 y 9, donde las primeras tres son las que tienen un mayor número de afluencia, por lo que un modelo como este puede ayudar para identificar a mayor detalle las insuficiencias con las que cuenta el STC Metro.

En lo que respecta al comercio dentro de las instalaciones, la mayor queja son los vagoneros, este problema de venta informal existe en todas las líneas del metro, sin embargo los usuarios se quejan en gran medida dentro de las líneas 1, 2 y 3, tomando en cuenta que son las de mayor afluencia, tener a un vendedor en el poco espacio que queda en el andén hace que la gente se queje de esta situación; pero es muy difícil erradicarlo ya que en muchos casos la delincuencia organizada está detrás de estos vendedores.

En general, la inseguridad es un tema constante dentro de la Ciudad de México, las quejas recurrentes van dirigidas a la línea 2 y 3, desafortunadamente esta práctica se lleva a cabo en todas las líneas, es preocupante que dentro de las instalaciones haya robos en los vagones, así como asaltos y acoso a las mujeres. En este rubro, los usuarios ocuparon 2 palabras, que son las más frecuentes, donde solicitan mayor seguridad por lo que para el STC Metro tendría que ser una prioridad poder salvaguardar la seguridad de los usuarios que ocupan a diario este transporte.

Ante la frecuencia con la que aparecen las líneas 1, 2 y 3 dentro de cada grupo, es posible señalar la existencia de un problema general dentro de las mismas, esto se puede atribuir a que son las más concurridas; cabe destacar que en el grupo de fallas, por la característica del mismo, la mayor frecuencia en la ubicación se centró dentro de las líneas 2 y 3, es decir existió un mayor número de fallas atribuidas al STC Metro en estas líneas.

Además de la información recabada en los primeros tres grupos, este último nos ayuda a dar una conclusión más certera de donde existen las deficiencias en el servicio del STC Metro. Las autoridades correspondientes deben dar mayor atención a las líneas 2 y 3 para realizar un mantenimiento en todos sus rubros de forma prioritaria.

A pesar de que los grupos de comercio y seguridad son importantes, parece ser que para los usuarios del metro les resulta importante contar con un servicio de calidad, donde se puedan desplazar de forma eficiente a sus destinos. Posiblemente ante el crecimiento de la zona urbana del Estado de México y la Ciudad de México, se esté considerando ampliar la infraestructura de las diferentes líneas, no obstante esta carga podría ser un mayor problema si no se arregla la problemática de movilidad que ya existe.

Bibliografía

Libros

- Lantz, B. (2015). *Machine Learning with R*, Birmingham, UK, Packt Publishing Ltd.
- Tonkin, E., Tourte, G. (2014). *Working with Text*, Cambridge, UK, Chandos Publishing.
- Aggarwal, C. (2015). *Data Mining*. Editorial Springer.
- Russell, S. (2004). *Inteligencia Artificial, un enfoque moderno*. Madrid, España: Pearson Prentice Hall.
- Pavan, B (2012), *Twitter: 5 años*, Hipertextual S.L.
- Blancas, S., Hernández, M.,Arellano, D. (2017) *Decisiones e implementación en la construcción de las primeras once líneas de la red del Metro en la Ciudad de México.*
- Moreno Galván, R. (2013). *Construcción de acceso de línea 12 con línea 2 en la estación Ermita del Metro.* (Tesis de licenciatura, Universidad Nacional Autónoma de México)
- Jurafsky, D., Martin, J. (2018). *Speech and Language Processing*, Pearson International Edition.

Páginas web

- https://metro.cdmx.gob.mx.
- https://marketing4ecommerce.net/historia-de-twitter/
- https://www.ig.com/es
- https://www.ig.com/es/ideas-de-trading-y-noticias/noticias-acciones/cambiando-las-reglas-del-juego--la-historia-de-twitter-190531#information-banner-dismiss