MCDI

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO
GERENCIA DE CAPITAL HUMANO
POSGRADOS

"IDENTIFICACIÓN Y ANÁLISIS DE QUEJAS EN TWITTER DE LOS PRINCIPALES BANCOS EN MÉXICO DE 2018 A 2019 MEDIANTE TÉCNICAS DE MINERÍA DE DATOS Y RECUPERACIÓN DE INFORMACIÓN"

REPORTE ANALÍTICO DE EXPERIENCIA LABORAL
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Daniel Antonio Armas Texta

Asesor:

Dr. Luis Guillermo Ruiz Velázquez

Asesor:

Dr. Eric Sadit Téllez Avila

Ciudad de México, Diciembre, 2020

CONACYT

INFOTEC POSGRADOS

INFOTEC
POSGRADOS

AUTORIZACIÓN DE IMPRESIÓN Y NO ADEUDO EN BIBLIOTECA

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

Ciudad de México, 17 de mayo de 2021
INFOTEC-DAIC-GCH-SE-160/2021.

La Gerencia de Capital Humano / Gerencia de Investigación hacen constar que el trabajo de titulación intitulado

IDENTIFICACIÓN Y ANÁLISIS DE QUEJAS EN TWITTER DE LOS PRINCIPALES BANCOS EN MÉXICO DE 2018 A 2019 MEDIANTE TÉCNICAS DE MINERÍA DE DATOS Y RECUPERACIÓN DE INFORMACIÓN

Desarrollado por el alumno **Daniel Antonio Armas Texta** y bajo la asesoría del **Dr. Luis Guillermo Ruiz Velázquez** y el **Dr. Eric Sadit Téllez Avila**; cumple con el formato de biblioteca. Por lo cual, se expide la presente autorización para impresión del proyecto terminal al que se ha hecho mención.

Asimismo se hace constar que no debe material de la biblioteca de INFOTEC.

Vo. Bo.

_____

Lic. Susana Argelia Salomón Jalili
Coordinadora de Biblioteca

Anexar a la presente autorización al inicio de la versión impresa del trabajo referido que ampara la misma.

C.p.p Servicios Escolares

Agradecimientos

Agradezco a Dios por haberme dado la oportunidad de superarme, a mi padre, Antonio Armas, cuyos sabios consejos me brindaron la disciplina para aprovecharla y a mi madre, Amparo Texta, cuyo ejemplo de vida me otorgó la confianza, el coraje y la determinación para lograrlo.

A mi *«Comita»*, Ricarda Carbajal, quien cariñosamente me quitó las cargas del día a día y a mi hermano de sangre, Diego Armas, cuya ambición y hambre de conocimiento me impulsó a buscar nuevos retos.

Agradezco a mis hermanos por elección, Iván Domínguez *«Trolo»*, Jesús Méndez *«Gigio»* y Roberto Flores *«Betito»* quiénes con su humor y afecto me dieron la fuerza y los ánimos en los momentos que más lo necesitaba.

Agradezco a mi querido amigo y mentor, Mario Hernández, cuya experiencia orientó mi camino, además de darme la facilidad y confianza para estudiar y trabajar a la vez.

A mis abuelitos Antonio Armas y Alicia Pluma, quienes pacientemente han esperado este momento.

Por último, agradezco a mi estimado compañero de generación Miguel Álvarez, por vivir esta experiencia junto a mí y ser tan buena mancuerna.

¡Muchas gracias a todos, sin ustedes no hubiera podido lograrlo!

Tabla de contenido

Índice de figuras

Índice de cuadros

Introducción

Los bancos son instituciones financieras que generan ganancias a través del otorgamiento de créditos. Cada individuo es libre de escoger a la institución financiera que más le agrade, es por esto que los bancos deben competir entre ellos para atraer a la mayor cantidad de clientes.

De acuerdo con [Hussain and Prieto, 2016] una aplicación potencial de técnicas de analítica avanzada para la industria financiera es la evaluación de la exposición al riesgo reputacional relacionado con los servicios ofrecidos por los bancos a sus clientes. Ya que una percepción negativa puede disminuir la capacidad de un banco para mantener las relaciones comerciales existentes, establecer nuevas relaciones o el acceso continuo a fuentes de financiamiento.

Por otro lado, [Culnan et al., 2010] muestra que el surgimiento de plataformas como *Twitter* o *Facebook* han generado comunidades *online* con intereses acerca de una empresa, marca o producto. Dichas comunidades impactan directamente a las marcas, las ventas, servicio al cliente, soporte y desarrollo de nuevos productos.

A través de técnicas de recuperación de la información, procesamiento de lenguaje natural y aprendizaje no supervisado, fue posible explorar los *tweets* correspondientes a los bancos más grandes de México en búsqueda de las principales quejas de los usuarios de servicios financieros de 2018 a 2019. Dentro de los hallazgos destacan las quejas al realizar/recibir transferencias electrónicas, las constantes llamadas para ofrecer productos y servicios así como los cargos no reconocidos.

Este trabajo es una muestra de las capacidades que se pueden implementar y los resultados que se pueden obtener al explorar fuentes de datos semi-estructuradas de texto libre.

Resumen

El presente proyecto busca determinar las principales quejas de los usuarios de servicios financieros en México a través de la explotación contenida en texto libre. Para lograrlo, se escogieron a los siete bancos más grandes de México (conocidos como G7). Con ayuda del `API` de *Twitter* y mediante técnicas de recuperación de información se extrajeron los *tweets* correspondientes a cada una de estas instituciones de 1 de enero 2018 a 31 diciembre 2019.

Una vez que se obtuvieron los datos de *Twitter*, se hizo uso de técnicas de procesamiento de lenguaje natural y aprendizaje no supervisado para detectar los puntos calientes de los *tweets* y así inducir las quejas de los usuarios de servicios financieros.

Las quejas fueron analizadas tanto de forma temporal (hora, día, trimestre y año) como de forma espacial (país, municipio) y por institución financiera. Se realizaron comparaciones por tipo de queja y banco para determinar tanto las fortalezas como las áreas de oportunidad de cada institución.

Por último, se realizaron comparaciones con los resultados reportados por la Comisión Nacional para la Protección y Defensa de los Usuarios de los Servicios Financieros (CONDUSEF) para determinar las convergencias y divergencias en la información obtenida tanto por denuncias formales como en redes sociales (*Twitter*).

Contexto

En palabras del Banco de México [BANXICO, 2018], el sistema financiero mexicano está constituido por un conjunto de instituciones que captan, administran y canalizan a la inversión, el ahorro tanto de nacionales como de extranjeros.

En México, existen diferentes organismos que regulan la actividad bancaria, de los cuales destacan:

- El **Banco de México (BANXICO)** tiene el objetivo prioritario de preservar el valor

de la moneda nacional a lo largo del tiempo y, de esta forma, contribuir a mejorar el bienestar económico de los mexicanos [BANXICO, 2019].

- La **Comisión Nacional Bancaria y de Valores (CNBV)**, la cual es un órgano desconcentrado de la Secretaría de Hacienda y Crédito Público (SHCP), con facultades en materia de autorización, regulación, supervisión y sanción sobre los diversos sectores y entidades que integran el sistema financiero en México [CNBV, 2018].
- La **Comisión Nacional para la Protección y Defensa de los Usuarios de los Servicios Financieros (CONDUSEF)** se dedica a promover y difundir la educación y la transparencia financiera para que los usuarios tomen decisiones informadas sobre los beneficios, costos y riesgos de los productos y servicios ofertados en el sistema financiero mexicano [CONDUSEF, 2018].

Dentro de sus actividades, la CNBV genera una designación de instituciones para identificar a las entidades de crédito, cuya quiebra potencial pudiera afectar la estabilidad del sistema financiero o de la economía del país, a fin de requerirles, en función de su nivel sistémico, un suplemento de capital. En 2018, la Comisión Nacional Bancaria y de Valores ratificó a los **siete bancos más grandes que operan en México** [Juárez, 2018]. A este conjunto de instituciones financieras se les conoce cómo **G7** y está compuesta por los siguientes bancos:

- BBVA
- Citibanamex
- Banorte
- Santander
- HSBC
- Scotiabank
- Inbursa

Problemática

De acuerdo con cifras de la CONDUSEF, los bancos tocaron récord en reclamaciones al cierre del 2018. Las reclamaciones de los bancos sumaron 9.4 millones, un incremento de 6.3 % a tasa anual, según cifras del Buró de Entidades Financieras. Por número de reclamaciones, Citibanamex, Santander, BBVA, Banorte y HSBC (que forman parte del G7) concentraron 81.1 % de las quejas totales del sector [Expansión, 2019].

Además [Hussain and Prieto, 2016] propone que una aplicación potencial de técnicas de analítica avanzada para la industria financiera es la evaluación de la exposición al riesgo reputacional relacionado con los servicios ofrecidos por los bancos a sus clientes. Ya que una percepción negativa puede disminuir la capacidad de un banco para mantener las relaciones comerciales existentes, establecer nuevas relaciones o el acceso continuo a fuentes de financiamiento.

México es un país donde una buena parte de la población hace uso de servicios financieros [Orozco, 2011]. Existen diversas entidades regulatorias que vigilan la actividad de los bancos para asegurar que estos laboren bajo las mejores prácticas ofreciendo valor a sus clientes. Sin embargo, pocas veces se explota lo que sucede fuera de estos medios y desaprovechando todo el potencial que generan las redes sociales.

Mediante herramientas de analítica avanzada, se pueden aprovechar los datos generados día a día en las redes sociales para apalancar estrategias que reduzcan el número de quejas, mejoren la experiencia de usuario, lo cual se traduce en fidelidad, posicionamiento de marca y rentabilidad para las empresas.

Objetivos

Objetivo General

Derivado de lo anterior, el objetivo del presente proyecto es **determinar y analizar las quejas en *Twitter* de los usuarios de servicios bancarios que integran el G7 y su con-**

traste con la información reportada por la CONDUSEF.

Objetivo Específico

Aplicar técnicas de recuperación de la información y aprendizaje computacional a mensajes obtenidos en *Twitter* para la detección de patrones en el comportamiento y distribución de quejas de los principales bancos en México. Para poder alcanzar el objetivo deseado, se visualizan tres objetivos específicos:

1. Realizar la extracción y estandarización de *tweets* relacionadas a quejas de los bancos que pertenecen al G7 de enero 2018 a diciembre 2019.
2. A los *tweets* obtenidos, aplicar técnicas de aprendizaje computacional desde un entorno no supervisado para la detección de tópicos y así determinar que aqueja al cliente.
3. Una vez generados los tópicos, realizar análisis de tendencias y distribución de las quejas por institución, fecha y zona geográfica, además de explorar la relación con los reportes generados por la CONDUSEF.

Es importante mencionar que la metodología seguida se puede implementar para cualquier empresa o industria que tengan interacciones con sus clientes por medio de redes sociales. En consecuencia, al replicar este desarrollo se pueden obtener análisis bajo demanda de sus propias quejas (o de sus competidores), lo cual puede brindar una ventaja estratégica dentro de la industria.

Estructura del documento

El primer capítulo está dedicado a introducir y contextualizar el problema, alcance y objetivos del manuscrito.

En el segundo capítulo se abordará el *estado del arte* referente a la recuperación de la información (*IR* por sus siglas en inglés[1]) poniendo especial atención a los índi-

[1] *IR: Information Retrival*

ces invertidos, representación vectorial del texto e identificación de tópicos en fuentes de datos semi-estructuradas. Una vez teniendo comprendidas dichas bases, se podrá realizar su aplicación en este caso de estudio.

Durante el tercer capítulo se presenta el trabajo relacionado con este proyecto. Se comparan diferentes autores y resultados que dan la guía y justificación de los las técnicas ocupadas a lo largo de este proyecto.

A lo largo del cuarto capítulo se presentará la metodología utilizada, recorriendo cada una de sus fases para la correcta aplicación de técnicas de analítica avanzada a un problema del día a día. Se aborda con especial atención la selección y extracción de datos, el preprocesamiento del texto, la representación vectorial, así como la aplicación de técnicas de aprendizaje no supervisado para la detección de tópicos.

Durante el quinto capítulo se analizarán los resultados obtenidos yendo de lo general a lo particular, comparando los resultados entre bancos, a través del tiempo y por zona geográfica. También se contrastarán los resultados con los datos reportados por la CONDUSEF en sus reportes trimestrales, observando similitudes y prestando principal atención a la información complementaria exclusiva de las redes sociales.

Para finalizar, después de analizar los resultados, en el último capítulo se presentarán las conclusiones del presente trabajo, mostrando las diferencias entre los resultados obtenidos y los resultados oficiales, contrastando las implicaciones de cada una y mostrando tanto el valor agregado como la riqueza que ofrece generar este tipo de explotación de fuentes semi-estructuradas de datos.

Capítulo 1

Marco Teórico

Capítulo 1. Marco Teórico

1.1. Recuperación de Información

1.1.1. Datos e información

Antes de explicar cuál es el objetivo de la *recuperación de información* es necesario **determinar las diferencias entre datos e información**. Para lograr definir dichos límites, se ocupará la *jerarquía de la sabiduría* propuesta por [Rowley, 2007] la cual es uno de los modelos fundamentales en la literatura de *Sistemas de información y administración del conocimiento.*

La jerarquía de la sabiduría (también conocida como jerarquía *DIKW* por sus siglas en inglés[1]) es usada para contextualizar datos, información, conocimiento y a veces sabiduría respecto uno del otro. Se asume implícitamente que los datos pueden ser usados para crear información; información puede ser usada para crear conocimiento y el conocimiento puede ser utilizado para crear sabiduría [Restrepo and Pacheco, 2019].

De acuerdo con [Rowley, 2007] la diferencia entre datos, información, conocimiento y sabiduría está dada por:

1. **Datos**. Se definen como símbolos que representan propiedades de objetos, eventos y su contexto. Es producto de la observación. Son inútiles hasta que son usados. Por lo anterior, **la diferencia entre datos e información es funcional, no estructural.**

2. **Información**. Contiene descripciones, respuestas a preguntas que inician con palabras como «quién», «qué» «cuándo» entre otras. Los sistemas de información generan, almacenan, recuperan y procesan datos. En otras palabras, **la información se deriva de los datos.**

[1] *DIKW: Data, Information, Knowlegde and Wisdom*

3. **Conocimiento**. Es el «saber cómo» y hace posible la transformación de la información en instrucciones. El conocimiento puede ser adquirido de un ente que lo posee a otro mediante instrucciones o directamente de la experiencia.

4. **Sabiduría**. Es la habilidad de aumentar la efectividad del conocimiento. La sabiduría agrega valor lo cual requiere de ética, juicio y estética lo cual implica que es inherente a la persona, único y personal.

Figura 1: Representación gráfica de la jerarquía de la sabiduría. Elaboración propia basada en [Rowley, 2007]

En la Figura 1 se puede apreciar la relación entre datos, información, conocimiento y sabiduría. Es importante mencionar que la base de todo son los datos, a partir de los cuales creamos información mediante transformaciones de los mismos.

A grandes rasgos, podemos clasificar los tipos de datos por su manera de estructurarse como:

- **Estructurados**. Son aquellos cuyo contenido es fijo y puede describirse como una tupla de datos. Este tipo de colecciones de datos suelen organizarse en tablas.

- **No estructurados.** Suelen ser textos escritos en un formato libre en algún lenguaje humano. Los datos no estructurados son llamados documentos, y las colecciones son conocidas como corpus de documentos. Las colecciones pueden estar organizadas de diferentes formas para facilitar su acceso a documentos individuales.

- **Semi estructurados.** suelen estar descritos en lenguajes especializados para tal fin. Entre los lenguajes más comunes para para definir este tipo de datos se encuentran JSON y XML. Tienen una estructura regular pero variable, suelen ser

auto-descritos. En bases de datos es común encontrar conjuntos de pares `llave→valor` pues son los motores más comunes en bases de datos *No-SQL*.

Por tanto, la *recuperación de información* buscará explotar los diferentes tipos de datos para extraer la información que hay en ellos.

Es importante precisar que, en el caso del presente proyecto, la fuente de datos es semi-estructurada (archivo tipo JSON), sin embargo la mayoría del análisis es sobre información no estructurada (*mensajes escritos en texto libre*).

Recuperación de información

De acuerdo con [Baeza-Yates et al., 1999], la recuperación de información (*IR* por sus siglas en inglés[2]) se encarga de la representación, almacenamiento, organización y acceso a los elementos de información (usualmente texto libre). La representación y organización de dichos elementos deben proveer al usuario un fácil acceso a la información que está interesado. Dada una consulta, la meta de los sistemas de recuperación de información es regresar la información que pueda ser **útil o relevante** para el usuario.

Se debe hacer énfasis que recuperar información no es lo mismo a recuperar datos. Los sistemas de recuperación de datos (*DR* por sus siglas en inglés[3]) proveen una solución al usuario de bases de datos, sin embargo no resuelven el problema de obtener información acerca de un tema o tópico.

Para poder ser eficientes en satisfacer la necesidad del usuario, los sistemas de recuperación de información deben de (alguna forma) **«interpretar»** el contenido de los objetos de información (documentos) en una colección y calificarlos de acuerdo a la relevancia que este pudiera tener para el usuario. Esta interpretación de documentos involucra la extracción sistemática y sistémica de la información del texto de los documentos además de su uso en la determinación de la relevancia para el usuario. Por lo tanto, **el concepto de relevancia es el centro de la recuperación de información**.

[2] *IR: Information Retrival*
[3] *DR: Data Retrival*

A lo largo del presente proyecto se hará énfasis en temas relacionados al *acceso y organización de documentos*, poniendo especial atención a estructuras de datos que permitan realizar búsquedas de forma eficiente como **índices invertidos** y técnicas de agrupación para **identificación de tópicos.**

1.2. Búsqueda

En palabras de [Manning et al., 2008], *«la IR es encontrar material (usualmente documentos) de naturaleza no estructurada (usualmente texto) que satisfaga la necesidad de extracción de la información dentro de largas colecciones de documentos»*. Por tanto, la *búsqueda* de dicha información se convierte en parte fundamental de la *IR*.

Para poder realizar búsquedas eficientes es necesario contar con un arreglo ordenado, es decir, dado un arreglo $A[1,n] = a_1, a_2, ..., a_n$ encontrar una permutación π tal que $\pi(A[1,n]) = a_{\pi(1)}, a_{\pi(2)}..., a_{\pi(n)}$ satisfaga que $a_{\pi(i)} < a_{\pi(i+1)}\ \forall\ i \in \{1,2,...,n-1\}$. Una vez generado dicho orden, es necesario generar estructuras de datos que permitan explotar el orden generado.

Esta sección está orientado a describir los algoritmos que son el motor de las máquinas de búsqueda. En particular, se presenta la estructura de un *índice invertido en modelo binario y en listas de posteo.*

1.2.1. Índices Invertidos

De acuerdo con [Yan et al., 2009], un **índice invertido** para una colección de documentos es una estructura de datos que almacena, para cada término (palabra) que aparece en algún lugar de la colección, información sobre las ubicaciones donde ocurre. En particular, para cada término w, el índice contiene una *lista invertida* I_w la cual consiste en una *serie de índices* de las publicaciones. Cada publicación en I_t contiene información sobre el apariciones de w en un documento en particular d, generalmente la información contenida es el *ID* del documento (*docID*), el número de ocurrencias de w en d (la frecuencia), y posiblemente otra tipo de información sobre las ubicaciones

de las ocurrencias dentro del documento y sus contextos. Las publicaciones de cada lista suelen ser ordenados por *docID*.

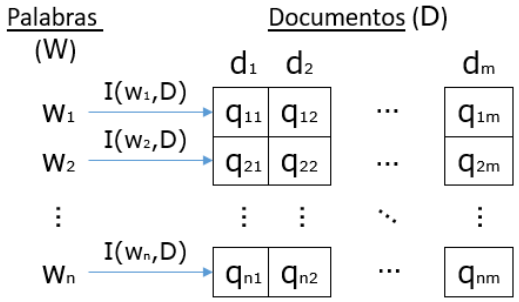


Figura 2: Representación gráfica de la estructura de un índice invertido. Elaboración propia.

La figura Figura 2 muestra la representación gráfica de la relación de un índice invertido I, las palabras W y la colección de documentos D, dónde q es la información contenida en el índice invertido. Dado lo anterior, un índice invertido se interpreta como es una estructura de datos de índice que almacena un mapeo del contenido, como palabras o números, a sus ubicaciones en un documento o un conjunto de documentos.

Existen diferentes tipos de índices invertidos dependiendo de la información contenida en ellos q. A continuación se presentaran las más populares: **modelo binario, modelo frecuencial** y **representación TF-IDF**.

1.2.2. Modelo binario

El modelo binario para un índice invertido consiste en considerar la información q cómo una representación de la existencia (o ausencia) de una palabra w en el documento d. Es decir,

$$q_{ij} = \begin{cases} 1 & \text{si la palabra } w_i \text{ está presente en el documentos } d_j \\ 0 & \text{en otro caso} \end{cases} \tag{1.1}$$

La figura Figura 3 muestra la representación gráfica de la relación de un índice invertido utilizando el modelo binario.

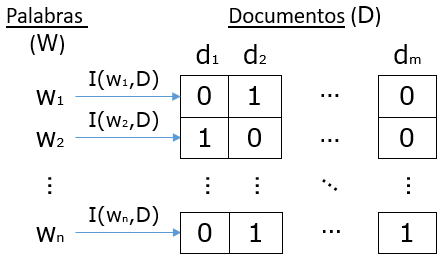


Figura 3: Representación gráfica de la estructura de un índice invertido utilizando el modelo binario. Elaboración propia.

Esta representación es útil para buscar documentos que contengan una palabra específica. Sin embargo al momento de realizar una búsqueda sobre un índice invertido binario, el orden en el cual se van a presentar los resultados está determinado por el orden adyacente a los documentos y se debe considerar para la recuperación de información es fundamental obtener *información relevante para el usuario.*

Para superar dicho problema, se puede utilizar definiciones alternas de q, por ejemplo: frecuencia.

1.2.3. Modelo frecuencial

El modelo frecuencial para un índice invertido consiste en considerar la información q la cantidad de veces que aparece una palabra w en el documento d. Es decir,

$$q_{ij} = \mathrm{N}(w_i, d_j) \tag{1.2}$$

Dónde: $\mathrm{N}(w_i, d_j)$ es el número de veces que la palabra w_i aparece en el documento d_j

La figura Figura 4 muestra la representación gráfica de la relación de un índice invertido utilizando el modelo frecuencial.

Con esta representación, el orden en el cual se van a presentar los resultados

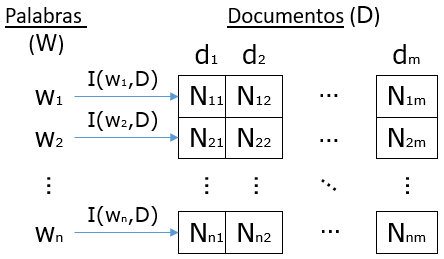


Figura 4: Representación gráfica de la estructura de un índice invertido utilizando el modelo frecuencial. Elaboración propia.

derivados de la búsqueda sobre el índice invertido está determinado por la cantidad de veces que aparece la palabra dentro del documentos. Ahora, es importante considerar que existen casos dónde una palabra se puede repetir varias veces dentro de un documento y esto sesgar el resultado de la búsqueda. Existe una técnica muy socorrida que ha mostrado ser de gran utilidad para sobrepasar este problema, la **representación *TF-IDF***

1.2.4. Modelo *TF-IDF*

De acuerdo con [Manning et al., 2008], ***TF-IDF*** (del inglés *Term Frequency – Inverse Document Frequency*), *frecuencia de término – frecuencia inversa de documento*, es una medida numérica que expresa **cuán relevante es una palabra para un documento en una colección**. Esta medida se utiliza a menudo como un factor de ponderación en la recuperación de información y la minería de texto. El valor *TF-IDF* aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos, lo que permite manejar el hecho de que algunas palabras son generalmente más comunes que otras.

El *TF-IDF* se calcula como:

$$\text{tfidf}_{ij}(w_i, d_j, D) = \text{tf}(w_i, d_j) \times \text{idf}(w_i, D)$$

Donde:

$$\text{tf}(w_i, d) = \frac{\text{f}(w_i, d_j)}{\text{máx}\{\text{f}(w, d_j) : w \in d_j\}}$$

$$\text{idf}(w_i, D) = \log \frac{|D|}{|\{d \in D : w_i \in d\}|}$$

Considerando a D como el número de documentos, t un término específico y d un documento específico.

Por tanto, la representación *TF-IDF* de un índice invertido se obtiene al utilizar la medida de relevancia $q_{ij} = \text{tfidf}(w_i, d_j, D)$.

1.2.5. Optimización de un índice invertido

Hasta ahora, se ha presentado un índice invertido como una matriz $n \times m$ dónde n es el tamaño del vocabulario, es decir, **la cantidad de palabras únicas en la colección de documentos** y m es la cantidad de documentos en la colección y cada entrada de la matriz, es el peso relacionado q.

Dado el gran tamaño de las colecciones de documentos así como del vocabulario, mantener una matriz de tales proporciones es inviable en la práctica. Basado en que no todos los términos se utilizan en todos los documentos, dicha matriz se puede representar de manera *dispersa*. Esta representación corresponde al vocabulario y una serie de listas ordenadas asociadas a cada término. Estas listas se conocen como *listas de posteo*.

Las listas de posteo son alternativas eficientes a todos los modelos de índices invertidos presentados en esta sección. Para el modelo binomial y frecuencial, se ignoran las entradas con valor 0, mientras que para una representación *TF-IDF* se pueden omitir los valores más pequeños, realizando un recorte de las listas de posteo.

La Figura 5 muestra la comparación entre una representación matricial y la representación dispersa (mediante listas de posteo) de un índice invertido.

En el día a día, las palabras `Hola`, `HOLA` y `hola` son «iguales» o «tienen el mismo significado» para la población en general. Sin embargo, desde un punto de vista

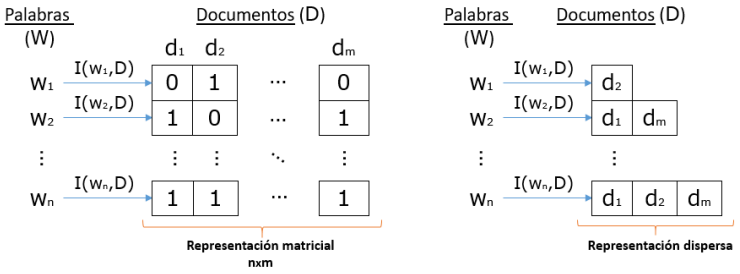


Figura 5: Comparación de la estructura de un índice invertido utilizando una representación matricial y una representación de listas de posteo. Elaboración propia.

computacional son diferentes pues cada una son cadenas de caracteres compuestas por diferentes símbolos. En consecuencia, antes de construir un índice invertido, es necesario modificar y reestructurar el texto contenido en los documentos, para que las herramientas computacionales los representen de la forma adecuada. Las técnicas asociadas a estas modificaciones se presentaran durante la siguiente sección.

1.3. Representación vectorial del texto

El proceso de recuperación de información requiere de una interacción directa entre el usuario y un sistema informático: mientras que el usuario tiene la necesidad de obtener información, el sistema informático debe traducir dicha necesidad en operaciones computacionales con el objetivo de regresar el mejor resultado al usuario y así satisfacer su necesidad. Desafortunadamente esta interacción entre el usuario y el sistema informático no es sencilla pues para dicho sistema, el texto es un conjunto de caracteres sin ningún sentido en particular, por tanto para realizar la traducción de la necesidad en términos de operaciones, es necesario transformar el texto en símbolos con significado para el sistema. No obstante, antes de realizar la traducción de texto a símbolos informáticos, es necesario saber si el texto de los documentos tiene «significado» para el usuario.

En la Figura 10 se muestra el proceso de representación vectorial del texto.

Por otro lado, es importante observar que *el índice invertido tiene como dominio el vocabulario inducido por la colección de documentos*, es decir, las **palabras únicas**

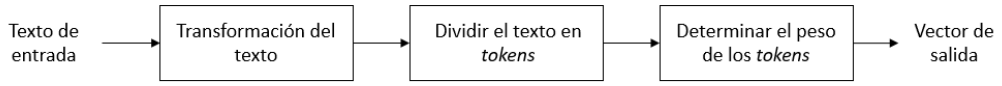


Figura 6: Tratamiento genérico del texto para obtener una representación vectorial. Elaboración propia basada en [Tellez et al., 2017]

que aparecen en los documentos. El vocabulario puede ser muy extenso lo cual puede derivar problemas computacionales para la manipulación de datos. Por lo anterior es importante utilizar técnicas para la reducción del vocabulario.

1.3.1. Preprocesamiento de texto

Al analizar textos en formato libre, existen diversos retos como: faltas ortográficas, signos de puntuación, uso de gentilicios, uso de jerga popular, entre otras. Cada uno de estos retos dificulta la transformación del texto a símbolos con significado computacional y por ende, dificulta el análisis y la recuperación de información a partir de ellos. Además, dentro de los textos, no todas las palabras son necesarias para transmitir una idea, por tanto el mantener dichas palabras generaría un vocabulario más grande del necesario.

Para lidiar con dicho problema, es necesario generar un preprocesamiento del texto con el fin de homologar las palabras así como eliminar elementos innecesarios. Utilizando como base a [Tellez et al., 2017] el preprocesamiento del texto que consta de los siguientes pasos:

- Cambiar el texto a minúsculas. De esta forma es posible homologar palabras como `Hola, HOLA` y `hOlA` a una sola cadena de caracteres: `hola`
- Eliminar acentos, signos de puntuación y caracteres especiales. Tanto los acentos, signos de puntuación y demás caracteres especiales sirven para transmitir de forma más clara y precisa una idea, sin embargo, en términos computacionales no son necesarios.
- Eliminar las «palabras vacías» (*stopwords*). De acuerdo con [Wilbur and Sirotkin, 1992], las *stopwords* son palabras que cumplen con una función gramatical pe-

ro carecen de significado en sí mismos. Los ejemplos más claros de este tipo de palabras en idioma español son los artículos, pronombres, preposiciones. Al eliminar dichas palabras, además de reducir el vocabulario se puede lograr mejorar la recuperación de información.

- Reducir a raíz (*stemming.*) Este proceso busca reducir las palabras a su raíz (o *lexema*) el cual es la parte fundamental de la palabra. Mediante este proceso se pueden homologar palabras como `Rojas, roja, rojas` y `rojo` a `roj`.
- Reducir a lema (*lemmatization*). Este proceso consiste en eliminar la conjugación, género y pluralidad de las palabras para hallar el lema correspondiente. Esta es la forma lingüística en que se encuentran las palabras del diccionario. Mediante este proceso se pueden homologar palabras como `frías, fríos, fría` y `frío` a `frío`.

Cada una de las técnicas anteriores son las más populares para el preprocesamiento del texto, sin embargo, dependiendo del objetivo del análisis, el idioma y los documentos se decide cuáles técnicas se aplican.

Una vez realizado el preprocesamiento del texto, se es posible realizar la traducción del texto a símbolos con sentido computacional. Esta traducción se realiza en términos de vectores y matrices. Durante las siguientes secciones se presentaran las principales representaciones vectoriales del texto.

1.3.2. Bolsa de palabras

Después de realizar el preprocesamiento del texto, se procede a partir el texto en unidades léxicas, generalmente palabras. A este proceso se le conoce como *tokenizado* y nos permite representar el texto como una bolsa de *tokens*.

Una de las formas más populares para generar la representación vectorial de los *tokens* es mediante una matriz A de dimensión $m \times n$, dónde m es la cantidad de documentos y n el tamaño del vocabulario. Cada una de las entradas a_{ij} contiene información sobre el documento d_i y la palabra w_j. A esta representación vectorial del

texto se le conoce como **bolsa de palabras** (*BoW* por sus siglas en inglés[4]).

$$A = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_m \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & ... & w_n \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix} \end{array}$$

Como se puede notar, existe una relación directa entre la representación del texto mediante una bolsa de palabras A y representación matricial de un índice invertido, definiendo $I := A^T$, en consecuencia, de forma análoga se puede utilizar la representación vectorial de bolsa de palabras para modelos **binario, frecuencial** y ***TF-IDF***.

Sin embargo, la representación por bolsa de palabras tiene dos desventajas: 1) no capturan la relación semántica de las palabras y 2) el tamaño de la matriz depende del tamaño del vocabulario, el cual puede ser muy extenso en documentos con texto libre. Afortunadamente, existen métodos modernos que no solo ayudan a reducir drásticamente la dimensión del espacio vectorial sino que también ayudan a capturar el contexto semántico de los documentos. A este tipo de representaciones se les conoce como: **palabras embebidas.**

1.3.3. Palabras embebidas

Las representaciones de textos presentadas hasta el momento, generan una dimensión por cada palabra en el vocabulario con lo cual se puede representar los documentos como vectores en grandes dimensiones. Citando a [Levy and Goldberg, 2014] al utilizar este tipo de técnicas (bolsa de palabras), los términos están definidos por vectores independientes entre sí, lo cual implica que son palabras no relacionadas aun cuando exista una relación gramatical entre ellas. En consecuencia, se busca una representación que capture la semántica y sintáctica de la similitud entre palabras.

[4] *BoW: Bag of Words*

Existen muchas representaciones de palabras, por ejemplo agrupando en clústers basados en su contexto ([Uszkoreit and Brants, 2008]) o utilizando descomposición en valores singulares ([Bullinaria and Levy, 2007]) sin embargo de acuerdo con los estudios realizados por [Turian et al., 2010], [Collobert et al., 2011], [Socher et al., 2011] y [Al-Rfou et al., 2013] las representaciones que han tenido un mayor éxito han sido aquellas llamadas *palabras embebidas* (*words embeddings*).

De acuerdo con [Liu et al., 2015], la representación mediante palabras embebidas juega un papel cada vez más vital en la construcción de vectores de palabras basados en sus contextos dentro de una gran colección de documentos. Este tipo de representación captura información semántica y sintáctica de las palabras, y se puede utilizar para medir similitudes de palabras, que son utilizadas ampliamente en diversas tareas de recuperación de información y procesamiento de lenguaje natural (*NLP* por sus siglas en inglés[5]). Además, permiten un cálculo eficiente de similitud entre palabras a través de operaciones matriciales de baja dimensión ([Levy and Goldberg, 2014]).

Uno de los ejemplos más famosos (en inglés) que permite entender la ventaja que ofrecen las palabras embebidas es el poder realizar la siguiente representación:

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

Es decir, el vector $\overrightarrow{queen}$ es parecido al vector $\overrightarrow{king}$ pero en mujer.

La generación de palabras embebidas parte de la idea de que el significado de una palabra es afectado por las palabras en torno suyo, es decir el contexto de cada palabra está dado por todas las palabras de las que aparece rodeada.

Los dos métodos más usados para trabajar con palabras embebidas son:

- **Bolsa de palabas continua (*CBOW*[6])**. Se puede pensar como en *completar una oración a la que le falta una palabra*, en consecuencia toma como entrada un

[5] *NLP: Natural Language Processing*
[6] *CBOW: Continous Bag of Words*

contexto C de tamaño $|C|$ y trata de predecir la palabra que corresponde. Por ejemplo si se tiene la siguiente sentencia:

«La manzana es de color __»

El contexto estaría dado por la representación mediante bolsa de palabras para el modelo binaria de los tokens $\{la, manzana, es, de, color\}$ ($\{x_1, x_2, ...x_5\}$) y se medirá el error con respecto a la representación binaria de la salida para la palabra y_i = «rojo». Se utiliza una red neuronal con una capa oculta para ajustar los pesos correspondientes a cada palabra usando contextos de tamaño $|C|$. La Figura 7 presenta de manera ilustrativa la arquitectura de la red neuronal para **CBOW**, donde W es la matriz de pesos que mapea las entradas x_i a la capa oculta. Generalmente se toma el promedio de todas las palabras en el contexto C y $W^{'}$ es la matriz de pesos que mapea las salidas de capa oculta a la capa de salida.

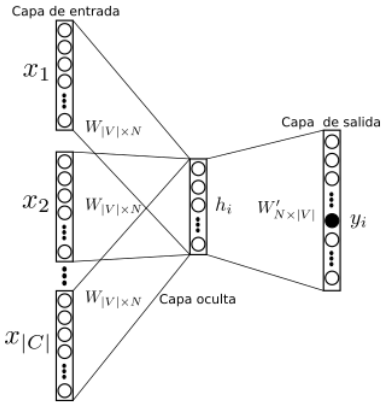


Figura 7: Arquitectura de la red neuronal para la representación **CBOW**.Elaboración propia basada en [Mikolov et al., 2013]

La palabra que se quiere aprender puede ser vista como una palabra «central» y las palabras en torno a ella, serían su contexto. Así podemos ver el contexto C en términos de un tamaño de ventana ws[7].

- **SKIP-GRAM**. Es el proceso inverso a CBOW, donde a partir de una palabra se predice su contexto. Es decir para cada posible posición de una palabra en un

[7] *ws: windows size*

contexto dado, se obtienen $|C|$ distribuciones de probabilidad con $|V|$ probabilidades cada una. La Figura 8 presenta de manera ilustrativa la arquitectura de la red neuronal para **SKIP-GRAM**.

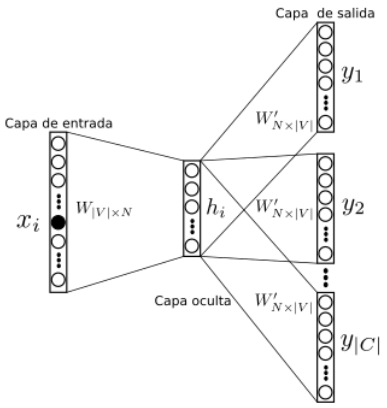


Figura 8: Arquitectura de la red neuronal para la representación **SKIP-GRAM**. Elaboración propia basada en [Mikolov et al., 2013]

De acuerdo con [Mikolov et al., 2013], algunas de las ventajas de un método con respecto de otro de acuerdo son: el modelo *SKIP-GRAM* es mejor cuando se tienen pocos datos y es bueno para representar palabras poco frecuentes, mientras que *CBOW* es más rápido y produce mejores representaciones de palabras frecuentes.

1.4. Detección de tópicos

En la recuperación de información es importante conocer el contenido de los documentos y si bien hasta el momento se han presentado estructuras que ayudan a realizar búsquedas dentro de los mismos, la necesidad de información del usuario puede ir dirigida a entender «¿de qué tratan los documentos?». Para esto, las estructuras de datos como los índices invertidos no son suficientes pues se requiere profundizar en la información contenida en todos los documentos. A este tipo de problemas se les conoce como *detección de tópicos*.

Para resolver el problema anterior es necesario encontrar las «aglomeraciones naturales dentro de los documentos», es decir, encontrar aquellos subgrupos donde los documentos pertenecientes a cada subgrupo comparten características que los ha-

cen «similares» entre ellos y diferentes entre los documentos que pertenecen a otros subgrupos.

Existen técnicas de aprendizaje no supervisado cuyo fin es realizar esta estratificación de los datos iniciales. A este conjunto de técnicas se les conoce como: algoritmos de agrupación o **clustering**.

En la recuperación de información, existe una hipótesis conocida como **hipótesis de clustering**, la cual dice (según [Manning et al., 2008]) que *«los documentos del mismo grupo se comportan de manera similar con respecto a la relevancia para las necesidades de información »*.

La hipótesis establece que si hay un documento de un grupo que es relevante para una solicitud de búsqueda, entonces es probable que otros documentos del mismo grupo también sean relevantes. Esto se debe a que los clústers agrupan documentos que comparten muchos términos.

1.4.1. Clustering

De acuerdo con [Manning et al., 2008], los algoritmos de clustering agrupan un conjunto de documentos en subconjuntos o agrupaciones. El objetivo de los algoritmos es crear clústeres que sean coherentes internamente, pero claramente diferentes entre sí. En otras palabras, los documentos dentro de un grupo deben ser lo más similares posible; y los documentos de un grupo deben ser lo más diferentes posible de los documentos de otros grupos. La agrupación es la forma más común de aprendizaje no supervisado. Sin supervisión significa que no hay ningún experto humano que haya asignado documentos a las clases. En la agrupación, es la distribución y composición de los datos lo que determinará la pertenencia al clúster.

Formalmente, el problema de clustering puede definirse cómo sigue: utilizando como entrada una colección de objetos X y un valor entero K y obtener un conjunto $G = \{G_1, G_2, ..., G_K\}$. Donde cada objeto $x_i \in X$ es un vector definido en $\mathbb{R}^N$, se cumple que $1 < K \leq |X|$. Al finalizar el proceso cada elemento x_i debe pertenecer a uno de los

grupos en G.

Es importante recalcar que:

- m_k es el número de elementos en el grupo G_k, es decir $m_k = |G_k|$
- c_k será utilizado para referirse al centroide de todos los elementos en G_k, el cuál puede ser:
  - La media geométrica:
    $$c_k = \frac{1}{m_k} \sum_{i=1}^{m_k} x_i$$
  - La mediana.
  - El *mediode*: uno de los elementos en el grupo (centro).
- Se utilizará M para hacer referencia al conjunto de centroides $\{c_1, c_2, ..., c_k\}$.
- En el problema de clustering cada objeto x_i solo pertenece a un grupo, por lo tanto para cualquier par de grupos G_A, G_B donde $A \neq B$, se cumple que:

$$G_A \cap G_B = \emptyset$$

.

Es posible resumir el proceso de clustering como sigue:

1. Definición del dominio, lo cual implica:
   - Identificar los objetos a analizar.
   - Determinar el propósito de generar grupos.
   - El conjunto de características que describen los objetos.
2. Definir una función de comparación, es decir cómo se determina que tan parecidos son los objetos
3. Definir cómo medir la calidad de los grupos encontrados.
4. Determinar el algoritmo de clustering a utilizar.

1.4.1.1. Medidas de similitud y distancias

Una vez definido el dominio del problema es necesario contar un mecanismo para comparar los objetos en la colección. Lo anterior con la finalidad de decir que tan similar o distante es un objeto x_i con respecto de otro x_j. Para esto se puede utilizar una función de distancia d o bien una función de similitud *sim*. Distancia y similitud están relacionados, desde que dos objetos que están a una distancia pequeña se puede decir que son similares. En la Tabla 1 se presentan las principales medidas de distancia y similitud presentadas en la literatura.

Nombre	**Función**
Minkowski	$L_q(x_i, x_j) = \sqrt[q]{\sum_{k=1}^{N}(x_{ik} - x_{jk})^q}$
Coseno (similitud)	$\cos(x_i, x_j) = \frac{\sum_{k=1}^{N} x_{ik} * x_{jk}}{\sqrt{\sum_{k=1}^{N} x_{ik}^2}\sqrt{\sum_{k=1}^{N} x_{jk}^2}}$
Coseno (distancia)	$dcos(x_i, x_j) = 1 - \cos(x_i, x_j)$
Divergencia de Kullback-Liebler	$kld(x_i, x_j) = \sum_{k=1}^{N} x_{ik} \times \log \frac{x_{ik}}{x_{jk}}$

Cuadro 1: Principales medidas de distancia y similitud presentadas en la literatura. Elaboración propia.

1.4.1.2. Medidas de calidad

Evaluar si un agrupamiento es **«bueno»** o no, no es una tarea simple, desde que no existe un criterio definitivo para determinarlo. Sin embargo se han propuesto muchos criterios de evaluación, los cuales podemos dividir en externos e internos. En este caso se presentarán las medidas de calidad *internas*. Este tipo de medidas evalúan que tan compactos son los clústers mediante el uso de una medida similitud o distancia. Esta clase de métricas generalmente intentan medir la cohesión en cada clúster (intra-clúster), la separación de los diferentes clústers (inter-clúster) o bien una combinación de ambas.

- **La suma de los errores al cuadrado** (SSE[8])**.** Consiste en calcular la suma de la distancia de todos los objetos con respecto de los centroides de sus respectivos

[8] *SSE: Sum of Squared Error*

grupos. Se calcula mediante la siguiente formula:

$$SSE = \sum_{k=1}^{K} \sum_{\forall x_i \in G_k} (x_i - c_k)^2$$

Donde G_k es el k-ésimo clúster y c_k es el centroide del grupo. SSE es una forma simple de medir que tan diferentes son los elementos en un clúster con respecto de su centro geométrico. Esta medida permite maximizar la similitud entre objetos en el mismo grupo mediante la minimización del SSE.

- **Silueta.** El coeficiente de silueta trata de minimizar la distancia intra-clúster mientras maximiza la distancia inter-clúster. El coeficiente $sil(G)$ para un conjunto de clústers G se determina como sigue:

  1. Para un elemento $x_i \in G_A$ su distancia promedio a_i con respecto de todo los elementos en su mismo clúster se calcula como sigue:

  $$a_i = \frac{1}{m_A - 1} \sum_{\forall x_j \in G_A; i \neq j} d(x_i, x_j)$$

  2. Se calcula su distancia promedio con mínima b_i con respecto a los elemento que no están en el mismo clúster al que pertenece x_i

  $$b_i = \min_{G_B \neq G_A} \frac{1}{m_B} \sum_{\forall x_j \in G_B} d(x_i, x_j)$$

  Con m_A y m_B definidos anteriormente.

  3. Con el fin de tratar de maximizar la separación entre x_i y los elementos que no están en su mismo grupo, mientras se minimiza la distancia promedio con los que se encuentran en su mismo grupo; se realiza el siguiente cociente:

  $$sil(x_i) = \frac{b_i - a_i}{\max(a_i, b_i)}$$

4. Se suma el coeficiente para todos los elementos en un mismo clústers

$$sil(G_k) = \frac{1}{m_k} \sum_{x_i} sil(x_i)$$

5. Finalmente se suman los coeficientes de cada grupo para obtener el coeficiente global

$$sil(G) = \frac{1}{K} \sum_{k=1}^{K} sil(G_k)$$

Está métrica tiene un costo mucho mayor que el SSE ya que en el peor caso se requiere conocer la distancia entre todos los elementos de la colección lo cual no es factible para colecciones de datos muy grandes.

1.4.1.3. Criterios para escoger K

Este problema ha sido objeto de muchos estudios, pero solo se mencionaran dos métodos populares propuestos en [Madhulatha, 2012] y [Aranganayagi and Thangavel, 2007] respectivamente.

- **Método del codo**. De acuerdo con, consiste en calcular el valor del SSE para valores continuos de K y elegir el valor en que se da el cambio máximo en el valor de SSE.
- **Método de la silueta**. Según con En este caso de calcula el valor del coeficiente de silueta y se elige el valor de K que presente el valor máximo del score.

Existen diferentes algoritmos de clustering, sin embargo en este proyecto solo se presentaran dos: **K-Means** y **Clustering Jerárquico**.

1.4.2. *K-Means*

Es un algoritmo basado en particionado. El algoritmo requiere cómo entrada la colección X y el número de clústers K. El proceso comienza seleccionando K elementos de

X, los cuales son utilizados como los centroides iniciales. Posteriormente cada punto es asignado a su centroide más cercano, todos los elementos asignados a un mismo centroide forman un clúster. Una vez asignados todos los elementos se calcula el centroide c_k para cada uno de los grupos. Se repiten el proceso de asignación y actualización de los centroides hasta que no hay cambios en los centroides.

Formalmente, el algoritmo *K-Means* es un algoritmo **iterativo** que dado un conjunto inicial de K centroides $c_1^{(1)}, c_2^{(1)}, ..., c_k^{(1)}$ se realizan los siguientes pasos:

- Asigna cada observación al grupo con la media más cercana

$$G_i^{(t)} = \{x_p : \|x_p - c_i^{(t)}\| \leq \|x_p - c_j^{(t)}\| \ \forall \ 1 \leq j \leq k\}$$

  Donde cada x_p va exactamente dentro de un $G_i^{(t)}$.

- Calcular los nuevos centroides como el centroide de las observaciones en el grupo.

$$\mathbf{c}_i^{(t+1)} = \frac{1}{|G_i^{(t)}|} \sum_{\mathbf{x}_j \in G_i^{(t)}} \mathbf{x}_j$$

  El algoritmo repetirá estos pasos hasta que $|c_i^{(t+1)} - c_i^{(t)}| < \epsilon$

1.4.3. Clustering Jerárquico

En este tipo de algoritmos, se asume que la medida de similitud se puede extender de tal manera que es posible medir la similitud entre dos subconjuntos. Es decir, si s es una medida de similitud y $A, B \subseteq \mathbf{X}$ siendo $\mathbf{X}$ el espacio a clusterizar, entonces se define S como:

$$S(A, B) = \sum_{(x,y) \in \mathbf{X} \times \mathbf{X} \ x \neq y} \frac{s(x, y)}{|A||B|}$$

Existen dos esquemas básicos de algoritmos jerárquicos:

- **Clustering Jerárquico Aglomerativo**. Este algoritmo trata a cada documento co-

mo un grupo único e individual desde el principio y luego combinan (o aglomeran) sucesivamente pares de grupos hasta que todos los grupos se fusionan en un solo grupo que contiene todos los documentos. Formalmente, el esquema aglomerativo sigue los siguientes pasos:

- Sea $G_0 = \{\{x_1\}, \{x_2\}, ..., \{x_m\}\}$
- Para todo $r \in \{2, ..., m\}$ se calcula:

$$G_r = (G_{r-1} \setminus \{A_r, B_r\}) \cup \{D_r\}$$

Dónde A_r, B_r y D_r son tales que:

$$S(A_r, B_r) = \text{máx}\{S(A_r, B_r) | (A_r, B_r) \in G_{r-1} \times G_{r-1}, A_r \neq B_r\}$$

$$D_r = A_r \cup B_r$$

- **Clustering Jerárquico Divisivo**. El esquema divisivo procede de manera opuesta al aglomerativo: en vez de unir los clústers que son más similares, separa los que menos se parecen. Formalmente, el esquema divisivo sigue los siguientes pasos:
  - Sea $G_m = \{\mathbf{X}\}$
  - Para todo $r \in \{m-1, ..., 1\}$ se calcula:

$$G_r = (G_{r-1} \setminus \{D_r\}) \cup \{A_r, B_r\}$$

Dónde A_r, B_r y D_r son tales que:

$$S(A_r, B_r) = \min_{G \in G_{r+1}} (\text{mín}\{S(A, B) | (A, B) \in G \times G, A \neq B, A \cup B = G\})$$

Es importante mencionar que en casos extremos, el algoritmo aglomerativo genera un único clúster en el cual están todos los elementos, mientras que el algoritmo divisivo genera m clústers únicos e individuales para cada elemento de **X**. A diferencia

del algoritmo de *K-Means*, el clustering jerárquico no requiere que pre-especifiquemos el número de grupos K.

1.5. Resumen

A lo largo de este capítulo se describieron las principales estructuras de datos para recuperar información de un conjunto de documentos no estructurados: los índices invertidos la cual se describe como una estructura de datos de índice que almacena un mapeo del contenido, como palabras o números, a sus ubicaciones en un documento o un conjunto de documentos. Adicionalmente, se presentaron índices invertidos binarios, que recuperan la información de existencia (1) o ausencia (0) de un término w_i dentro de un documento d_j. De igual forma, se presentó el modelo frecuencial, el cual incorpora la información de la *cantidad de veces* que aparece un término w_i en un documento d_j. Para lidiar con el problema de palabras «muy comunes», se introdujo el modelo *TF-IDF*, el cual realiza una ponderación del término w_i con base a su frecuencia en d_j relativo a su relevancia dentro de la colección de documentos D. Estas diferentes representaciones de índices invertidos se obtienen variando la definición de q, que representa la relevancia de la información contenida en los documentos. Por último, se mostró una representación alterna de las índices invertidos en términos de listas de posteo. Dicha representación es una forma eficiente de generar índices invertidos pues permite reducir su almacenamiento en memoria.

Adicionalmente se abordó la importancia de realizar una traducción del texto en símbolos con significado para el sistema informático. Así mismo, se presentaron las principales técnicas de preprocesamiento de texto que tienen como objetivo tanto disminuir el tamaño del vocabulario, como mejorar el proceso de recuperación de información a partir de fuentes de datos semi estructuradas. Una vez realizada la limpieza al texto, se mostraron diferentes dos formas de representar el texto en términos matriciales: bolsas de palabras y palabras embebidas. Para las bolsas de palabras se presentaron los modelos binarios, frecuencial y *TF-IDF* haciendo notar que existe una relación directa entre una representación por bolsa de palabras y un índice invertido en

su forma matricial. Para palabras embebidas se presentaron las dos técnicas más populares conocidas como *CBOW* y *SKIP-GRAM*. De acuerdo con [Mikolov et al., 2013], el modelo *SKIP-GRAM* es mejor cuando se tienen pocos datos y es bueno para representar palabras poco frecuentes, mientras que *CBOW* es más rápido y produce mejores representaciones de palabras frecuentes.

Después, se introdujo la necesidad de poder conocer el contenido de los documentos sin ayuda de un humano, es decir, es necesario agrupar los documentos de acuerdo a su contenido para saber «*¿de qué tratan?*». Dentro de recuperación de información, a este proceso se le conoce como *detección de tópicos* en datos semiestructurados. Para resolver este problema se presentaron algoritmos de *clustering* además se mostró la *hipótesis de clustering*. Se generaron las definiciones necesarias para poder describir diferentes algoritmos para generar clústers, así como las medidas de similitud/distancia más populares, criterios para evaluar los clústers obtenidos y criterios para escoger la cantidad de clústers. Por último se presentaron dos tipos de algoritmos: *K-Means* y *Clustering jerárquico*. El primero se caracteriza por ser un proceso iterativo donde a cada paso se recalculan los centroides de acuerdo a las observaciones más cercanas. Por otro lado, el *clustering jerárquico* busca unir (aglomerativo) o dividir (divisivo) de acuerdo a una extensión de la medida de similitud que permite evaluar la similitud no solo entre observaciones sino entre conjuntos de observaciones (subconjuntos de $\mathbf{X}$).

Capítulo 2

Trabajo Relacionado

Capítulo 2. Trabajo Relacionado

La analítica de texto es un tema que se ha popularizado en los últimos años, en particular los tópicos relacionados a análisis de sentimientos. Este último se aborda como un problema de aprendizaje supervisado, dónde a partir de una base de conocimiento, se desea construir un clasificardor de tal forma que aprenda y generalice los resultados. Por ejemplo: [Danisman and Alpkocak, 2008] utilizan como fuente de conocimiento el conjunto de datos `ISEAR` [1] sobre la cual, utilizando una representación vectorial *TF-IDF*, comparan el desempeño de dos clasificadores: *Naïve-Bayes* y *Máquinas de soporte vectorial*. El autor hace hincapié en el tipo de procesamiento y muestra que las técnicas de *stemming* y eliminación de *stop words* se deben manejar con precaución, pues se puede eliminar información relevante para el clasificador.

Por otro lado, [Dave et al., 2003] realizó un proyecto de clasificación sobre comentarios en internet. Usando las reseñas de productos de *Amazon* y *C|net* como fuente de datos, el autor ocupa técnicas de clasificación para poder diferenciar los comentarios positivos de los negativos. Es importante mencionar que a lo largo de este trabajo se muestra el impacto que tienen en la clasificación las diferentes técnicas de preprocesamiento de texto y de representación vectorial.

De igual forma, [Lula and Wójcik, 2011] muestra como una alternativa el uso del análisis de sentimiento y analítica de texto como una opción para analizar las opiniones de los celulares en polaco. Los autores desarrollaron un sistema que permite retroalimentar al consumidor sobre las principales características de los celulares.

De acuerdo con [Pak and Paroubek, 2010], los *microblogs* se han vuelto sumamente populares en los últimos años; los autores de dichos *microblogs* escriben en ellos sobre su vida, comparten opiniones sobre una gran variedad de temas y discuten sobre los problemas actuales. En este trabajo, los autores utilizan *Twitter* como fuente de conocimiento y muestra la importancia de la analítica de texto sobre esta plataforma. Es importante mencionar que el autor decidió utilizar *Twitter* principalmente por

[1] *ISEAR: International Survey on Emotion Antecedents and Reactions*

su tamaño y variedad en los usuarios de la plataforma. A lo largo de este trabajo, el autor muestra de forma detallada los pasos para extraer, procesar, analizar y clasificar los datos contenidos en *Twitter*.

De forma similar, [Thelwall et al., 2011] utilizó *Twitter* para analizar la opinión correspondiente a los eventos más significativos a lo largo de un mes, con lo cual se muestra una vez más la versatilidad de aplicaciones que tiene la información contenida en dicha plataforma.

Así como se puede utilizar analítica de texto y procesamiento de lenguaje natural para entender más sobre productos y servicios, esta herramienta en conjunto con la información presentada en *Twitter* se puede ocupar para realizar comparativos entre empresas, tal es el caso de [Vidya et al., 2015], dónde los autores utilizaron análisis de sentimiento para comparar la reputación de diferentes compañías de telefonía móvil al contrastar las opiniones de los usuarios respecto a los servicios y productos ofrecidos.

De forma similar, en [Guercini et al., 2014] los autores utilizaron *Twitter* para encontrar los aspectos más relevantes para los clientes de las aerolíneas. A través de la construcción de una base de conocimiento, preprocesamiento de texto y generación de clasificadores, el trabajo logró determinar lo «*bueno*» y lo «*malo*» de cuatro aerolíneas y así realizar propuestas para mejorar la experiencia del usuario. Es interesante observar que en este artículo, no se construyen clasificadores de aprendizaje computacional, sino a través de una clasificación de símbolos y palabras y un conteo de las mismas en cada *tweet* se determinó la polaridad del mismo.

Todos los ejemplos anteriores muestran aplicaciones de la analítica de texto en inglés, sin embargo es importante mencionar que el tema del análisis de sentimiento y extracción de información de textos también ha sido un tema relevante trabajado en idioma español: [Miranda and Guzmán, 2017] muestra el estado del arte así como una metodología básica para realizar análisis de sentimiento en español. Por su parte, [Costumero et al., 2014] muestra una aplicación de la analítica de texto y recuperación de la información enfocada a registros electrónicos de salud en España.

Además, de los trabajos relacionados a analítica de texto en *Twitter* en espa-

ñol, resalta [Tellez et al., 2017] dónde los autores muestran una metodología para realizar análisis de sentimiento en *Twitter* utilizando las bases de *SemEval*. Además, se presentan los impactos generados a los clasificadores al utilizar diferentes técnicas de *n-gramas* y *q-gramas*.

De acuerdo con [Culnan et al., 2010] el surgimiento de plataformas como *Twitter* o *Facebook* han generado comunidades *online* con intereses acerca de una empresa, marca o producto, dichas comunidades reciben el nombre de *ambientes virtuales de clientes*. Los *ambientes virtuales de clientes* impactan directamente a las marcas, las ventas, servicio al cliente, soporte y desarrollo de nuevos productos. A lo largo de su trabajo, [Culnan et al., 2010] muestra el valor que generan las redes sociales para las empresas

Por su parte, [Hussain and Prieto, 2016] afirma que una aplicación potencial de técnicas de analítica avanzada para la industria financiera es la evaluación de la exposición al riesgo reputacionales relacionado con los servicios ofrecidos por los bancos a sus clientes, ya que una percepción negativa puede disminuir la capacidad de un banco para mantener las relaciones comerciales existentes, establecer nuevas relaciones o el acceso continuo a fuentes de financiamiento.

Como se puede apreciar, gran parte de los trabajos asociados a analítica de texto abordan el problema como una aplicación de aprendizaje supervisado, la cual parte de tener una base de conocimiento a partir de la cual el modelo se va a entrenar y adquirir el conocimiento. En los ejemplos citados anteriormente, esta base de conocimiento contiene información previamente clasificada con el sentimiento o la polaridad del comentario. Por tanto, en caso de no existir dicha base, este acercamiento no se puede utilizar. En el caso del presente trabajo, el objetivo es determinar las principales quejas de servicios bancarios en México, de lo cual no existe una fuente de conocimiento a partir de la cual podamos realizar el aprendizaje. Es por ese motivo que este proyecto tiene un enfoque exploratorio y por tanto, de aprendizaje no supervisado. Además, a la fecha de realizar este manuscrito, no existen trabajos relacionados con analítica de texto dirigidos a bancos en México. Los resultados obtenidos en este trabajo pueden ayudar a mitigar el riesgo reputacionales que menciona [Hussain and Prieto, 2016].

2.1. Resumen

A lo largo de este capítulo se presentó una variedad de artículos dónde se puede apreciar la importancia que tiene la analítica de texto para abordar problemas de la vida cotidiana. Además se presenta *Twitter* como una fuente muy relevante para realizar minera de opiniones por su tamaño y versatilidad en sus usuarios. Se presentaron artículos que muestran el estado del arte del análisis de sentimientos en español así como un par de trabajos que muestran opciones hacia donde se pueden dirigir las empresas para explotar los datos en medios sociales. Por último se mostró y justificó la forma en cómo se abordó el problema relacionado a este trabajo.

Durante el siguiente capítulo se detallará la metodología con la cual se abordó el problema partiendo desde el contexto de los datos hasta la definición de las quejas.

Capítulo 3

Metodología

Capítulo 3. Metodología

Durante este capítulo se explicará la metodología utilizada para realizar el proyecto, incluyendo la extracción del conjunto de los datos, las trasformaciones, limpiezas y preprocesamiento de los mismos. Así como los análisis exploratorios, representación del texto y entrenamiento de los modelos para identificar las quejas en *Twitter*.

Para realizar el proyecto se utilizó la metodología *CRISP-DM (Cross Industry Standard Process for Data Mining)* la cual según [Brown, 2015] es la más usada para minería de datos y de acuerdo con [Shearer, 2000] consta de las siguientes fases:

- **Entendimiento de negocio**. El objetivo de esta fase es definir y entender el objetivo del negocio.
- **Entendimiento de los datos**. Durante esta fase se busca generar familiaridad con los datos a utilizar.
- **Preparación de los datos**. Tiene como meta realizar la preparación necesaria de los datos previo a generar un modelo analítico.
- **Modelado**. El objetivo es aplicar diferentes modelos analíticos para resolver la problemática inicial.
- **Evaluación**. En esta fase, se evalúa a mayor detalle el modelo y se da una explicación respecto al problema planteado.
- **Despliegue**. Se realiza el despliegue y seguimiento del modelo.

Para fines de este trabajo, se consideran únicamente las primeras cinco fases. La etapa de *evaluación* se presentará en el capítulo de **Resultados.**

En la Figura 9 se puede aprecias la relación entre las diferentes fases de la metodología. De acuerdo con [Shearer, 2000], las flechas indican las dependencias más importantes y frecuentes entre las fases, mientras que el círculo exterior simboliza la naturaleza cíclica de la minería de datos en sí e ilustra que las lecciones aprendidas durante el proceso de minería de datos y de la solución implementada puede desencadenar nuevas preguntas comerciales, a menudo más centradas.

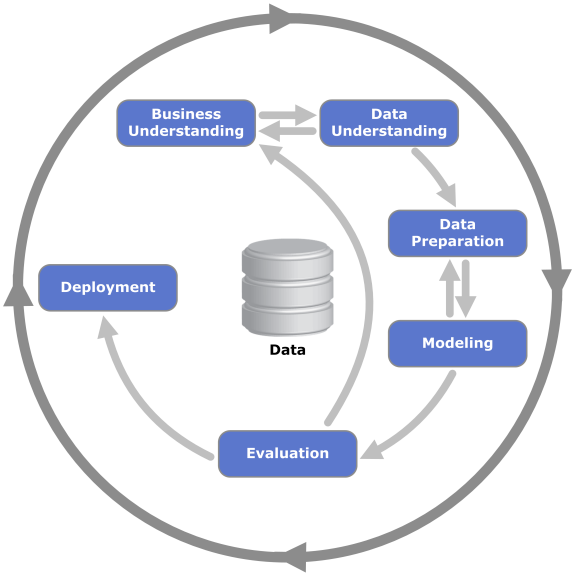


Figura 9: Diagrama de proceso que muestra la relación entre las diferentes fases de *CRISP-DM*. Tomado de [Jensen, 2012]

3.1. Entendimiento de negocio

Esta fase consiste en entender el objetivo del negocio (instituciones financieras) para poder plantear este último como un problema de minería de datos.

Desde el punto de vista de una institución financiera, es relevante poder conocer los puntos de dolor que tiene con el cliente, pues con esta información es posible tomar acciones, generar políticas o diseñar productos *ad-hoc* a las necesidades y demandas de los clientes con lo cual se puede aumentar la rentabilidad de la empresa.

Para poder lograr el objetivo anterior se requieren datos de cada institución, sin embargo estos datos pueden ser privados. En consecuencia, el análisis se enfoca en fuentes de datos públicas.

Las fuentes de datos públicos a utilizar son:

- Organismos públicos.
- Redes sociales.

3.1.1. Organismos públicos

La CONDUSEF busca proteger los intereses de los usuarios de servicios financieros mediante la supervisión y regulación a las instituciones financieras [CONDUSEF, 2018]. Como parte de sus labores, la Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros cada trimestre publica el BALANCE SOBRE LAS ACCIONES DE DEFENSA AL USUARIO, el cual tiene como objetivo informar a la población de las acciones que se ha tomado la CONDUSEF tomado en materia de controversias, dictámenes y defensa legal y gratuita. Para fines de este proyecto, se analizaron los resultados presentados en **materia de controversias**.

Con base en la información presentada en el BALANCE SOBRE LAS ACCIONES DE DEFENSA AL USUARIO correspondiente al primer trimestre del 2020 [CONDUSEF, 20Q1], se obtuvieron los siguientes cuadros:

Causas	2019	2020	%Var (2020 vs 2019)
Gestión de Cobranza	7,874	7,705	-2.1 %
Consumos no reconocidos	10,924	6,823	-37.5 %
Disposición de efectivo en cajero automático no reconocida	1,811	2,954	63.1 %
Negativa en el pago de la indemnización	4,036	2,893	-28.3 %
Transferencia electrónica no reconocida	1,547	2,844	83.8 %

Cuadro 2: Principales causas de controversias durante primer trimestre 2020. Elaboración propia.

Institución	2019	2020	%Var (2020 vs 2019)
Banorte	6,839	7,661	12.0 %
BBVA	8,476	6,976	-17.7 %
Banamex	7,352	6,278	-14.6 %
Santander	5,619	6,113	8.8 %
HSBC	3,380	2,839	-16.0 %

Cuadro 3: Número de controversias por institución financiera durante primer trimestre 2020. Elaboración propia.

Producto	2019	2020	%Var (2020 vs 2019)
Tarjeta de crédito	13,517	10,930	-19.1 %
Crédito personal	8,788	8,639	-1.7 %
Tarjeta de débito	8,896	8,554	-3.8 %
Reporte de crédito especial	8,987	7,903	-12.1 %
Daños - Automóviles	4,642	3,586	-22.7 %

Cuadro 4: Número de controversias por producto durante primer trimestre 2020. Elaboración propia.

De la Tabla 2 se observa un incremento considerable en las causas *«Disposición de efectivo en*

cajero automático no reconocida» y *«Transferencia electrónica no reconocida»* (63.1 % y 83.8 % respectivamente) lo cual podría sugerir **problemas relacionadas a canales digitales**.

Por otro lado, en la Tabla 3, Banorte resalta como el banco con mayor número de quejas y un incremento del 12 % respecto al año anterior. Así mismo es importante notar que durante 2019 BBVA fue el banco con mayores controversias, con un total de 8,476 durante el primer trimestre del 2019. Además, dicho valor es el mayor de toda la tabla.

En la Tabla 4, se observa una disminución constante en todos los productos ofrecidos, no obstante es relevante mencionar que las controversias están relacionadas con «*dinero propio*» (tarjeta de débito) y «*dinero prestado*» (tarjetas de crédito o créditos personales).

Es importante mencionar que según [CONDUSEF, 20Q1] los cajeros automáticos aparecen en la posición 8 del reporte, sin embargo ha tenido un incremento en quejas de 61.9 % del primer trimestre 2019 al primer trimestre 2020.

Esta información es relevante, sin embargo es superficial pues no otorga información que involucre el cruce de las tres dimensiones presentadas en el reporte (institución, producto y causa), además que estos datos solo se publican de manera trimestral. Una forma de resolver este problema es complementando esta información con datos contenidos en **redes sociales**.

3.1.2. Redes sociales

Según [Ayala, 2014] la aparición de las redes sociales y comunidades virtuales modificaron profundamente los hábitos comunicativos de los usuarios de la Red. *Facebook* y *Twitter*, usados por millones de individuos, han permitido que grupos de personas se sientan permanentemente comunicados.

De acuerdo a una encuesta nacional en Estados Unidos analizada por [Jones and Fox, 2009], determinó que el 87 % de los individuos entre 18 y 32 años estaban regularmente en línea y que el 60 % de las personas en este grupo de edad habían creado un perfil personal en una red social. Este punto nos habla de la penetración que tienes las redes sociales en el día a día de la sociedad.

Por su parte [Zapatero et al., 2013], indica que las redes sociales han creado un canal de comunicación que permite la posibilidad de interactuar y relacionarse con otros usuarios, con los que se comparte alguna inquietud, motivación o afición. Particularmente, este tipo de relación se puede dar entre las marcas y los consumidores finales.

Para fines de este trabajo se eligió a *Twitter* como fuente de datos complementaria, pues mediante su servicio de consumo de datos en *streaming*[1] es posible extraer los datos directamente de la plataforma en formato JSON mediante el consumo de su API pública. Sin embargo, la complejidad de utilizar *Twitter* como una fuente de datos complementaria radica en la habilidad de poder extraer la información a partir de datos *semi-estructurados*.

3.2. Entendimiento de los datos

El objetivo de esta etapa de la metodología es aumentar la familiaridad con los datos, identificar problemas de calidad con los mismos y obtener hallazgos iniciales [Shearer, 2000].

Para poder lograr la familiaridad deseada, es necesario pasar por las siguientes fases:

- *Extracción de datos*. Describe la forma de obtener los datos.
- *Descripción de datos*. Se describen las variables obtenidas.
- *Exploración de datos*. Se realiza una exploración inicial de los datos en búsqueda de hallazgos relevantes.
- *Calidad de datos*. Mide la completitud de los datos obtenidos.

[1] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data

3.2.1. Extracción de datos

Para hacer uso del `API stream` de *Twitter* es necesario delimitar la búsqueda al menos en términos de idioma y temporalidad. Dado que el análisis va dirigido a los principales bancos en México, se consideró únicamente **idioma español**. Además, con el fin de buscar tendencias y detectar comportamientos estacionales, se decidió tener **dos años completos de historia**, mismos que comprenden un umbral de tiempo de **01 de enero 2018 a 31 de diciembre 2019**.

Se utilizó la base de datos del laboratorio de BigData de Infotec tomada del *stream* público de *Twitter* que corresponde a lo más al 1% de los *tweets* disponibles. Se tienen cerca de 1,576,800,000 *tweets* en esos años).

Para asegurar que los *tweets* fueran dirigidos a las instituciones financieras de interés, se utilizó la siguiente expresión regular en lenguaje de programación `Julia`:

```
context = r"@BBVA.Mex|@BBVABancomer|@Citibanamex|@SantanderMx|
@HSBC.MX|@ScotiabankMX|@BancoInbursa|@Banorte.mx|@BanjercitoSNC|
@BanBajioMX|@BancoAzteca\b|@BancoAfirme|@BanCoppel|@CondusefMX"imx
```

Es importante mencionar que en `Julia` se puede modificar el comportamiento de las expresiones regulares mediante alguna combinación de los indicadores `i, m, s` y `x` después de la comilla doble de cierre. Estas banderas tienen el siguiente significado:

- `i`. Hace una coincidencia de patrones que no distingue entre mayúsculas y minúsculas.
- `m`. Trata la cadena como varias líneas.
- `s`. Trata la cuerda como una sola línea.
- `x`. Le dice al analizador de expresiones regulares que ignore la mayoría de los espacios en blanco que no está ni con barra invertida ni dentro de una clase de personaje.

Por tanto, con los comandos `imx`, la expresión regular no distingirá entre mayúsculas y minúsculas, dividirá la cadena como varias líneas y eliminará los espacios

en blanco.

Con lo anterior se obtuvo un **universo inicial de 235,360 *tweets***.

Por otro lado, la información presentada en el Tabla 2 sugiere que el uso de expresiones como «cargo no reconocido», «transferencia(s)», «no funciona», etc., pueden ser usadas para detectar la casusa de las quejas. Así mismo, el Tabla 4 sugiere el uso de palabras clave como «crédito(s)», «préstamo(s)», «dinero», entre otras para identificar quejas relacionadas a los productos. Además, usando adjetivos como «pésimo», «asco», «peor», etc., se puede construir un conjunto de expresiones que denoten quejas. En consecuencia, se aplicó un segundo filtro para obtener los *tweets* correspondientes a las quejas, que consiste en la siguiente expresión regular:

```
pat = r"\#codi|\bcodi\bcr.ditos|soluci.n|profeco|transferen|interbanc
|pago|tarjetas|dep.sit|pr.stamos|fraude|no reconocido|tasas de
inter.s| cobros|cargo|contratar producto|llevo esperando|muy molesto|
p.simo servicio| malas pr.cticas|peor|tardan|ratero|ladron|no
funciona|maldito| adeudo| hasta la madr| asco|denunc|confiabl|tiempos|
aclaraci|robo| lo bueno| saldo|en l.nea|denunci|condusef|fall"imx
```

Con las expresiones anteriores se extrajeron los *tweets* que «etiquetaran» a algún banco de interés y además contengan las frases escogidas. Este último archivo **contiene 75,883 *tweets*** en formato `.json` donde cada línea es un JSON válido.

3.2.2. Descripción de datos

Mediante el uso de la librería `pandas` se generó una tabla a partir del archivo `.json`. En la Tabla 5 se puede apreciar la columna y la descripción[2] de las variables contenidas en la tabla resultante:

Columna	Descripción
`created_at`	Fecha de creación del *tweet*
`id`	Identificador único del usuario de *Twitter*
`id_str`	Identificador único del *tweet*
`text`	Texto del `tweet`
`truncated`	Indicador de *tweet* truncado
`extended_tweet`	Diccionario con información adicional del *tweet* completo
`entities`	Diccionario con información adicional del *tweet*
`metadata`	Diccionario con información de metadatos del *tweet*
`source`	Fuente de donde fue publicado el *tweet*
`in_reply_to_status_id`	Si el *tweet* es una respuesta, este campo contendrá el ID del *tweet* original en formato `int`
`in_reply_to_status_id_str`	Si el *tweet* es una respuesta, este campo contendrá el ID del *tweet* original en formato `str`
`in_reply_to_user_id`	Si el *tweet* es una respuesta, este campo contendrá el ID de usuario del *tweet* original en formato `int`
`in_reply_to_user_id_str`	Si el *tweet* es una respuesta, este campo contendrá el ID de usuario del *tweet* original en formato `str`
`in_reply_to_screen_name`	Si el *tweet* es una respuesta, este campo contendrá el nombre de usuario del *tweet* original
`user`	Diccionario con información adicional del usuario que publicó el *tweet*
`coordinates`	Diccionario con información de las coordenadas asociadas al *tweet*
`place`	Diccionario con información del lugar donde se publicó el *tweet*
`quoted_status_id`	Si el *tweet* es citado, este campo contendrá el ID del *tweet* original en formato `int`
`quoted_status_id_str`	Si el *tweet* es citado, este campo contendrá el ID del *tweet* original en formato `str`
`is_quote_status`	Indica si el *tweet* es citado
`quoted_status`	Diccionario con *tweet* citado
`retweeted_status`	Diccionario con *tweet retweeteado*
`quote_count`	Número de veces que el *tweet* ha sido citado
`reply_count`	Número de veces que el *tweet* ha sido reenviado
`retweet_count`	Número de veces que el *tweet* ha sido *retweeteado*
`favorite_count`	Número de veces que el *tweet* ha sido marcado como favorito
`favorited`	Indica si el *tweet* ha sido marcado como favorito
`retweeted`	Indica si el *tweet* ha sido *retweeteado*
`possibly_sensitive`	Si el *tweet* contiene un *link*, indica si dicho *link* pudiera contener información sensible
`filter_level`	Indica el máximo nivel de filtro que se pueda usar y seguir visualizando el *tweet*
`lang`	Idioma asociado al *tweet*
`matching_rules`	Proporciona el *id* y *tag* asociado a la regla que encontró el *tweet*

Cuadro 5: Variables obtenidas de la extracción de datos usando la `API` de *Twitter*.

[2] https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/tweet-object

Es importante resaltar lo siguiente:

- La variable `created_at` contiene la información de cuando fue creado el *tweet*.
- Las variables `extended_tweet, place` y `user` son **diccionarios** en sí mismos.
- El texto del *tweet* puede venir en la variable `text` o en la variable `full_text` del diccionario `extended_tweet` dependiendo del valor de la variable `truncated`.
- El diccionario `user` contiene datos del usuario, dentro de los cuales resaltan:
  - `id`: identificador único de *Twitter*.
  - `screen_name`: nombre del usuario de *Twitter*.
- El diccionario `place` contiene datos del lugar geográfico de dónde se originó el *tweet*, dentro de los cuales resaltan:
  - `full_name`: ciudad y estado de dónde se originó el *tweet*.
  - `screen_name`: país de donde se originó el *tweet*.

3.2.3. Exploración de datos

Con el objetivo de empezar a tener hallazgos de los datos obtenidos, se decidió explorar el campo `text`.

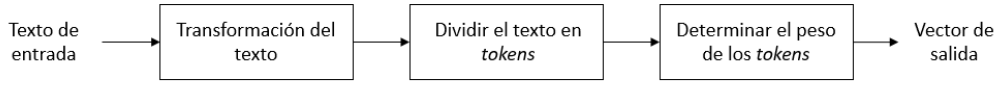


Figura 10: Tratamiento genérico del texto para obtener una representación vectorial. Elaboración propia basada en [Tellez et al., 2017]

En la Figura 10 se muestra el proceso de representación vectorial del texto. Utilizando como base a [Tellez et al., 2017] la trasformación del texto que consta de los siguientes pasos:

- Cambiar el texto a minúsculas.
- Eliminar acentos, caracteres especiales y *URLs*.

- Eliminar las *stopwords*.

Texto original	**Texto transformado**
Y de nuevo no funciona mi tarjeta!!! Eres un asco @Citibanamex!!!	nuevo funciona tarjeta asco citibanamex

Cuadro 6: Ejemplo de preprocesamiento aplicado a un *tweet*. Elaboración propia.

Dentro de la Tabla 6 se puede apreciar el cambio que existe entre el texto original y el texto preprocesado.

Luego, al dividir los *tokens*, se generó el vocabulario inducido por el *corpus de tweets*, el cual consta de **30,328 palabras únicas**.

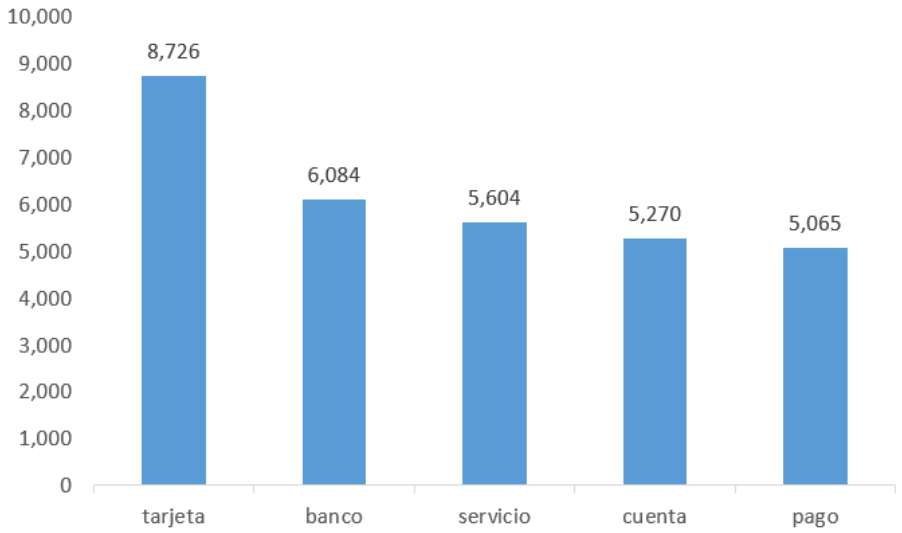


Figura 11: Las cinco palabras más frecuentes dentro del *corpus de tweets*. Elaboración propia.

En la Figura 11 se pueden observar las palabras más frecuentes. Es importante observar que la palabra más repetida es *«tarjeta»*, lo cual es consistente con la información presentada en el Tabla 4.

Por último, para determinar el peso de los *tokens* se hizo uso de la ponderación *TF-IDF*[3]. Con esto se logró tener una representación vectorial de los *tweets*.

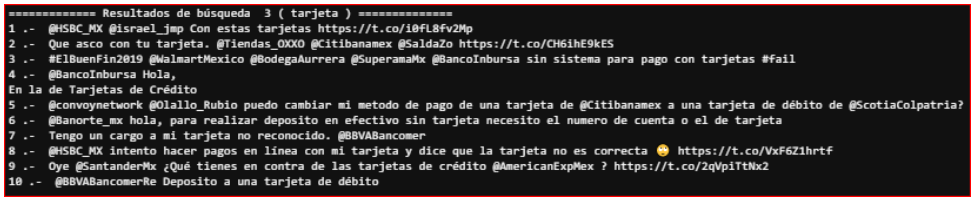
Como ejemplo **se construyó un índice invertido sobre listas de posteo** utilizando un pesado mediante *TF-IDF* como factor de relevancia. Se realizó la búsqueda de las

[3] *TF-IDF: Term Frequency - Inverse Document Frequency*

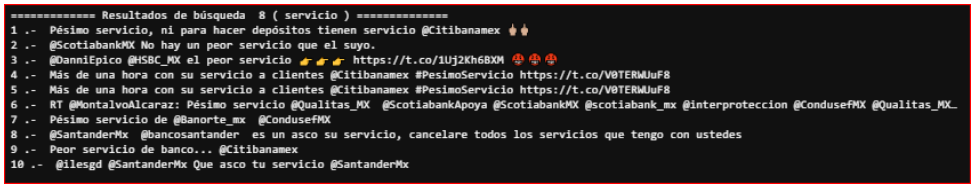
siguientes palabras: `pago`, `pagar`, `tarjeta`, `crédito`, `préstamo`, `llamada`, `llamar`, `servicio`, `atención`, `dinero`, `robo`, `fraude`, `cliente`, `usuario`, `cobro`, `tiempo`.

Palabra	Resultados obtenidos
pago	3,085
pagar	865
tarjeta	5,473
crédito	1,888
préstamo	125
llamada	1,569
llamar	421
servicio	3,563
atención	1,234
dinero	2,420
robo	650
fraude	1,028
cliente	1,550
usuario	432
cobro	440
tiempo	1,085

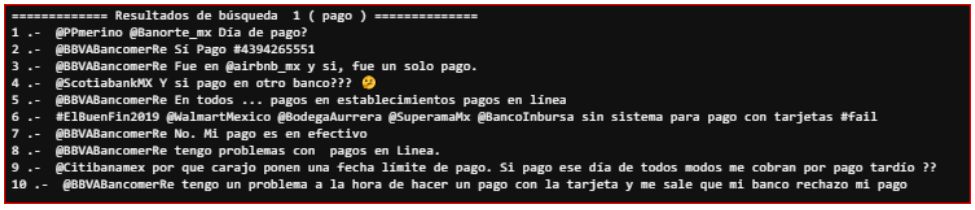
Cuadro 7: Resultados para la lista de palabras usada como filtros de búsqueda. Elaboración propia.



(a) Primeros 10 resultados de la palabra *tarjeta*



(b) Primeros 10 resultados de la palabra *servicio*



(c) Primeros 10 resultados de la palabra *pago*

Figura 12: Palabras con la mayor cantidad de resultados al realizar la búsqueda sobre el índice invertido. Elaboración propia.

Al observar la Figura 12 es notorio que los resultados de la *«tarjeta»* hacen referencia a **problemas con el uso o pago de la tarjeta de crédito** mientras que los resultados de *«servicio»* se relacionan con quejas sobre el **tiempo de espera y servicio al cliente**. Por último, los resultados de la palabra *«pago»* hacen referencia a problemas en **los sistemas al realizar pagos**.

3.2.4. Calidad de datos

En términos de completitud de datos, solo 16,670 *tweets* contienen datos de geo localización (de un total de 75,883), lo cual será una limitante para el análisis por zona geográfica. Para las demás variables de interés (`user.id`, `created_at`, `place.country`, `place.full_name`, `truncated`, `extended_tweet.full_text`, `text`) se cuenta con datos completos.

Por otro lado, tal como se vió en el **Capítulo 2** y **Capítulo 3**, al trabajar con datos de *Twitter* se debe considerar que cada uno de los *tweets* es **texto en formato libre**,

lo cual implica un **alto uso de gentilicios, jerga popular y faltas ortográficas**. Con el fin de generar vectores densos que mantengan la relación semántica de los textos, se utilizó un modelo de palabras embebidas (*word embeddings*) propuesto por [Mikolov et al., 2013] para controlar el tamaño de la matriz resultante y generar una mejor representación vectorial de los *tweets*.

3.3. Preparación de datos

Siguiendo la metodología propuesta por [Shearer, 2000], la fase de *preparación de datos* consiste en todas aquellas actividades destinadas en construir el conjunto de datos que alimentará el modelo. Los pasos que sugiere el autor son las siguientes:

- Selección de datos. Acotar los datos con los cuales se van a trabajar.
- Limpieza de datos. Se genera la limpieza necesaria en los datos.
- Construcción de datos. Se generan variables adicionales en caso de ser necesarias.
- Integración de datos. Se integran todos los datos en una tabla final que servirá para el modelado.

3.3.1. Selección de datos

Tal como se mencionó en la Sección 3.2, la limitante que existe al utilizar los datos de la CONDUSEF van dirigidos a la granularidad y frecuencia de los mismos, por tal motivo se eligió una red social y en específico *Twitter* como fuente de datos complementaria dada la factibilidad de acceso a los datos históricos. Los datos que van a permitir realizar un análisis más detallado de las quejas, su frecuencia, temporalidad y ubicación son las siguientes (presentadas en la Sección 3.2):

- `created_at` y `place.full_name`. Servirán para generar análisis temporales de las quejas.

- `place.country` y `place.full_name`. Servirán para generar análisis espaciales de las quejas.
- `text` y `extended_tweet.full_text`. Servirán para detectar las quejas.
- `truncated`. Servirá para saber qué campo utilizar entre `text` y `extended_tweet.full_text`.

```
Data columns (total 7 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   user.id                 75883 non-null  int64
 1   created_at              75883 non-null  object
 2   timestamp_ms            75883 non-null  object
 3   place.country           16670 non-null  object
 4   place.full_name         16670 non-null  object
 5   extended_tweet.full_text  33872 non-null  object
 6   text                    75883 non-null  object
dtypes: int64(1), object(6)
memory usage: 4.1+ MB
```

Figura 13: Campos seleccionados de *Twitter* para realizar el proyecto. Elaboración propia.

La tabla con los campos seleccionados se puede apreciar en Figura 13.

3.3.2. Limpieza de datos

En la Sección 3.2 se mostró que menos del 22 % de los datos cuenta con información que permita ubicarlos de forma geoespacial. Desafortunadamente no hay forma de deducir dichos valores, por lo cual esta será una limitante para el análisis geoespacial de las quejas.

Por otro lado, los campos `place.country` y `place.full_name` son texto que no está homologado, es decir, para un mismo nombre (por ejemplo *Veracruz*) aparecen distintos resultados (*Veracruz, Veracrus, Ver., Veracruz de Ignacio de la Llave*). Por tanto, se decidió aplicar las siguientes reglas de normalización al texto con el fin de limpiar estos campos:

- Transformar el texto a minúsculas.
- *Parsear* el texto y mantener el nombre popular.
- Eliminar signos de puntuación.

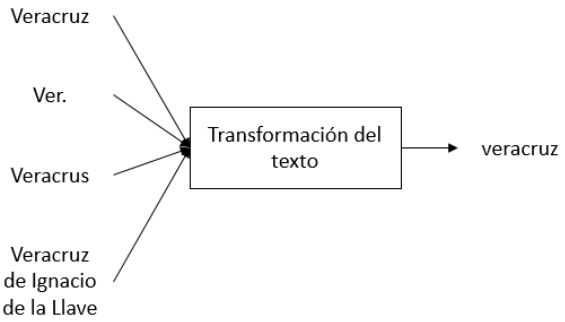


Figura 14: Ejemplo de transformación de texto aplicado a las variables *place.country* y *place.full_name*. Elaboración propia.

La figura Figura 14 muestra un ejemplo de las transformaciones realizadas. Dicha limpieza se aplicó para los estados pertenecientes a México.

Ningún otro tipo de limpieza se aplicó a los otros campos.

3.3.3. Construcción de información tabular auxiliar para el análisis

A partir de las variables iniciales, se construyeron las siguientes columnas:

- `texto`: usando `text` y `extended_tweet.full_text` dependiendo del valor de `truncated`.
- `pais`: a partir de la variable `place.country`.
- `estado` y `municipio`: utilizando la variable `place.full_name`.
- `id_user`: es un cambio de nombre de la variable `user.id`.
- `fecha2`: por la diferencia de huso horario, se restaron 5 horas a la variable `created_at` para que coincida con el horario de la Ciudad de México (*GMT-5*).

Por otro lado, de acuerdo con el objetivo del proyecto (mencionado en la Sección 3.1), para poder enfocar el análisis en los bancos del G7 es necesario identificar la institución bancaria a la cual está dirigida el *tweet*, sin embargo esto no es posible realizarlo directamente de los datos obtenidos. Entonces, mediante el uso de expresiones regulares y reglas de asociación se logró identificar a qué banco hace referencia cada *tweet* con lo cual **se construyó una variable indicadora para cada banco** que muestra

si el *tweet* mencionó a la institución financiera. Es decir, se construyeron las siguientes variables:

- `BBVA`. Toma el valor de 1 si el *tweet* menciona al banco BBVA.
- `Citibanamex`. Toma el valor de 1 si el *tweet* menciona al banco Citibanamex.
- `Santander`. Toma el valor de 1 si el *tweet* menciona al banco Santander.
- `HSBC`. Toma el valor de 1 si el *tweet* menciona al banco HSBC.
- `Scotiabank`. Toma el valor de 1 si el *tweet* menciona al banco Scotiabank.
- `Inbursa`. Toma el valor de 1 si el *tweet* menciona al banco Inbursa.
- `Banorte`. Toma el valor de 1 si el *tweet* menciona al banco Banorte.
- `Banjercito`. Toma el valor de 1 si el *tweet* menciona al banco Banjercito.
- `Banbajio`. Toma el valor de 1 si el *tweet* menciona al banco Banbajio.
- `BancoAzteca`. Toma el valor de 1 si el *tweet* menciona a Banco Azteca.
- `Afirme`. Toma el valor de 1 si el *tweet* menciona al banco Afirme.
- `Bancoppel`. Toma el valor de 1 si el *tweet* menciona al banco Bancoppel.
- `Condusef`. Toma el valor de 1 si el *tweet* menciona al banco CONDUSEF.

Con base en lo anterior, se generó la variable $g7_i$ definida como **el número de bancos del G7 mencionados en el i-ésimo *tweet***.

Dada la definición anterior, es importante resaltar lo siguiente:

- Si $g7_i = 0$ implica que en el *i-ésimo tweet* no hubo mención de ningún banco del G7.
- Si $g7_i > 0$ implica que en el *i-ésimo tweet* se mencionó a más de un banco del G7.

Con el fin de enfocar los esfuerzos del análisis a los bancos más importantes de México y **tener *tweets* ajenos entre instituciones bancarias**, se filtró considerando `g7==1` obteniendo así un **universo final de 45,324 *tweets***.

Por último, se generó la variable `banco` la cual contiene el nombre (etiqueta) de la institución bancaria a la cual va dirigida el *tweet*.

Con lo realizado hasta ahora es posible generar diferentes tipos de análisis específicos por institución, tiempo y geoespaciales, lo cual ya genera mayor granularidad que los datos presentados por la CONDUSEF. Sin embargo falta poder identificar las quejas de los usuarios de los servicios bancarios, para esto es necesario realizar análisis del texto de cada *tweet*. Adicionalmente, no todo el texto es relevante, por lo tanto se realizó el siguiente preprocesamiento a la variable `texto`:

1. Cambiar el texto a letras minúsculas.
2. Cambiar la *«ñ»* por *«ni»*. De esta forma, palabras como *«daño»* o *«años»* fueron transformadas a *«danios»* y *«anios»* respectivamente.
3. Eliminar la última *«s»* de las palabras que terminan en *«s»*. Con lo anterior, palabras como *«pagos»* y *«tiempos»* fueron cambiadas a *«pago»* y *«tiempo»*.
4. Eliminar los acentos.
5. Eliminar los caracteres especiales.
6. Eliminar *URLs*.
7. Eliminar *hashtags* y etiquetas de usuario, pues ya existe forma de identificar a qué institución financiera va dirigido el *tweet* (usando la variable `banco`). Además, el análisis de usuarios está fuera del alcance de este proyecto.
8. Eliminar *stopwords*. La lista de *stopwords* para idioma español contenida en la librería `nltk.corpus` se complementó con las palabras: *w, si, mas, rt, dia, solo, hora, puede, hoy, hace, estan, q, asi, vez* y *dicen*, ya que estas palabras tampoco muestran un valor en sí mismas.

Se decidió mantener los *emojis* dentro del texto, pues de acuerdo con [Shiha and Ayvaz, 2017], los *emojis* tienen un impacto fuerte al realizar análisis del texto.

La Figura 15 muestra un ejemplo de las trasformaciones realizadas sobre el texto orinal. El texto preprocesado se almacenó en una nueva variable llamada `texto_limpio`.

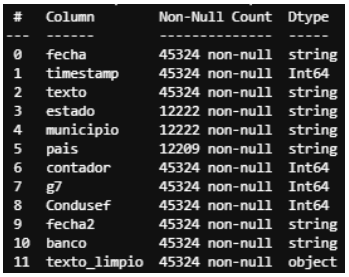
(a) Texto original

(b) Texto preprocesado

Figura 15: Ejemplo del preprocesamiento realizado sobre la variable `texto`. Elaboración propia.

3.3.4. Integración de datos

Por último, se integraron las variables: `fecha2`, `timestamp`, `texto`, `estado`, `municipio`, `pais`, `contador`, `g7`, `Condusef`, `banco` y `texto_limpio` en una solo conjunto de datos. La Figura 16 muestra la estructura de la tabla final.



#	Column	Non-Null Count	Dtype
0	fecha	45324 non-null	string
1	timestamp	45324 non-null	Int64
2	texto	45324 non-null	string
3	estado	12222 non-null	string
4	municipio	12222 non-null	string
5	pais	12209 non-null	string
6	contador	45324 non-null	Int64
7	g7	45324 non-null	Int64
8	Condusef	45324 non-null	Int64
9	fecha2	45324 non-null	string
10	banco	45324 non-null	string
11	texto_limpio	45324 non-null	object

Figura 16: Estructura de la tabla final sobre la cual se realizó el proyecto. Elaboración propia.

3.4. Modelado

En esta fase se seleccionan y aplican diferentes técnicas de modelado con el objetivo de resolver el problema inicial. Para poder recuperar las quejas a partir del texto, primero es necesario realizar una representación vectorial del mismo. Después, es necesario aplicar técnicas de aprendizaje no supervisado para identificar las quejas, pues el objetivo es detectar aquellas «aglomeraciones naturales» en los datos.

Según [Shearer, 2000], esta fase comprende los siguientes pasos:

1. Selección y ejecución de la técnica de modelado. Aquí se define el modelo analítico a utilizar.

2. Prueba del modelo. En este punto se evalúa el modelo para saber qué tan bueno es para realizar las predicciones.

3. Construcción del modelo. Se construye el modelo final.

En este caso, al ser una aplicación de técnicas de aprendizaje no supervisado, no es posible realizar una evaluación del modelo en términos de la predicción, sin embargo se puede evaluar la interpretabilidad de los resultados para validar que éstos tengan sentido de acuerdo al Entendimiento del negocio (Sección 3.1).

Para realizar la detección de quejas dentro de los *tweets*, el modelado de los datos se dividirá en tres partes:

1. Representación vectorial del texto.

2. Identificación de tópicos.

3. Identificación de subtópicos.

En cada sección se aplicarán los pasos correspondientes a la fase de *Modelado* propuesta por [Shearer, 2000].

3.4.1. Representación vectorial del texto

3.4.1.1. Selección y ejecución de la técnica de modelado para la representación vectorial de textos

Al ser texto en formato libre, una representación clásica mediante el uso de bolsa de palabras puede ser ineficiente ya que la matriz resultante será demasiado grande y dispersa. En consecuencia, se decidió utilizar un modelo de palabras embebidas (*word embeddings*) para controlar el tamaño de la matriz resultante y generar una mejor representación vectorial de los *tweets*.

Para llevar a cabo dicha tarea, se hizo uso de `fastText`, la cual es una librería abierta desarrollada por *Facebook* para clasificación y representación de texto [Joulin et al., 2016].

Los hiperparámetros importantes que requiere el modelo son los siguientes:

- `size` (dimensión del vector de salida). De acuerdo con [Facebook, 2020] los valores más ocupados rondan entre 100 y 300. Dado que el modelo pre-entrenado para idioma español es de dimensión 300, se decidió utilizar `size=300`.
- `window` (ventana de contexto). En su trabajo, [Church and Hanks, 1990] sugieren una ventana `w=5`, sin embargo al ser los `tweets` textos cortos, se requiere una ventana pequeña. En este caso se utilizó `window=2`.
- `min_n` y `max_n` (tamaño de *q-gramas*). En [Tellez et al., 2017] podemos encontrar los experimentos para diferentes valores de *q*. Dado que la ventana a utilizar es pequeña y el texto está en formato libre, se decidió utilizar `min_n=2` y `max_n=5`.
- `epochs` (número de épocas que el modelo se ajustará a los datos). Dado que el *corpus* es pequeño (45,324 documentos), se decidió `epochs=15`.

Con lo anterior se obtuvo un modelo para palabras embebidas, con el cual se tiene la representación de cada palabra en un espacio vectorial de dimensión trescientos. Sin embargo es necesario obtener no solo la representación de cada palabra sino de cada *tweet*. Por tanto se decidió utilizar la representación básica presentada por [Kenter et al., 2016]. En consecuencia, se definió el TE_i (*tweet embedding* para el documento i) como:

$$TE_i = \frac{1}{n_i} \sum_{j=1}^{n_i} WE_{ij}$$

Donde:

- n_i es la cantidad de palabras en el *i-ésimo tweet*.
- WE_{ij} es el *word embedding* correspondiente a la *j-ésima* palabra del *i-ésimo tweet*

Por último, se calcularon los ***tweets embeddings* normalizados** (TEN_i) de la siguiente forma:

$$TEN_i = \frac{TE_i}{||TE_i||}$$

Donde:

- $||TE_i||$ es la norma $p = 2$ del *i-ésimo tweet embedding*. Lo anterior es necesario pues se utilizará la *distancia coseno* ya que como se mostró en el **Capítulo 1**, se ha demostrado que es una gran herramienta para problemas de analítica de texto.

Los ***tweets embeddings* normalizados** se almacenaron en la variable `embeddings`. Sobre estos valores se aplicarán técnicas de clusterización para la detección de tópicos.

En la representación vectorial de textos no es posible generar una evaluación contextual de los resultados, esto se realizará al identificar los tópicos dentro de dicha representación.

3.4.2. Identificación de tópicos

3.4.2.1. Selección y ejecución de la técnica de modelado para la detección de tópicos

Una vez generada la representación vectorial de cada *tweet*, el siguiente paso es encontrar las «aglomeraciones naturales» dentro de los datos. Las técnicas de *Clustering* se encargan de dividir los datos en subconjuntos (G_i) con la característica que los datos que pertenecen a cada clústers son lo **similares entre ellos** ($d(x_o, x_p) < \varepsilon$ si $x_o, x_p \in G_i$) mientras que si las observaciones pertenecen a diferentes clústers, las elementos son **diferentes entre ellos** ($d(x_o, x_p) >> \varepsilon$ si $x_o \in G_i$ y $x_p \in G_j$).

En este caso se hizo uso de algoritmo de clusterización *K-Means*, pues es un algoritmo de complejidad computacional lineal ([Madhulatha, 2012]) dada por $\mathbf{O}(nkl)$, dónde n es la cantidad de observaciones, k es el número de clústers y l es el número de iteraciones que requiere el algoritmo para converger.

La métrica utilizada para este proyecto fue la **métrica coseno** ya que es la más popular para este tipo de problemas ([Salton, 1989], [Larsen and Aone, 1999] , [Baeza-Yates et al., 1999]) además de ser una con los mejores rendimientos de acuerdo con [Huang, 2008].

Para determinar el valor óptimo de K se hizo uso del *criterio del codo*. Esta medida permite maximizar la similitud entre objetos en el mismo grupo, minimizando *Sum Squared Errors*. La estadística SSE se calcula como:

$$SSE = \sum_{k=1}^{K} \sum_{\forall x_i \in G_k} (x_i - c_k)^2$$

Donde:

- G_k es el *k-ésimo* clúster.
- c_k es el centroide del *k-ésimo* clúster.
- x_i es la *i-ésima* observación.

De acuerdo con [Madhulatha, 2012], el *criterio del codo* indica que *«se debe elegir k de tal forma que al agregar otro clúster (k+1), esta última configuración no agregue suficiente información»*.

Por otro lado, [Bejar, 2020] nos presenta la siguiente forma de calcular k usando el *criterio del codo*: sea

$$L : y = ax + b$$

Donde:

- $a = \frac{SSE(G_1)}{1-m}$
- $b = \frac{nSSE(G_1)}{1-m}$
- m son los posibles clústers

Después, para todo $i \in \{1, ..., m\}$ se calculan las distancias D_i como:

$$D_i = d(SSE(i), L) = \frac{|a - SSE(i) + b|}{\sqrt{a^2 + 1}}$$

Por tanto, el valor de k obtenido por el *criterio del codo* está dado por:

$$k = max_i D_i$$

Utilizando el criterio anterior, **se obtuvo** $k = 7$**.**

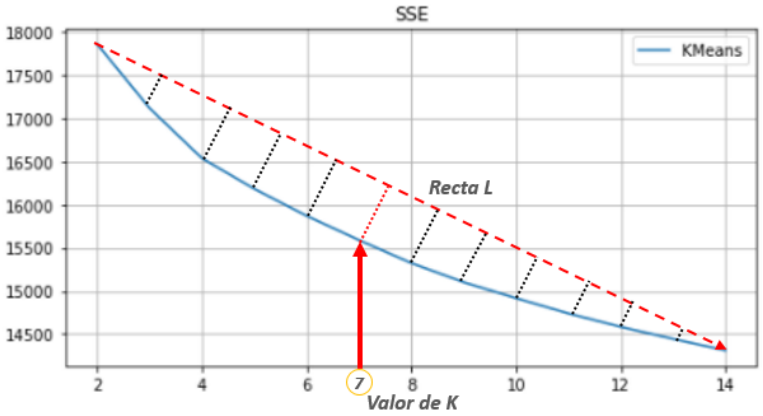


Figura 17: Elección del valor de k usando el *criterio del codo*. Elaboración propia basada en [Bejar, 2020] Fig 5.1

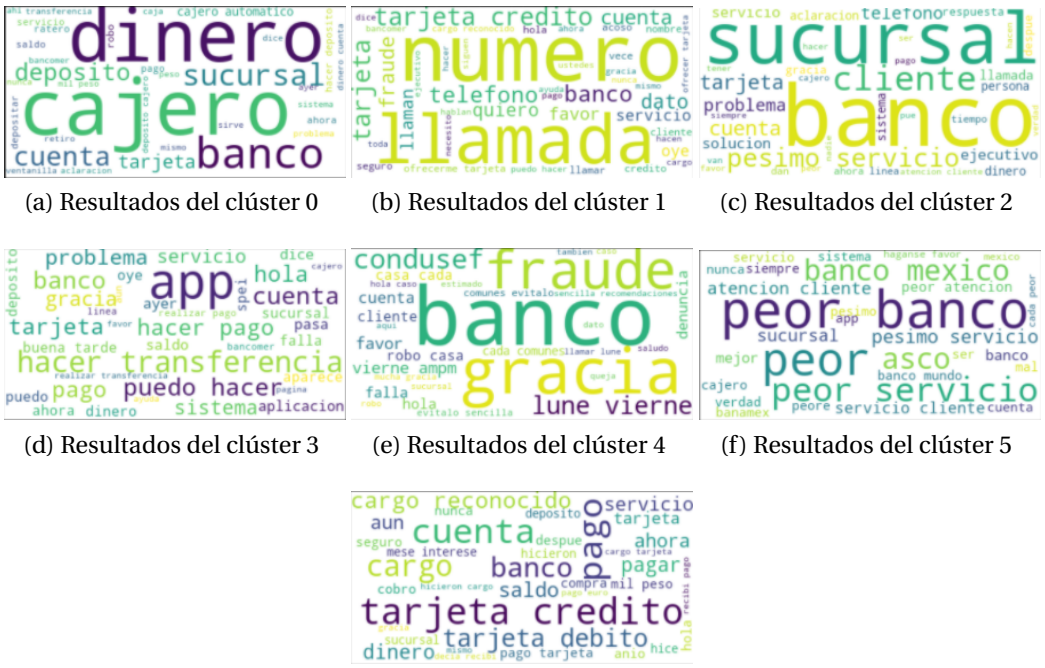
La Figura 17 muestra de forma gráfica la elección del valor k usando el *criterio del codo*.

La distribución de los clústers resultantes utilizando el algoritmo de *KMeans* con $k = 7$ y usando la similitud coseno se puede observar en Tabla 8.

Id de Clúster	**Número de elementos**
0	5,248
1	5,825
2	9,158
3	7,752
4	5,143
5	3,572
6	8,626

Cuadro 8: Distribución de clústers sobre los *tweets embeddings normalizados* usando *KMeans* con $k = 7$. Elaboración propia.

3.4.2.2. Evaluación de la interpretabilidad de los resultados de la detección de tópicos



(a) Resultados del clúster 0 (b) Resultados del clúster 1 (c) Resultados del clúster 2

(d) Resultados del clúster 3 (e) Resultados del clúster 4 (f) Resultados del clúster 5

(g) Resultados del clúster 6

Figura 18: Nubes de palabras obtenidas al aplicar *KMeans* a los *tweets embeddings normalizados*. Elaboración propia

Una vez obtenidos los Clúster, se construyeron las nubes de palabras con el fin de brindar una interpretación de la información contenida en cada uno de los subconjuntos resultantes.

De acuerdo a la Figura 18, los tópicos identificados en los *tweets* son los siguientes:

- Clúster 0. Este clúster agrupa quejas relacionadas a *cajero, dinero y banco*. La etiqueta de este conjunto de documentos será **Cajero.**
- Clúster 1. Este clúster agrupa quejas relacionadas a *número y llamadas*. La etiqueta de este conjunto de documentos será **Llamadas.**
- Clúster 2. Este clúster agrupa quejas relacionadas a *servicio, sucursal y banco*.La etiqueta de este conjunto de documentos será **Servicio al Cliente.**

- Clúster 3. Este clúster agrupa quejas relacionadas a *app, transferencias y sistemas*. La etiqueta de este conjunto de documentos será **Servicios Digitales.**
- Clúster 4. Este clúster agrupa quejas relacionadas a *banco y fraudes*. La etiqueta de este conjunto de documentos será **Fraudes.**
- Clúster 5. Este clúster agrupa documentos relacionadas a *peor, servicio y banco*. La etiqueta de este conjunto de documentos será **Reputación.**
- Clúster 6. Este clúster reporta quejas relacionadas a *tarjetas, crédito y débito*. La etiqueta de este conjunto de documentos será **Tarjetas.**

La etiqueta de cada tópico se guardó en la variable `tema`.

3.4.3. Identificación de subtópicos

3.4.3.1. Selección y ejecución de la técnica de modelado para la detección de tópicos

Con el objetivo de generar una mayor explicabilidad de las quejas, se generaron subtópicos a partir de los tópicos presentados anteriormente.

A diferencia de la detección de tópicos, en este caso se debe considerar una relación entre el tópico previamente detectado y sus *«hijos»*. Por tal motivo, se decidió ocupar un algoritmo de clustering jerárquico (*HCA* por sus siglas en inglés[4]).

De acuerdo con [Rokach and Maimon, 2005], existen dos versiones relevantes del *HCA*:

1. Aglomerativo. Inicia con clústers individuales y comienza a unir de dos en dos los que *más se parecen* hasta llegar a un macro clúster (en el caso extremo).
2. Divisivo. Parte de un macro clúster y comienza a dividir de dos en dos los elementos que *menos se parecen* hasta llegar a clústers individuales (en el caso extremo).

[4] HCA: Hierarchical Clustering Algorithm

En este caso, dado que ya se parte de un clúster específico (el tópico detectado), se decidió usar *HCA divisivo*.

Para generar los clústers, se decidió usar el *criterio de Ward* [Ward Jr, 1963], el cual genera clústers de tal forma que se reduzca la mayor cantidad de varianza (SSE) dentro de cada uno. El algoritmo dejará de dividir en el momento que la reducción de la varianza entre la cantidad de clústers actual y generar uno adicional sea insignificante (similar al *criterio del codo*).

La Figura 19 presenta de forma ilustrativa el dendograma obtenido para el tópico **Tarjeta**. Todos los dendogramas se pueden apreciar en el **Anexo 1.**

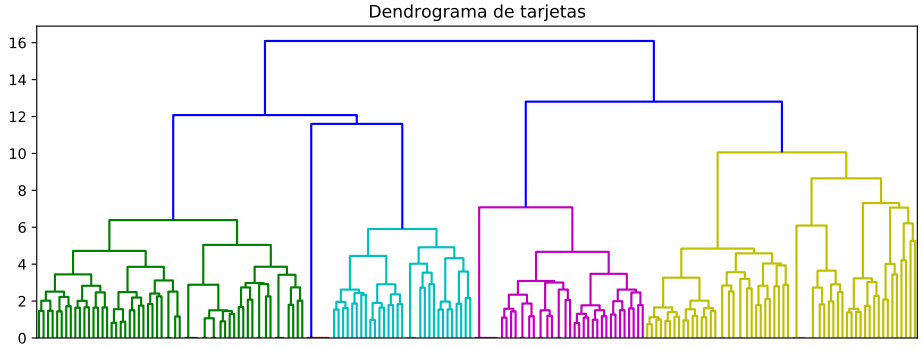


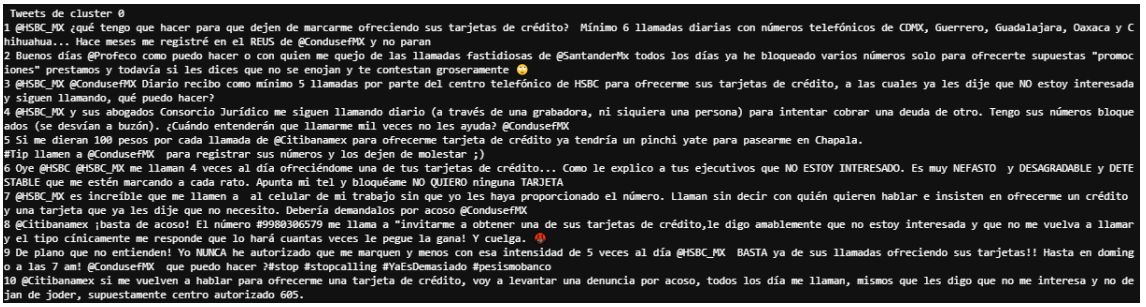
Figura 19: Dendograma obtenido al aplicar *HCA divisivo* al tópico **Tarjetas**. Elaboración propia.

3.4.3.2. Evaluación de la interpretabilidad de los resultados de la detección de subtópicos

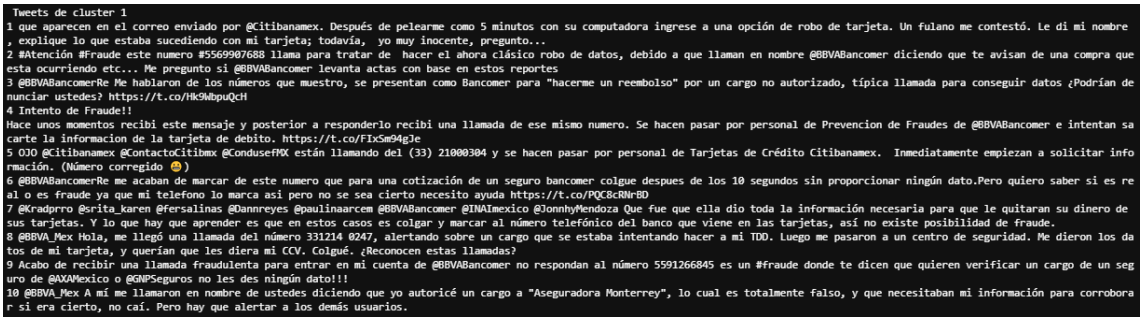
Para determinar la etiqueta de cada uno de los subtópicos generados se analizaron los 10 *tweets* más cercanos al centroide de cada subclúster. En la Figura 20 se muestra de manera ilustrativa un ejemplo de los *tweets* obtenidos para el tópico **Llamadas**.

Los resultados de los subtópicos fueron:

- Para el tópico **Cajero**, se obtuvieron los siguientes subclústers:
  - Subclúster 0. Son quejas relacionadas con problemas al realizar depósitos o



(a) Los 10 *tweets* más cercanos al centroide del primer subclúster del tópico **Llamadas**



(b) Los 10 *tweets* más cercanos al centroide del segundo subclúster del tópico **Llamadas**

Figura 20: Ejemplo de los 10 *tweets* más cercanos al correspondiente centroide para cada subclúster del tópico **Llamadas**. Elaboración propia

retiros en los cajeros automáticos. La etiqueta de este subclúster será **«Problemas con depósitos/retiros en cajeros»**.

- Subclúster 1. Son quejas relacionadas a la nula de funcionalidad de los cajeros automáticos. La etiqueta de este subclúster será **«El cajero no funciona»**.

- Subclúster 2. Son quejas relacionadas al robo por parte de los cajeros, esta queja menciona tanto cajeros automáticos, como al personal de la sucursal bancaria. La etiqueta de este subclúster será **«El cajero me robó»**.

- Subclúster 3. Son quejas relacionadas con problemas al retirar dinero en los cajeros automáticos pues no otorgan el efectivo. La etiqueta de este subclúster será **«El cajero se quedó mi dinero»**.

- Para el tópico **Llamadas**, se obtuvieron los siguientes subclústers:

  - Subclúster 0. Son quejas relacionadas a la frecuencia en las llamadas recibidas por parte de los bancos con el fin de vender sus productos. La etiqueta

de este subclúster será **«Llamadas para ofrecer productos»**.

- Subclúster 1. Son quejas reportando llamadas fraudulentas en nombre de la institución financiera. La etiqueta de este subclúster será **«Llamadas fraudulentas»**.

- Para el tópico **Tarjetas**, se obtuvieron los siguientes subclústers:

  - Subclúster 0. Son quejas relacionadas con cargos no reconocidos en tarjetas de crédito y tarjetas de débito. La etiqueta de este subclúster será **«Cargos no reconocidos en TDC y TDD»**.

  - Subclúster 1. Son quejas relacionadas a la falta de cumplimiento en promociones y ofertas con tarjeta de crédito. La etiqueta de este subclúster será **«Promociones en TDC y TDD»**.

  - Subclúster 2. Son quejas relacionadas al tiempo de atención respecto a las aclaraciones de tarjeta de crédito y débito. La etiqueta de este subclúster será **«Aclaraciones TDC y TDD»**.

  - Subclúster 3. Son quejas relacionadas con problemas al realizar el pago a la tarjeta de crédito. La etiqueta de este subclúster será **«Pago en TDC»**.

  - Subclúster 4. Son quejas relacionadas con problemas a las comisiones cobradas en relación a la tarjeta de débito. La etiqueta de este subclúster será **«Comisiones TDD»**. Cabe resaltar que este subclúster se compone de una serie de *retweets* donde el usuario se queja de una comisión que tuvo que pagar a BBVA por recibir un depósito en euros.

- Para el tópico **Servicios digitales**, se obtuvieron los siguientes subclústers:

  - Subclúster 0. Son quejas relacionadas a problemas con los sistemas al momento de querer realizar transferencias interbancarias. La etiqueta de este subclúster será **«Transferencias en canales digitales»**.

  - Subclúster 1. Son quejas relacionadas a la falla de la app móvil y banca en línea. La etiqueta de este subclúster será **«Falla en servicios digitales»**.

- Para el tópico **Servicio al cliente**, se obtuvieron los siguientes subclústers:
    - Subclúster 0. Son quejas relacionadas a un mal servicio por parte del personal del banco. La etiqueta de este subclúster será **«Mal servicio»**.
    - Subclúster 1. Son quejas relacionadas a la falta de solución de problemas reportados. La etiqueta de este subclúster será **«No soluciona problemas»**.
    - Subclúster 2. Son quejas relacionadas al tiempo de atención en sucursales bancarias. La etiqueta de este subclúster será **«Atención lenta en sucursal»**.
    - Subclúster 3. Son quejas relacionadas al tiempo de espera para ser atendidos en el *Call Center*. La etiqueta de este subclúster será **«Tiempo de espera en *call center*»**.
    - Subclúster 4. Son quejas relacionadas a la falta de solución y seguimiento de problemas reportados. La etiqueta de este subclúster será **«No resuelve problemas»**.
- Para el tópico **Fraudes**, se obtuvieron los siguientes subclústers:
    - Subclúster 0. Son *tweets* de seguimiento a reportes de fraude en línea. A diferencia de los demás subclústers, este no subclúster no refleja quejas, más bien muestra seguimiento y atención al cliente tanto de los bancos como de la CONDUSEF. La etiqueta de este subclúster será **«Seguimientos en línea»**.
    - Subclúster 1. Son acusaciones directas a la institución financiera o bien a alguno de sus trabajadores. La etiqueta de este subclúster será **«Acusaciones de fraude»**.
    - Subclúster 2. Son quejas respecto a recibir mensajes fraudulentos por parte del banco. La etiqueta de este subclúster será **«Mensajes fraudulentos»**.
    - Subclúster 3. Son respuestas de la `Condusef` de los *tweets* en los que fueron etiquetados. La etiqueta de este subclúster será **«Respuestas Condusef»**.
    - Subclúster 4. Son quejas relacionadas al robo de identidad. La etiqueta de este subclúster será **Robo identidad**.

- Subclúster 5. Este subclúster se compone de una serie de *retweets* donde se reporta una falla en el sistema de Santander. La etiqueta de este subclúster será **«Falla sistema Santander»**.
- Subclúster 6. Es una serie de *retweets* de BBVA para prevenir el robo a casa habitación. La etiqueta de este subclúster será **«Tips BBVA»**.

- Para el tópico **Reputación**, se obtuvieron los siguientes subclústers:
  - Subclúster 0. Son ataques directos a la institución etiquetándolo como *«el peor banco»*. La etiqueta de este subclúster será **« El peor banco»**.
  - Subclúster 1. Son quejas relacionadas al lento servicio que ofrece la institución. La etiqueta de este subclúster será **«Servicio lento de la institución»**.

En la Figura 21 se muestra de manera gráfica un resumen de los tópicos y subtópicos encontrados en los *tweets*.

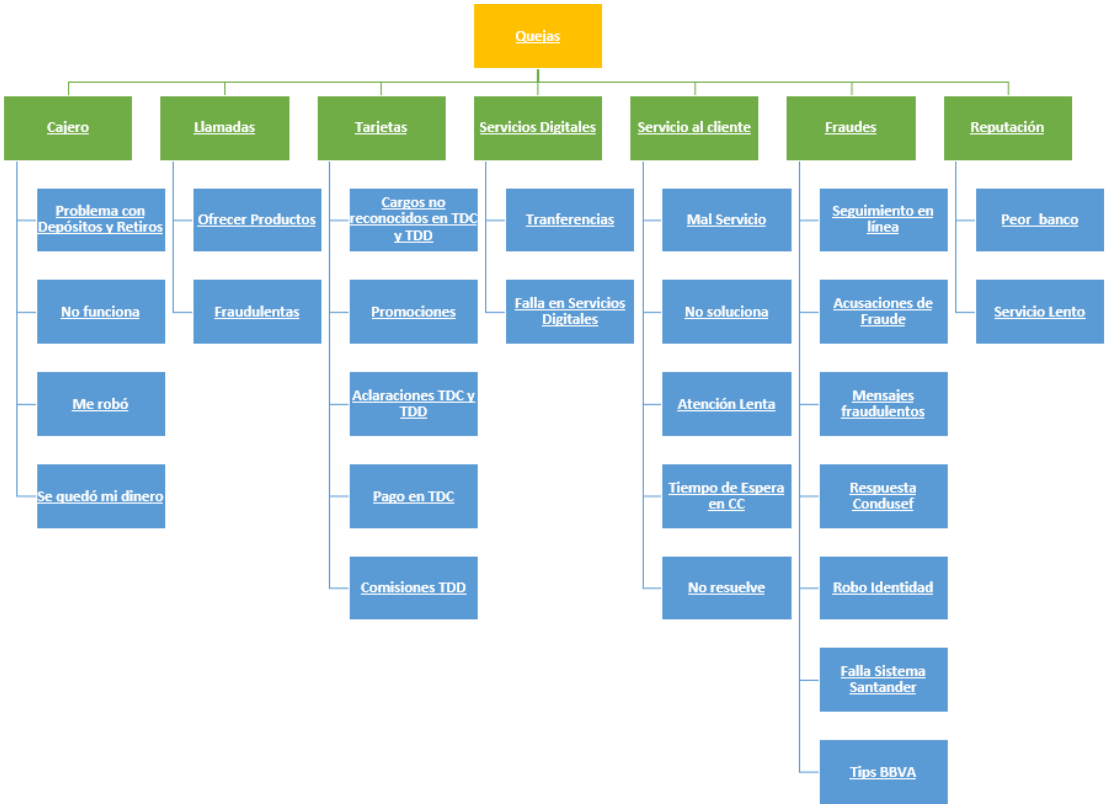


Figura 21: Resumen de tópicos y subtópicos encontrados en los *tweets*. Elaboración propia

Es importante mencionar que a pesar de que los subclúster pueden tener con-

tenido o información similar (por ejemplo el subclúster 1 y subclúster 4 del tópico *Servicio al cliente*) el modelo lo divide en grupos diferentes pues por las palabras y el contexto utilizado en cada uno, lo anterior sucede porque una misma idea puede ser expresada con diferentes palabras.

3.5. Resumen

A lo largo de esta capítulo se discutió sobre el interés de las instituciones financieras en conocer los puntos de dolor (quejas) de sus clientes y cómo pueden utilizar esta información para incrementar su rentabilidad. Así mismo, se presentó la información reportada por la CONDUSEF referente a las controversias reportadas durante el primer trimestre de 2020 y su comparativo con el primer trimestre de 2019. Por último, se mostró la relevancia de las redes sociales dentro del día a día de la sociedad y se eligió a *Twitter* como fuente de datos complementaria para el proyecto.

Además se presentó la forma en cual se realizó la extracción de datos utilizando la `API` *stream* de *Twitter*. Adicional, se describieron las variables obtenidas, haciendo especial énfasis en las variables `user.id, created_at, place.country, place.full_name, truncated, extended_tweet.full_text` y `text`. Luego, a través de modificaciones básicas del texto fue posible determinar que la mayoría de las quejas están relacionadas a *«tarjetas»*, *«pagos»* y *«servicios»*. Por último, se observó que solo el 21.96 % de los datos contiene información relacionada a la geo localización.

Así mismo realizó la preparación final de los datos para el análisis: se eligieron los campos base a partir de los cuales se construyeron nuevas columnas tales como `fecha2, estado, municipio` y `pais`. Mediante el uso de expresiones regulares y reglas de asociación se determinó la institución bancaria a la cual va dirigido cada *tweet*. Se conservó aquellos *tweets* que están dirigidos a los bancos del G7. Por otro lado, se aplicó una limpieza y trasformación a los textos con el objetivo de mantener aquellas palabras y símbolos (*emojis*) relevantes. Al final, se generó el conjunto de datos final sobre el cual se realizará la detección de quejas de los usuarios de servicios financie-

ros.

Por último se aplicó un algoritmo de *word embeddings* usando `fastText` para obtener la representación vectorial de los *tweets*. Adicional, mediante el uso de *KMeans*, se generaron tópicos sobre el texto procesado y se categorizaron para darle un contexto respecto al planteamiento del problema. Por último, con el objetivo de poder tener mayor visibilidad de las quejas de los usuarios, a cada tópico se le aplicó el algoritmo *HCA divisivo* generando un total de 27 subtópicos, mismos que cuentan con una explicación contextual.

Con todo lo anterior, es posible generar análisis temporales-espaciales por institución bancaria, tipo de queja (tópico) y queja particular (subtópico). Dichos análisis serán presentados en el **capítulo 4. Resultados**.

Capítulo 4

Resultados

Capítulo 4. Resultados

Este capítulo esta dedicado a la exploración de la información identificada como quejas que los usuarios han manifestado en redes sociales. Es posible realizar la exploración de las quejas desde diferentes puntos de vista como tiempo, espacio, tipo o institución financiera. Así mismo, es posible responder a preguntas relacionadas a los *tweets* como: ¿cuándo hay más quejas? ¿Qué estado de la República Mexicana concentra la mayor cantidad de quejas? ¿Qué institución financiera recibe más quejas? ¿Cuál es la principal queja de los usuarios de servicios financieros? ¿Existe concordancia con los datos presentados por la CONDUSEF?.

La Sección 4.1 está enfocada a analizar la cantidad de quejas en *Twitter*; se realizarán exploraciones temporales, espaciales y por institución financiera, con lo cual se busca dar un contexto general sobre la distribución del número de quejas analizadas. A lo largo de la Sección 4.2 se realizará el análisis por tipo de quejas e institución financiera para poder descubrir cuál es el punto de dolor más grande que tienen los clientes con cada banco. Por último, en la Sección 4.3 se muestra un comparativo de los resultados obtenidos en contraste con los presentados por la CONDUSEF.

4.1. Análisis descriptivo de la distribución de quejas en *Twitter*

Antes de generar un análisis granular por tipo de queja detectada en *Twitter* es necesario presentar un contexto general de los datos analizados, motivo por el cual es necesario mostrar un análisis descriptivo de los datos. Dicho análisis consta de tres vertientes: temporal, bancaria y espacial.

4.1.1. Distribución temporal de las quejas en *Twitter*

Para entender el comportamiento de las quejas de los clientes de instituciones financieras, una variable importante a considerar es el **momento** en el que se realiza dicha queja pues nos permite apreciar qué tanto han aumentado las quejas. Adicionalmente es posible observar comportamientos que ayuden a prevenir la misma, o bien, estar preparados para resolverla.

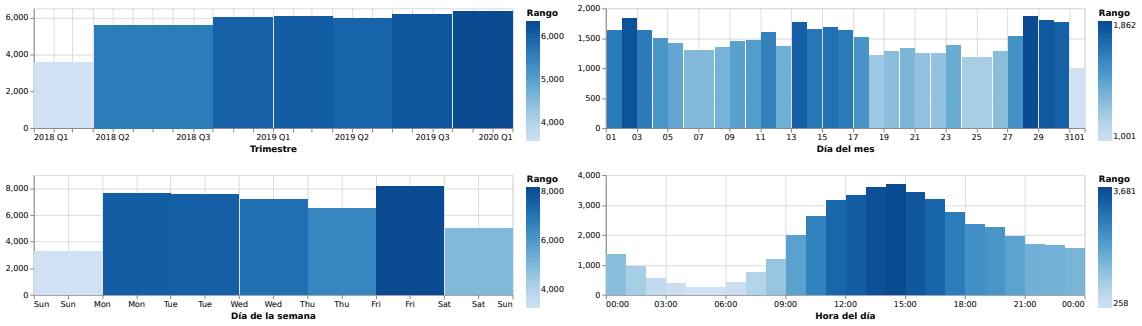


Figura 22: Análisis descriptivo temporal del número de quejas en *Twitter*. Elaboración propia.

La Figura 22 muestra la distribución de quejas por trimestre del año, día del mes, día de la semana y hora del día. A partir de dicha información surgen las siguientes observaciones:

1. Las quejas tienen una tendencia creciente de 2018 a 2019. Este fenómeno puede explicarse tanto por la evolución tecnológica y una mayor adopción de las redes sociales en la vida diaria [Ayala, 2014], como por un creciente número de clientes bancarizados en México [Orozco, 2011].

2. Los días donde se presentan más quejas es a inicios (o finales) y mediados de mes, fechas que coinciden con los días de pago de la mayoría de las empresas en el país (quincenal).

3. Los días que se presenta una mayor actividad de quejas son los viernes, seguidos por lunes y martes. Por otro lado, los días con menor actividad son los fines de semana, particularmente los domingos.

4. Las horas con mayor actividad de quejas es entre 10:00 horas a 19:00 horas. Este

umbral de tiempo coincide con la mayoría de los horarios de oficina del país.

4.1.2. Distribución por institución financiera de las quejas en *Twitter*

Otra variable relevante es la **institución financiera a la cual va dirigida la queja**, este tipo de información será relevante tanto para el banco *per se*, pues será capaz de **medir** la cantidad de quejas y generar nuevos «indicadores claves de comportamiento» (*KPI* [1]), como para los usuarios de servicios bancarios ya que se pueden tomar una mejor decisión al momento de contratar algún servicio o producto de dichas instituciones.

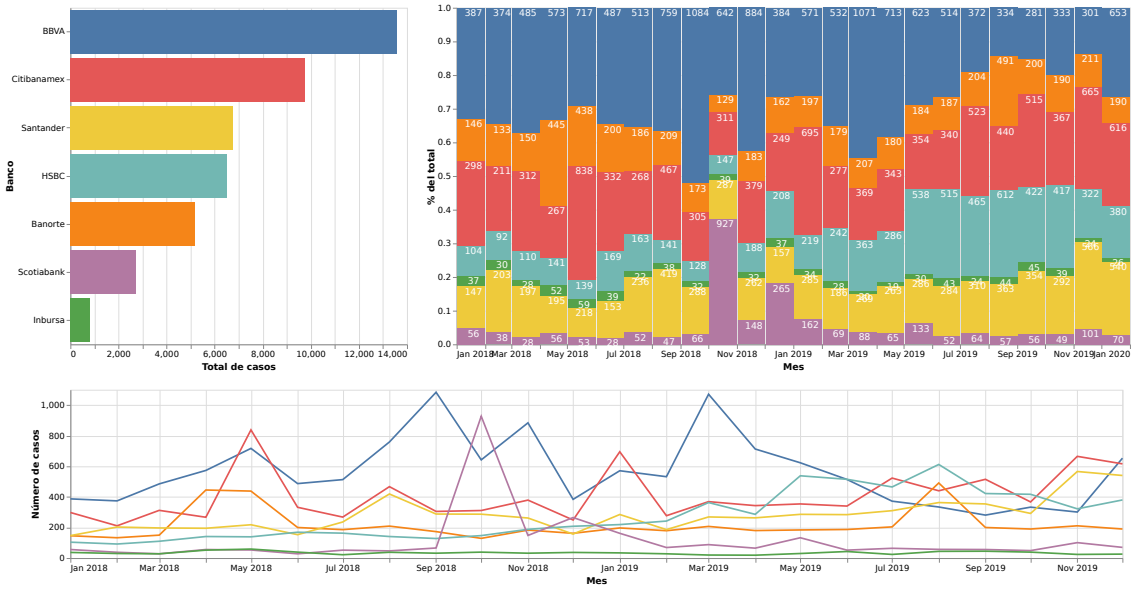


Figura 23: Análisis descriptivo por institución financiera del número de quejas en *Twitter*. Elaboración propia.

La Figura 23 muestra la distribución de quejas por institución financiera, participación del banco en las quejas de cada trimestre del año y la frecuencia de quejas por mes e institución. Se puede notar lo siguiente:

1. BBVA es el banco que presenta mayor cantidad de quejas durante el periodo de tiempo, seguida por Citibanamex y Santander. Mientras que Inbursa es la institución con menor número de quejas asociadas.

[1] KPI: Key Performance Indicator

2. Durante Noviembre 2018, Scotiabank tuvo una alza importante de quejas, misma que no volvió a tener en el periodo de tiempo analizado.

3. De marzo a noviembre 2019, las quejas de BBVA disminuyeron notablemente, mientras que Santander, HSBC y Banamex fueron a la alza.

Es importante mencionar que no es posible obtener de forma directa el *número de clientes* de cada institución financiera (pues son datos privados de cada banco), no obstante con el objetivo de realizar una comparación imparcial entre las instituciones financieras, se puede considerar el *tamaño de cada banco*. Para determinar el **tamaño del banco** se utilizó el indicador de **cartera vigente**, mismo que de acuerdo con [CNBV, 2020a] *«representa a todos los usuarios que están al corriente en los pagos del crédito que han adquirido, tanto del monto original como de los intereses»*. Este dato se representa en millones de pesos y por regulación de la Comisión Nacional Bancaria y de Valores, todos los bancos de México están obligados a reportar esta cifra en sus *Estados financieros*.

Entonces, considerando la cartera vigente, se generó el indicador ***Tasa de quejas del banco*** x ***durante el periodo*** t (Q_x^t) como:

$$Q_x^t = \frac{Quejas_x^t}{Cartera\ Vigente_x^t} \times 100000.$$

El indicador Q_x^t nos representa la cantidad de quejas del banco x por cada 100,000 millones de pesos de cartera vigente durante el periodo de tiempo t.

La Figura 24 presenta un comparativo entre el tamaño del banco y la cantidad de quejas recibidas de 2018 a 2019.

A partir de la Figura 24 se pueden realizar las siguientes observaciones:

- Al analizar la primer gráfica de la Figura 24, se nota que no existe una variación importante en el tamaño de las carteras activas de las instituciones financieras del G7 en los últimos dos años.

- Así mismo, el banco con mayor cartera vigente es BBVA seguida por Banorte,

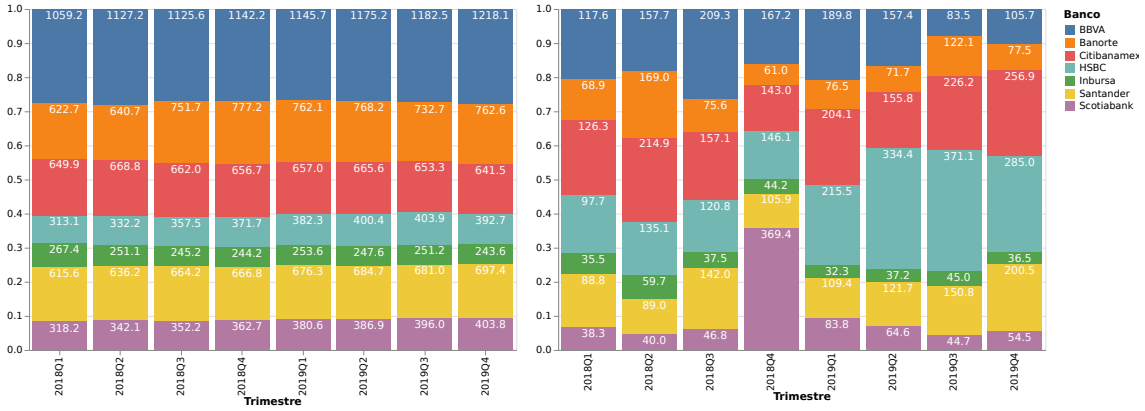


Figura 24: Comparativo trimestral por institución financiera del valor de la cartera activa y Q_x^t de 2018 a 2019. Elaboración propia con base a la información presentada en [CNBV, 2020b, Scotiabank, 2020, Santander, 2018]

Santander y Citibanamex. En contraste, la institución con menor cartera vigente es Inbursa.

- Por último, dado que **BBVA es el banco más grande del G7** es natural que concentre la mayor cantidad de quejas en *Twitter*, sin embargo al analizar el comportamiento de Q_x^t en la segunda gráfica de la Figura 24, se observa un comportamiento crítico para Citibanamex y HSBC, pues su tasa de quejas por cada 100,000 millones de cartera vigente es mucho mayor en comparación a otros bancos con tamaños de cartera equiparables como Scotiabank y Banorte.

4.1.3. Distribución espacial de las quejas en *Twitter*

Por último, para poder saber *¿dónde se concentran la mayor cantidad de quejas?* se realizó un análisis espacial. La distribución espacial de quejas se presenta en la Figura 25.

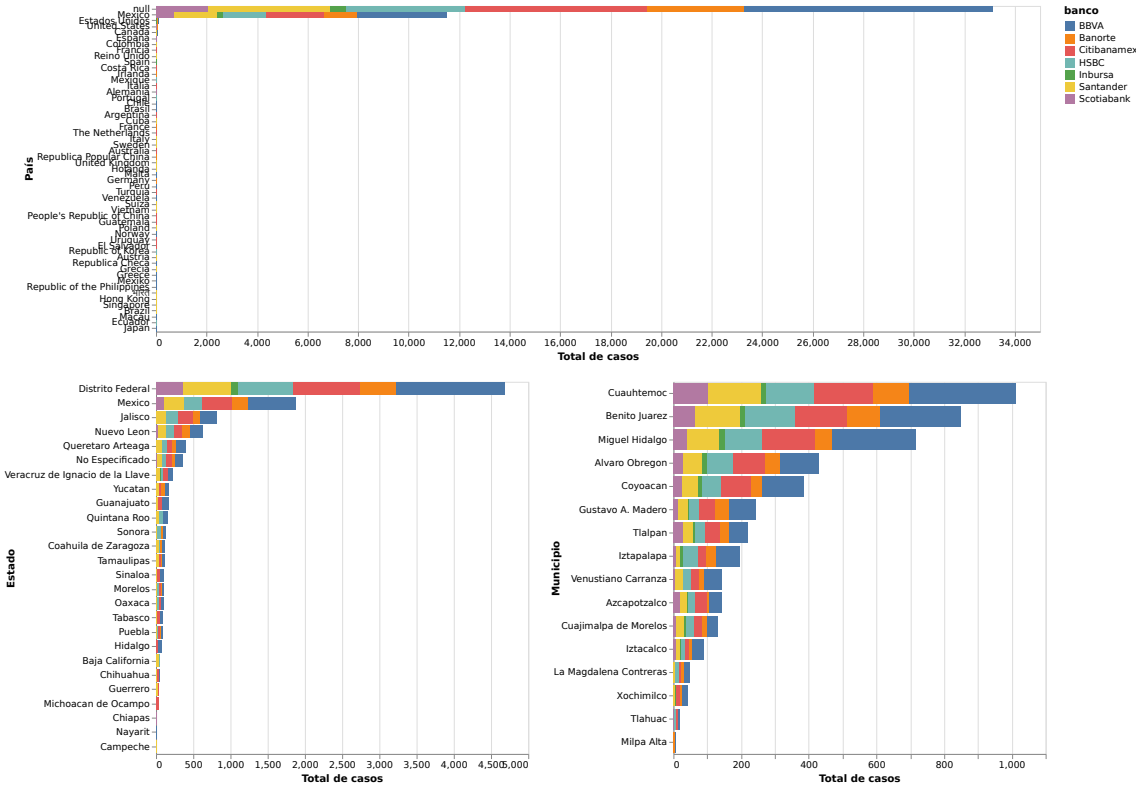


Figura 25: Análisis descriptivo espacial del número de quejas en *Twitter*. Elaboración propia.

Con base a la información presentada en la Figura 25, es importante resaltar:

1. Un total de 33,115 (73.06 %) de los *tweets* no cuentan con valor alguno en la variable de *País* que permita ubicarlos geográficamente. Este resultado se anticipaba pues se estudió durante el **Capítulo 3**.

2. De los 12,209 de los *tweets* restantes, un 96.15 % (11,740) pertenecen a México por lo cual se conjetura que el resto son mexicanos en el extranjero o bien, extranjeros con capital en México.

3. De los *tweets* que corresponden a México, los *tweets* de los estados: CDMX (4,681),

Estado de México (1,126), Jalisco (566), Nuevo León (393) y Querétaro (291) suman el 72.15 % de las quejas a nivel nacional.

4. Dentro de la CDMX, las alcaldías que más quejas generan (en orden descendente) son: Cuauhtémoc, Benito Juárez, Miguel Hidalgo, Álvaro Obregón y Coyoacán.

Ahora que existe un contexto general, es importante contestar las preguntas:

- ¿De qué se quejan los usuarios de servicios bancarios?
- ¿Qué banco otorga un mejor servicio?
- ¿Cuáles son las áreas de oportunidad para cada institución bancaria?

4.2. Análisis por tipos de quejas detectadas en *Twitter*

Gracias a los esfuerzos realizados durante el **Capítulo 3** es posible conocer los *puntos de dolor* (tópicos) y de las quejas (subtópicos) que tienen los usuarios de servicios financieros con las instituciones que pertenecen al G7. Lo anterior permite analizar los resultados obtenidos y determinar acciones a tomar con el objetivo de brindar un mejor servicio a los clientes.

Con el objetivo de tener un punto de partida, la Figura 26 presenta los resultados de los puntos de dolor y quejas (tópicos y subtópicos) detectados en *Twitter* de forma temporal y espacial, con lo cual se pueden realizar las siguientes observaciones:

- Las dos quejas más frecuentes son:
    1. Problemas al realizar transferencias mediante el uso de canales digitales.
    2. Cargos no reconocidos en tarjetas.
- Durante 2019, las quejas relacionadas a **Servicios Digitales** han tenido una tendencia creciente. Por otro lado, las quejas relacionadas con **Tarjetas** han disminuido durante el mismo periodo de tiempo.

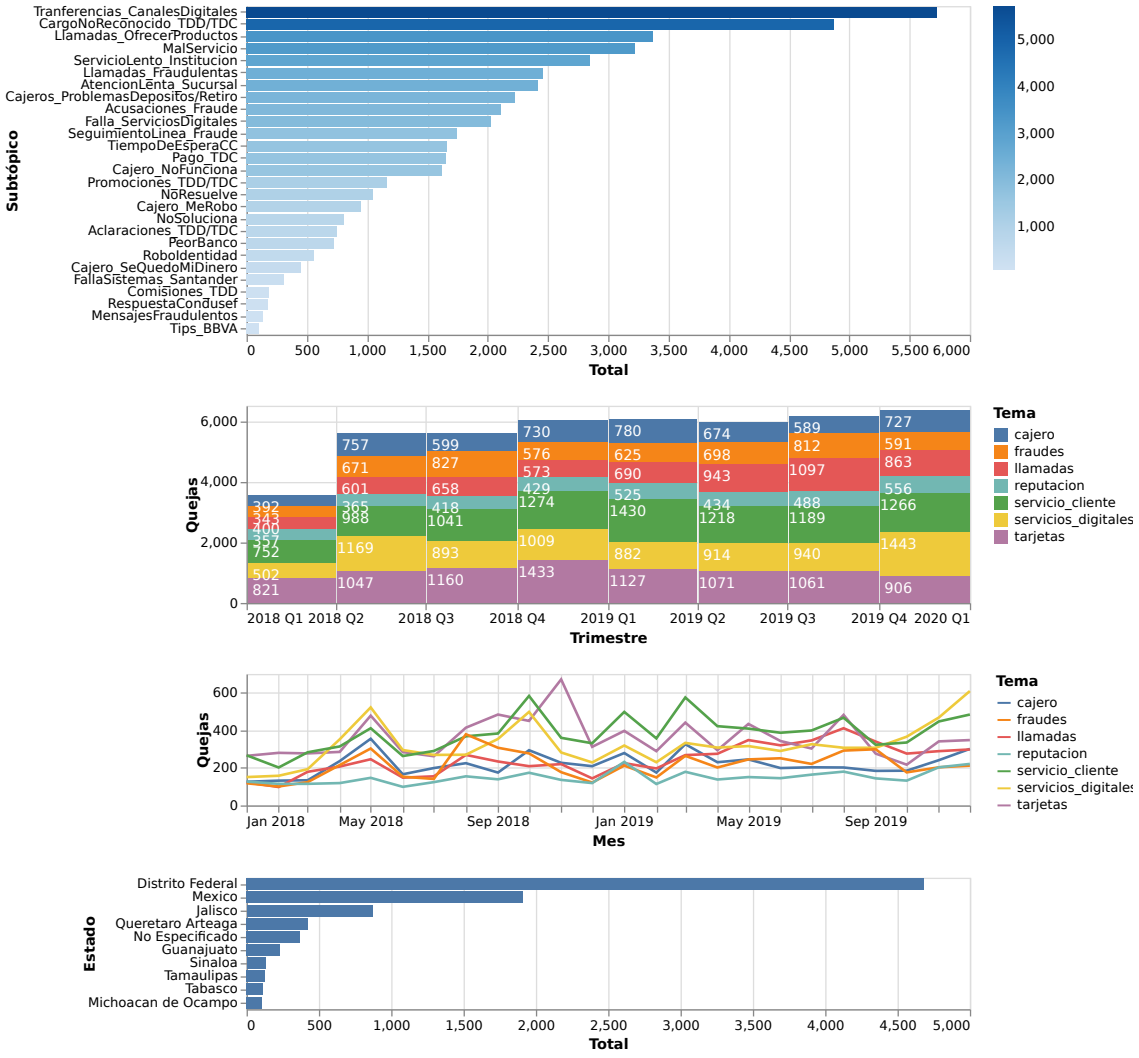


Figura 26: Análisis temporal y espacial por tipo de queja encontrada en *Twitter*. Elaboración propia.

- En general, los puntos de dolor más mencionados en *Twitter* son:
    - **Tarjetas.**
    - **Servicios Digitales.**
    - **Servicio al Cliente.**

Durante las siguientes secciones se presentan las quejas por cada uno de los bancos de forma individual así como el comparativo entre las instituciones pertenecientes al G7.

4.2.1. Análisis de las quejas detectadas en *Twitter* por institución financiera

A continuación se presentan los resultados obtenidos por cada una de los bancos pertenecientes al G7, las figuras relacionadas a cada institución financiera se pueden encontrar en el **Anexo B**.

De la Figura 43 se puede observar que la queja principal para BBVA se detona al realizar o recibir transferencias interbancarias mediante el uso de canales digitales, seguido por problemas derivados de cargos no reconocidos en las tarjetas. En general, las problemáticas relacionadas con **Canales Digitales** han aumentado su concurrencia durante el último trimestre de 2019. Sin embargo, también se puede notar un comportamiento decreciente en el número de quejas dirigidas a BBVA desde el tercer trimestre de 2018, en particular se nota un decrecimiento en las quejas relacionadas con tarjetas a partir del cuarto trimestre de 2019. Además, se observa que las quejas provienen principalmente de **CDMX, Estado de México y Jalisco**. Por último, es posible inferir un problema con las tarjetas durante los meses de septiembre y noviembre 2019.

En la Figura 44 es importante observar que para Citibanamex no existe una diferencia importante entre las tres principales quejas de los clientes: **enviar o recibir transferencias en canales digitales, cargos no reconocidos en tarjetas** y **atención lenta por parte de la institución**. Además, a lo largo del periodo estudiado se presenta un **crecimiento en el número de quejas** dirigidas a este banco, principalmente en proble-

máticas relacionadas con los canales digitales y servicio al cliente. La principal fuente de quejas proviene de **CDMX, Estado de México y Jalisco**. Por último, es posible inferir fallas generales en el banco durante los meses de mayo 2018 y enero 2019.

Con la información presentada en la Figura 45 es posible observar que la queja principal de los clientes de Scotiabank está relacionada con problemas al realizar o recibir transferencias mediante el uso de canales digitales. Por otro lado, se nota un comportamiento atípico en las quejas pues se nota un crecimiento muy abrupto en octubre 2018, mismo que desaparece en noviembre del mismo año y regresa a los valores regulares. Este crecimiento podría estar relacionado con un problema general en los sistemas del banco o bien un ataque a través de *Twitter*[2].

Con ayuda de la Figura 46 se observa que las dos principales quejas de Banorte son relacionadas a: problemas al enviar y recibir transferencias usando los canales digitales y cargos no reconocidos en tarjetas. En general, las áreas de oportunidad están enfocadas en tarjetas, servicios digitales y servicio al cliente. Además, se notan dos momentos en el tiempo que acumulan más quejas en comparación a su historia (abril-mayo 2018 y agosto 2019). Sin embargo, no se observa una tendencia creciente o decreciente en el número de quejas durante el periodo analizado. Las quejas provienen principalmente de **CDMX, Estado de México y Nuevo León**.

La Figura 47 muestra los resultados obtenidos para Inbursa, de donde es importante observar que la principal queja de este banco es relacionada a **cargo no reconocido** seguida por problemas en **transferencias al usar servicios digitales**. Su punto de dolor más grande está relacionado a tarjetas, sin embargo no existe una tendencia creciente o decreciente para las quejas recibidas. Primordialmente tiene presencia en la **CDMX**.

En la Figura 48 se observa un comportamiento muy particular, pues la principal queja de los clientes de HSBC son las constantes llamadas de este banco para ofrecer sus productos y servicios. Se puede observar un crecimiento acelerado en las quejas de 2018 a 2019 siendo particularmente alarmante aquellas relacionadas con llamadas y servicio al cliente. Las quejas recibidas son principalmente de **CDMX**.

[2] Al momento de este análisis no es claro el origen de este fenómeno y requiere un análisis posterior

Por último, en la Figura 49 se puede apreciar que las principales quejas de los clientes de Santander son relacionada **cargos no reconocidos en tarjetas** y **fallas en servicios digitales para enviar o recibir transferencias interbancarias**. Se puede observar un comportamiento creciente en el número de quejas recibidas durante 2018 y 2019, siendo servicios digitales y atención al cliente los principales puntos de dolor. Las quejas provienen principalmente de **CDMX** y **Estado de México**. Derivado del incremento de quejas en agosto 2018, se puede inferir alguna falla generalizada durante ese mismo mes.

4.2.2. Comparación de los resultados entre instituciones financieras

Una vez presentados los resultados por cada banco, se prosigue a comparar las quejas detectadas por instituciones financieras pertenecientes al G7.

De forma análoga a como se hizo en la Sección 4.1, con el fin de realizar una comparación imparcial entre instituciones financieras, se generó el indicador ***Tasa de quejas del banco*** x (Q_x) como:

$$Q_x = \frac{Quejas_x}{\overline{Cartera\ Vigente}_x} \times 100000.$$

Donde, $\overline{Cartera\ Vigente}_x$ es la media aritmética de la cartera vigente para el banco x de 1 de enero 2018 a 31 de diciembre 2019.

El indicador Q_x nos representa la cantidad de quejas del banco x por cada 100,000 millones de pesos de cartera vigente promedio durante 2018 y 2019.

En la Figura 27 se puede apreciar que los bancos que más quejas reciben relacionadas a los cajeros y funcionamiento de los mismos son BBVA, Citibanamex y Santander, tanto por número total de quejas como al normalizar las quejas utilizando la cartera vigente. Mientras para BBVA y Citibanamex la principal queja de los cajeros son los **problemas con depósitos o retiros** la queja más común, para Santander la principal queja es **la falta de funcionamiento de los mismos**. En contraste, Inbursa es la institución que menos quejas recibe en cuanto a cajeros. Por otro lado, en relación

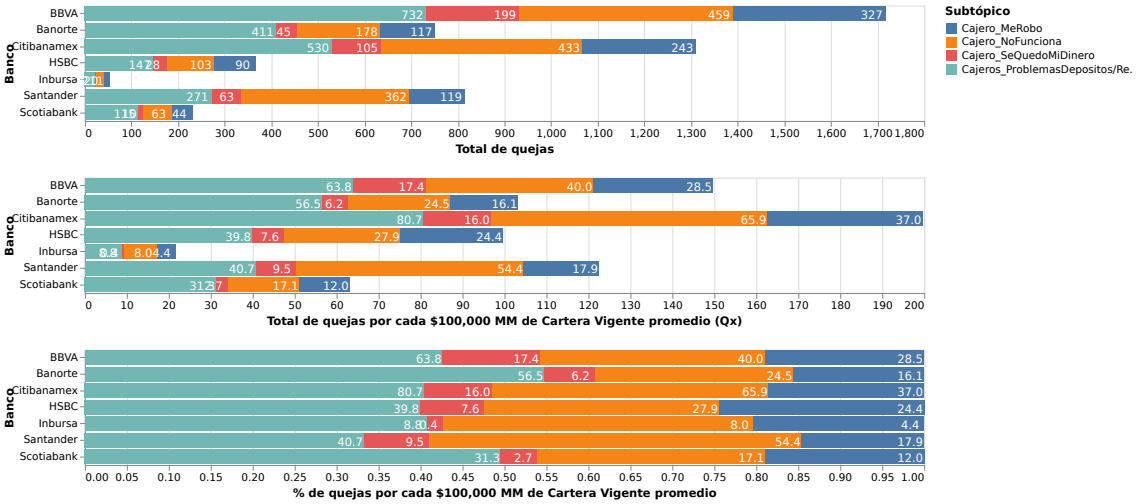


Figura 27: Comparación de resultados de todas las instituciones financieras relacionadas con quejas de **Cajeros**. Elaboración propia.

a la proporción de quejas recibidas de este tópico, Banorte y Scotiabank son los bancos con los cuales los clientes tienen más problemas al realizar depósitos o retiros en los cajeros automáticos. Una vez más, Santander es el banco que en proporción recibe más quejas por falta de funcionamiento.

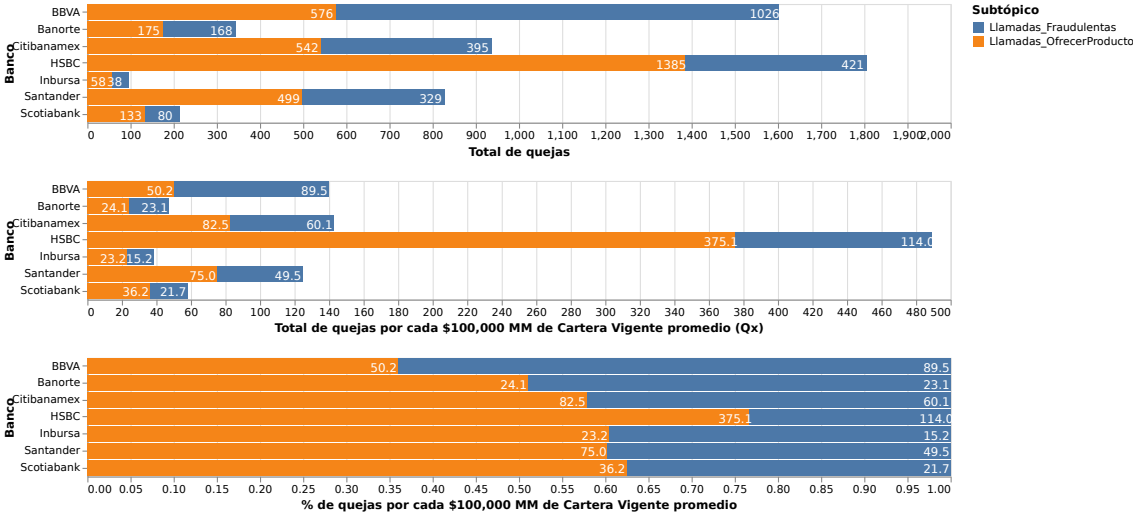


Figura 28: Comparación de resultados de todas las instituciones financieras relacionadas con quejas de **Llamadas**. Elaboración propia.

Así mismo, en la Figura 28 se observa que los banco con más quejas en el rubro de llamadas, por cantidad de quejas son: HSBC, BBVA y Citibanamex. No obstante, al considerar la tasa de quejas Q_x, HSBC es el banco más representativo para este tópico, siendo la queja principal las constantes **llamadas para ofrecer productos y servicios**,

este es un hallazgo crítico para HSBC. Por otro lado, se puede notar que la cantidad de quejas relacionadas a llamadas para BBVA y Citibanamex tiene una composición muy similar al utilizar la tasa de quejas por cada 100,000 millones de pesos en cartera vigente: mientras que los clientes de BBVA se quejan más de recibir llamadas fraudulentas, los clientes de Citibanamex también se quejan las llamadas recibidas para ofrecer productos y servicios. Es importante mencionar que los usuarios de BBVA reciben una mayor cantidad de llamadas fraudulentas. Una vez más, Inbursa es el banco que recibe menor cantidad de quejas además de poseer la tasa de quejas más baja.

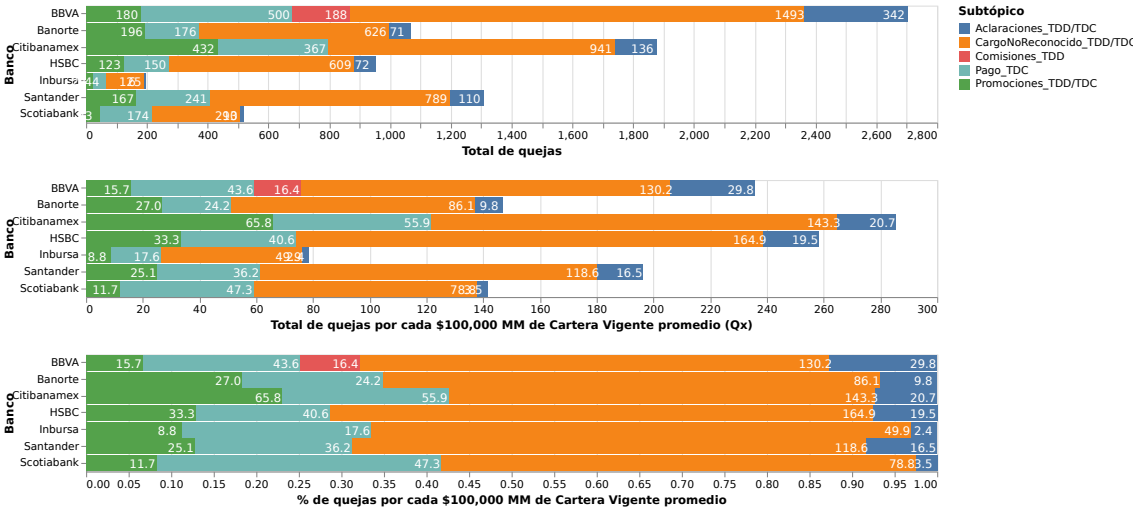


Figura 29: Comparación de resultados de todas las instituciones financieras relacionadas con quejas de **Tarjetas**. Elaboración propia.

En la Figura 29 al considerar el total de quejas recibidas por tarjetas de crédito o débito, se puede notar que los bancos que más quejas reciben (por cantidad de quejas) son BBVA, Citibanamex y Santander. No obstante, al considerar la tasa de quejas por cada 100,000 millones de pesos en cartera vigente, la instituciones que muestran mayores problemas al respecto son: Citibanamex, HSBC y BBVA. En general se observa que todos los clientes se quejan de tener **cargos no reconocidos en sus tarjetas**. Al omitir el punto anterior y analizar los datos de forma relativa a la cantidad de quejas recibidas, los clientes de BBVA reportan problemas en las aclaraciones en sus tarjetas de crédito y débito, mientras que los clientes de Citibanamex se quejan de las promociones en las tarjetas de crédito y los clientes de Scotiabank reportan problemas al realizar el pago de sus tarjetas de crédito.

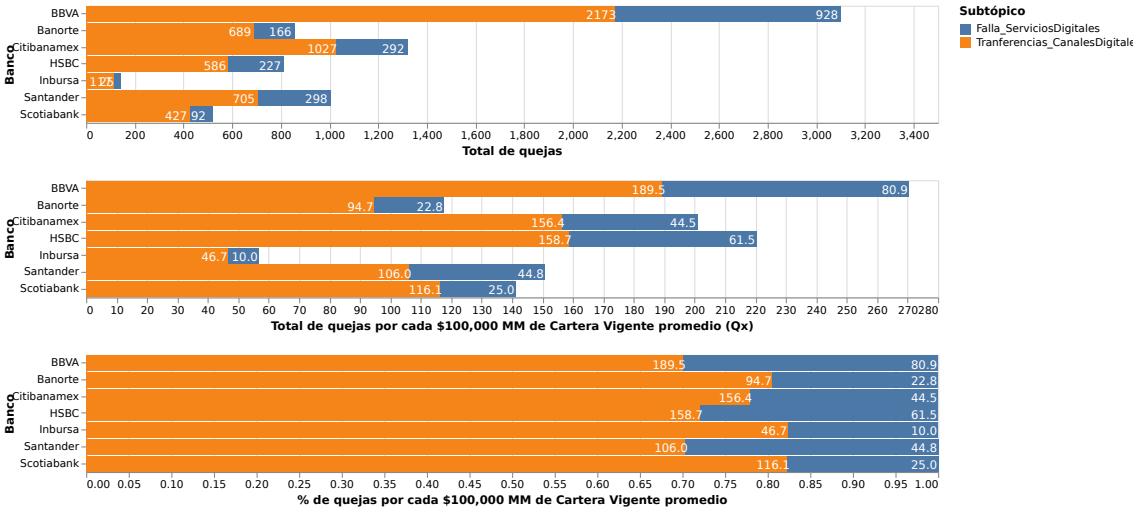


Figura 30: Comparación de resultados de todas las instituciones financieras relacionadas con quejas de **Servicios Digitales**. Elaboración propia.

Ahora bien, en la Figura 30 se muestra que la institución con más quejas relacionadas a servicios digitales es BBVA, seguida por Citibanamex y Santander. No obstante, utilizando Q_x, **BBVA** se mantiene en la posición de banco con **más quejas de servicios digitales** pero en segundo lugar se encuentra HSBC, seguido por Citibanamex. Al considerar los resultados en términos relativos, Inbursa y Scotiabank son los bancos que presentan más quejas relacionadas a problemas en las transferencias al usar canales digitales mientras que los clientes de HSBC y BBVA reportan más problemas derivados de la falla en dichos servicios.

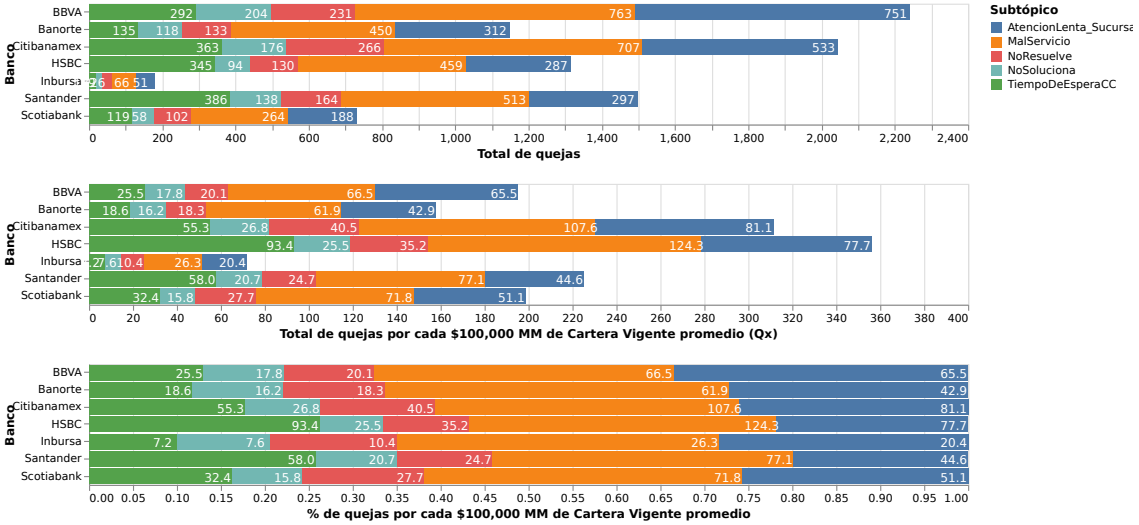


Figura 31: Comparación de resultados de todas las instituciones financieras relacionadas con quejas de **Servicio al Cliente**. Elaboración propia.

Una vez más, en la Figura 31 se muestra que en términos absolutos, los bancos que reciben más quejas relacionadas con el servicio al cliente son: BBVA, Citibanamex y Santander pero al considerar la tasa de quejas por cada 100,000 millones de cartera vigente, **el banco con más quejas respecto al servicio al cliente es HSBC**, seguido por Citibanamex y Santander. Al considerar los datos relativos, los clientes de HSBC y Santander son los más afectados en términos del tiempo que deben esperar para ser atendidos en el *call center*. Inbursa es el banco que más quejas recibe por su falta de solución a los problemas presentados. De la misma forma, los clientes de Inbursa y Banorte son los más afectados por un mal servicio mientras que los clientes de BBVA reportan recibir una atención lenta en las sucursales. Es importante observar que BBVA tiene una tasa de quejas Q_x muy similar a la de Scotiabank, pero con tres veces más de cartera vigente.

La Figura 32 muestra que de forma absoluta, los clientes de BBVA son los que más quejas relacionadas a fraude presentan, sin embargo utilizando la tasa de quejas por cada 100,000 millones de pesos en cartera vigente, los bancos más preocupantes son HSBC y Citibanamex. El banco más afectado por robo de identidad es BBVA, mientras tanto los clientes de Inbursa son los que reciben mayor seguimiento en línea. Por otro lado, los clientes de Banorte son quienes más acusaciones de fraude por parte de la institución realizan.

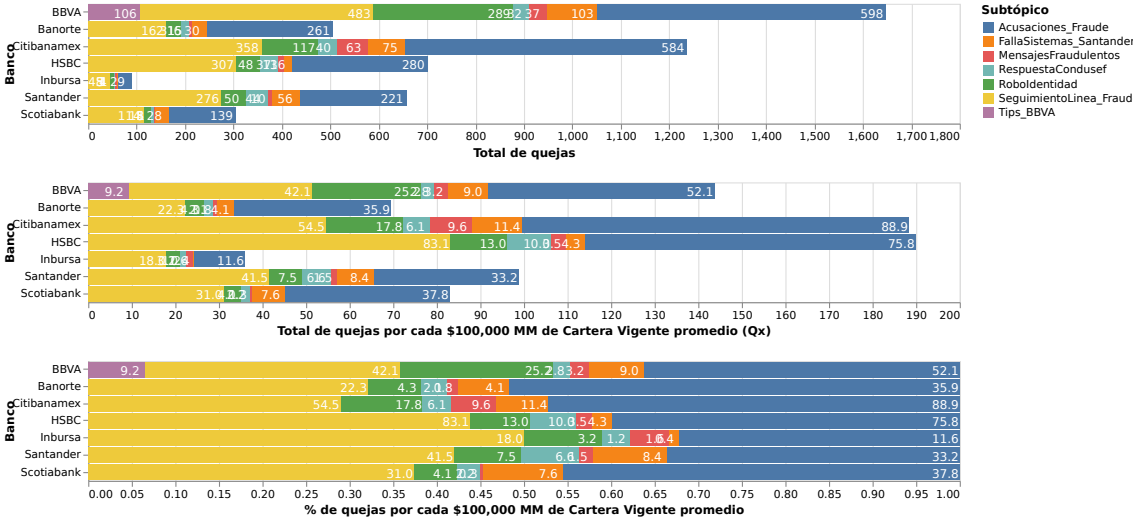


Figura 32: Comparación de resultados de todas las instituciones financieras relacionadas con quejas de **Fraudes**. Elaboración propia.

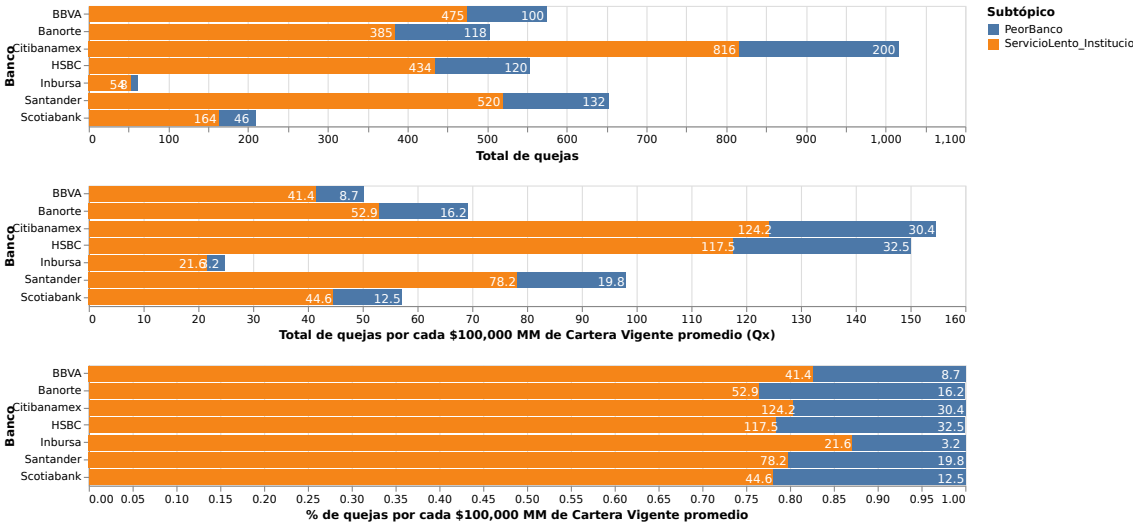


Figura 33: Comparación de resultados de todas las instituciones financieras relacionadas con quejas de **Reputación**. Elaboración propia.

Por último, la Figura 33 muestra que Citibanamex es el banco que más quejas recibe en términos reputacionales. En términos de la tasa Q_x, una vez más son Citibanamex y HSBC los bancos que presentan más problemas en este rubro. Además, la tasa de quejas muestra que HSBC es la institución que recibe más comentarios de ser «el peor banco», seguido por Citibanamex y Santander, mientras que en términos relativos el título es de Banorte. Por su parte, los clientes de Inbursa y BBVA se quejan de recibir un servicio lento. Es importante mencionar que para este tópico, BBVA tiene una tasa de quejas Q_x muy similar a la de Scotiabank, pero con tres veces más de cartera vigente.

Dicho lo anterior, resta contrastar los resultados obtenidos con los reportados por la CONDUSEF para determinas si existe consistencia entre las quejas encontradas en *Twitter* y los datos que tiene la entidad regulatoria.

4.3. Comparación con los resultados presentados por la CONDUSEF

Con base en la información presentada por la CONDUSEF en [CONDUSEF, 2019a], [CONDUSEF, 2019b], [CONDUSEF, 2019c] y [CONDUSEF, 2020] se determinó que las causas de controversias que son constantes en todos los reportes son:

- Consumos no reconocidos.
- Gestión de Cobranza.
- Negativa en el pago de la indemnización.
- Actualización de historial crediticio no realizada.
- Cargos no reconocidos en la cuenta (por parte de la institución otorgante del crédito).
- Cancelación no atendida.
- Crédito no reconocido.
- Disposición de efectivo en ATM no reconocida.
- Cancelación del contrato no atendida.

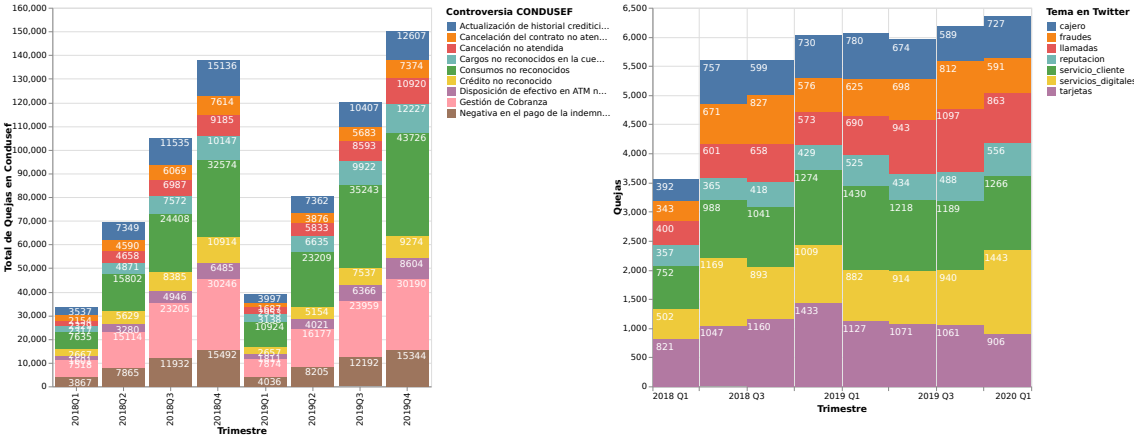


Figura 34: Comparación entre las controversias reportadas por la CONDUSEF y los resultados obtenidos en *Twitter* de 2018 a 2019. Elaboración propia.

La Figura 34 muestra por cada uno de los trimestres las controversias más reportadas ante la CONDUSEF en contraste con los resultados obtenidos en *Twitter*. Se pueden realizar las siguientes observaciones:

- A diferencia de los resultados obtenidos de *Twitter*, en los resultados reportados por la CONDUSEF existe un componente estacional en las quejas que se repite cada 12 meses.
- En ambos casos se observa un incremento de quejas, siendo los últimos trimestres del año los puntos con más concentración de controversias.
- La principal controversia está asociada a **Consumos no reconocidos**, seguida por **Gestión de Cobranza** y **Negativa en el pago de la indemnización.** Por otro lado, en los resultados encontrados en *Twitter* la principal queja es relacionada con **servicios digitales**, seguida por **servicio al cliente** y **tarjetas**.
- Las controversias relacionadas con servicios digitales (cajeros automáticos) son de las menos reportadas para la CONDUSEF, mientras que en *Twitter* se encuentra dentro de las 5 de quejas más frecuentes.

Por otro lado, en la Figura 35 muestra un comparativo en los resultados obtenidos en *Twitter* y los resultados presentados por la CONDUSEF a nivel institución bancaria. Se observa que BBVA y Citibanamex ocupan las primeras dos posiciones tanto en ambos resultados pero en el caso de BBVA no se observa un decremento en la cantidad de controversias reportadas ante la CONDUSEF. Así mismo, se observa que Banorte ocupa el tercer lugar en la CONDUSEF mientras que en *Twitter* se ubica en la quinta posición. Tanto Santander como HSBC muestran un incremento en quejas en ambos reportes. Por último, se puede notar que Inbursa no aparece en el reporte de la CONDUSEF.

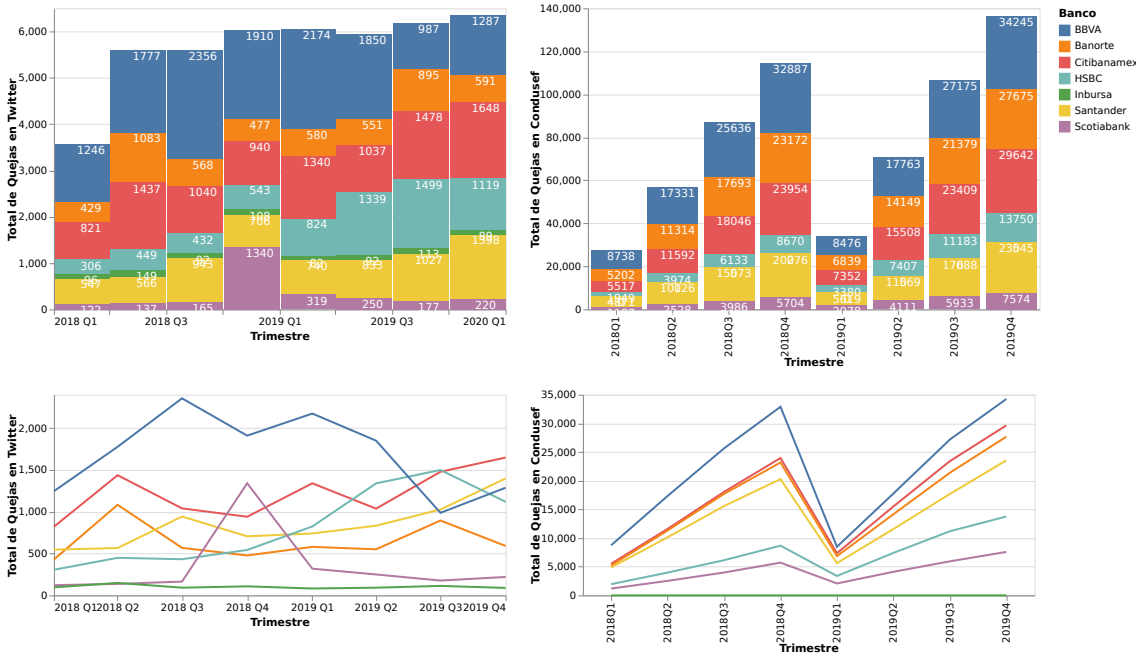


Figura 35: Comparativo de instituciones financieras del G7 con más controversias reportadas por la CONDUSEF de 2018 a 2019. Elaboración propia.

4.4. Resumen

A lo largo de este capítulo se presentaron diferentes análisis descriptivos. Se determinó que existe una tendencia creciente de las quejas de servicios bancarios de 2018 a 2019, además se encontró que se reciben más quejas durante mediados y finales de mes. Por otro lado, se observó que BBVA concentra el mayor número de quejas pero también tiene la mayor cartera vigente en el periodo de tiempo analizado. En contraste, Citibanamex se presenta en la posición número dos de los bancos con mayor número de quejas pero en la posición cuatro (de un total de siete) en cuanto a tamaño. Por último, se notó que solo el 26.04 % de los *tweets* cuentan con datos que permitan geolocalizarlos, de los cuales el 96.15 % de ellos pertenece a México. Además se observó que los estados más capitalizados son los que concentran el 72.15 % de las quejas.

Además se mostraron los resultados obtenidos por tipo de queja para cada una de las instituciones financieras que conforman el G7. Se observó que en general los clientes se quejan principalmente de los problemas que tienen al realizar las transferencias usando canales digitales así como de los cargos no reconocidos en sus tarjetas.

Así mismo, se determinó que los principales puntos de dolor de los clientes están relacionados con tarjetas, servicios digitales y servicio al cliente.

Por último se presentaron los datos reportados por la CONDUSEF en su reporte trimestral *Balance Sobre Las Acciones De Defensa Al Usuario*. Se observó que una de las principales causas de controversias es **Consumos no reconocidos** lo cual tiene relación con la segunda causa detectada en *Twitter* (**Cargos No Reconocidos en TDC/TDD**). Por otro lado, se observó que el comportamiento de controversias dentro de las reportes de la CONDUSEF es similar a los resultados obtenidos en *Twitter* tanto para BBVA como para Citibanamex mientras que Banorte aparece como un banco con menos quejas en *Twitter* pero más controversias reportadas ante la CONDUSEF.

Conclusiones

Conclusiones

Este trabajo presenta una aplicación de recuperación de la información (*IR* por sus siglas en inglés[3]) sobre texto libre para la detección de quejas asociadas a los principales proveedores de servicios bancarios en México. La detección de tópicos (quejas) contenidas en texto libre (*tweets*) consiste en dos fases importantes:

- Representación vectorial del texto libre.
- Encontrar las aglomeraciones naturales en la representación vectorial del texto.

Es así como, mediante el uso de técnicas avanzadas de analítica de texto y modelos de aprendizaje no supervisado es posible extraer la información a partir de datos semi estructurados.

Para realizar el análisis se utilizó la metodología *CRISP-DM*, la cual contempla fases específicas y exhaustivas para la extracción, limpieza y modelado de datos. Además de ser una metodología orientada a resolver problemas como una aplicación de la minería de datos.

Los resultados obtenidos muestran que existe una gran inconformidad por parte de los clientes al usar los servicios de trasferencias electrónicas, así como por cargos no reconocidos en sus tarjetas de crédito o débito. Lo anterior resalta problemas en los sistemas tecnológicos y de seguridad de los bancos.

Por otro lado, las quejas suelen presentarse de forma más frecuente los días de pago (inicio y mediados de mes), además los días de la semana con mayor actividad en términos de quejas son Viernes, Lunes y Martes en un horario de crítico de 12:00 a 19:00 horas. Con la información anterior se puede concluir que los Viernes de quincena de 12:00 a 19:00 son los momentos críticos para los bancos.

Las instituciones financieras del G7 que reciben de forma recurrente una mayor cantidad de quejas son: BBVA, Citibanamex y Santander.

[3] *IR: Information Retrival*

Además, de los *tweets* con información geográfica, la cantidad de quejas se concentra principalmente en estados que cuentan con las principales ciudades del país, como CDMX, Jalisco y Nuevo León. En el caso de CDMX, las quejas se agrupan en las alcaldías con un nivel socioeconómico mayor. En consecuencia, se observa de manera recurrente que un mayor número de quejas lo cual se traduce en un comportamiento más exigente en términos de servicio y atención en zonas con mayores ingresos.

A nivel institución se mostró lo siguiente:

- **BBVA.** Es el banco que concentra un número mayor de quejas. No obstante, también es la institución financiera más grande términos de cartera activa. Por otro lado, su principal dolor son los **servicios digitales** siendo los problemas al realizar o recibir trasferencias su principal queja, lo anterior indica problemas en la calidad de su banca móvil y banca en línea. Mientras tanto, recibe un número bajo de quejas relacionadas a su reputación, lo cual indica que tiene buena relación con sus clientes. Es importante mencionar que BBVA ha logrado disminuir la tendencia de quejas en *Twitter* lo cual habla de una estrategia y acciones para mejorar la experiencia de sus clientes.
- **Citibanamex.** Es el segundo banco que recibe más quejas en *Twitter* y ocupa la cuarta posición en términos de cartera vigente durante el último trimestre del 2019 (con menos de la mitad de respecto a BBVA). Su principal dolor es el **servicio al cliente** siendo su principal queja el lento servicio que reciben los usuarios con lo que se infiere una mala experiencia del usuario. Por otro lado, recibe un número pequeño de quejas relacionadas a las llamadas que realiza para colocar sus productos, con lo cual se infiere que es un banco que no tiene campañas agresivas de colocación via telefónica, lo cual agradecen los clientes. Desafortunadamente se observa un número creciente en la cantidad de quejas asociadas a Citibanamex, lo cual indica que si han tomado acciones, éstas no han tenido el efecto esperado.
- **Santander.** Es el tercer banco con mayor número de quejas así como el tercero en términos de cartera vigente. El dolor principal de sus clientes es el **servicio al**

cliente siendo la principal queja el lento servicio por parte de la institución con lo que se infiere una mala experiencia del usuario. Por otro lado, recibe un bajo número de quejas relacionadas a fraudes y se puede inferir que es un banco cuyos clientes no suelen ser victimas de fraudes. Desafortunadamente, se observa un comportamiento creciente en el número de quejas en todas sus aristas, lo anterior indica que la institución está teniendo grandes áreas de oportunidad conforme avanza el tiempo.

- **HSBC.** Es el cuarto banco con mayor cantidad de quejas y el penúltimo en términos de cartera vigente. Su **principal queja son las llamadas** constantes que realiza esta institución para ofrecer sus productos y servicios, lo anterior muestra cansancio, fastidio y hartazgo por parte de los clientes de HSBC. Por otro lado, el número de quejas que recibe en términos de sus cajeros es baja y se puede inferir que tienen una infraestructura robusta para cajeros automáticos. Durante todo 2018 y 2019 se observa un comportamiento acelerado en la cantidad de quejas recibidas, detonados principalmente por las llamadas y el servicio al cliente, sin embargo en ambos rubros pudo mejorar durante el último trimestre de 2019.
- **Banorte.** Es el tercer banco con menor número de quejas y durante el último trimestre del 2019 fue el segundo banco con la cartera vigente más grande, lo cual es un buen indicador. Su principal dolor es el **servicio al cliente** cuya queja mas representativa es el mal servicio por parte de la institución con lo que se infiere una mala experiencia del usuario. Mientras tanto, recibe un bajo número de quejas relacionadas a las llamadas que realiza para colocar sus productos con lo cual se infiere que es un banco que no tiene campañas agresivas de colocación via telefónica, lo cual agradecen los clientes. En el caso de Banorte, no se observa una tendencia creciente o descendiente del número de quejas lo cual indica que tiene controladas sus áreas de oportunidad, sin embargo no ha tomado acciones para corregirlas.
- **Scotiabank.** Es el penúltimo banco con mayor cantidad de quejas y ocupa la quinta posición en términos de cartera vigente. Su principal dolor es el **servicio al cliente** siendo el mal servicio la queja principal, con lo que se infiere una

mala experiencia del usuario. Por otro lado, recibe un bajo número de quejas relacionadas a sus cajeros y se puede inferir que tienen una infraestructura robusta para cajeros automáticos. Es importante mencionar que este banco tuvo un comportamiento atípico en octubre 2018, lo cual aumentó la cantidad de quejas recibidas en más de un 300 %, de no haber sido por dicho comportamiento se muestra como un banco que recibe pocas quejas por parte de los usuarios. En cuanto a la tendencia de las quejas recibidas, no se observa una tendencia creciente o descendiente del número de quejas lo cual indica que el banco tiene controladas sus áreas de oportunidad, sin embargo no ha tomado acciones para corregirlas.

- **Inbursa.** Es el banco con menor número de quejas así como con la menor cartera vigente de los bancos del G7. Su dolor más fuerte es en término de **tarjetas de crédito y débito** cuya principal queja son los cargos no reconocidos en tarjetas de crédito y debido, este es un punto crítico pues muestra una vulnerabilidad en la seguridad de sus productos lo cual genera desconfianza por parte de los clientes. Por otro lado, no recibe un número importante de quejas respecto a los cajeros y se puede inferir que tienen una infraestructura robusta para cajeros automáticos. A lo largo del periodo analizado, no se observa una tendencia creciente o descendiente del número de quejas lo cual indica que el banco tiene controladas sus áreas de oportunidad, sin embargo no ha tomado acciones para corregirlas.

En comparación con la información presentada por la CONDUSEF se observa que existen una relación en términos de la posición en la que se presentan las instituciones bancarias por número de controversias, así como en algunas de las controversias reportadas, sin embargo la diferencia más relevante se observa en los servicios digitales ya que este rubro no forma parte de las principales causas de controversias recibidas por la CONDUSEF. Además, el servicio al cliente es otro punto fundamental de quejas en *Twitter* que no contempla la comisión. Por lo tanto, se puede concluir que dentro de *Twitter* existen quejas que no son reportadas a la entidad regulatoria o bien, no es relevante para ésta. En consecuencia, se puede tomar los resultados aquí presentados como información complementaria a la generada por la CONDUSEF. Adicional-

mente, el producto de datos generado permite entender la frecuencia de las quejas a través del tiempo y ubicarlas geográficamente. Por último, es importante mencionar que este tipo de análisis pueden ser ejecutados bajo demanda en cualquier momento del tiempo, lo cual genera la posibilidad de tener un seguimiento muy puntual de las quejas de todas las instituciones financieras del país.

La información presentada en este trabajo es relevante para las instituciones financieras de México pues les permite conocer y monitorear los puntos de dolor de sus clientes para tomar las acciones necesarias con el fin de mejorar la experiencia del usuario. Además, estos resultados pueden ser usados por la población general, pues complementa la información generada por la CONDUSEF con lo cual se pueden tomar decisiones mejores informadas para escoger un proveedor de servicios financieros en términos de las necesidades y expectativas de cada uno.

Bibliografía

[Al-Rfou et al., 2013] Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.

[Aranganayagi and Thangavel, 2007] Aranganayagi, S. and Thangavel, K. (2007). Clustering categorical data using silhouette coefficient as a relocating measure. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, volume 2, pages 13–17. IEEE.

[Ayala, 2014] Ayala, T. (2014). Redes sociales, poder y participación ciudadana. *Revista Austral de Ciencias Sociales*, (26):23–48.

[Baeza-Yates et al., 1999] Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

[BANXICO, 2018] BANXICO (2018). Glosario. Consultado: 2020-05-01.

[BANXICO, 2019] BANXICO (2019). Misión y visión. Consultado: 2020-05-01.

[Bejar, 2020] Bejar, J. O. (2020). A family of classifiers based on feature space transformations and model selection.

[Brown, 2015] Brown, M. S. (2015). What it needs to know about the data mining process. *Published by Forbes*, 29.

[Bullinaria and Levy, 2007] Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

[Church and Hanks, 1990] Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

[CNBV, 2018] CNBV (2018). ¿qué hacemos? Consultado: 2020-05-01.

[CNBV, 2020a] CNBV (2020a). Glosario de términos portafolio de información. Consultado: 2020-05-01.

[CNBV, 2020b] CNBV (2020b). Portafolio de información. Consultado: 2020-05-01.

[Collobert et al., 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

[CONDUSEF, 2018] CONDUSEF (2018). ¿qué hacemos? Consultado: 2020-05-01.

[CONDUSEF, 2019a] CONDUSEF (2019a). Balance sobre las acciones de defensa al usuario [primer trimestre 2019]. Consultado: 2020-05-01.

[CONDUSEF, 2019b] CONDUSEF (2019b). Balance sobre las acciones de defensa al usuario [segundo trimestre 2019]. Consultado: 2020-05-01.

[CONDUSEF, 2019c] CONDUSEF (2019c). Balance sobre las acciones de defensa al usuario [tercer trimestre 2019]. Consultado: 2020-05-01.

[CONDUSEF, 2020] CONDUSEF (2020). Balance sobre las acciones de defensa al usuario [cuarto trimestre 2019]. Consultado: 2020-05-01.

[CONDUSEF, 20Q1] CONDUSEF (2020Q1). Balance sobre las acciones de defensa al usuario [primer trimestre 2020]. Consultado: 2020-05-01.

[Costumero et al., 2014] Costumero, R., García-Pedrero, Á., Gonzalo-Martín, C., Menasalvas, E., and Millan, S. (2014). Text analysis and information extraction from spanish written documents. In *International Conference on Brain Informatics and Health*, pages 188–197. Springer.

[Culnan et al., 2010] Culnan, M. J., McHugh, P. J., and Zubillaga, J. I. (2010). How large us companies can use twitter and other social media to gain business value. *MIS Quarterly Executive*, 9(4).

[Danisman and Alpkocak, 2008] Danisman, T. and Alpkocak, A. (2008). Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, page 53.

[Dave et al., 2003] Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528.

[Expansión, 2019] Expansión (2019). Estos son los bancos con más quejas, según condusef. Consultado: 2020-05-01.

[Facebook, 2020] Facebook (2020). Advanced readers: playing with the parameters. Consultado: 2020-05-01.

[Guercini et al., 2014] Guercini, S., Misopoulos, F., Mitic, M., Kapoulas, A., and Karapiperis, C. (2014). Uncovering customer service experiences with twitter: the case of airline industry. *Management Decision*.

[Huang, 2008] Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56.

[Hussain and Prieto, 2016] Hussain, K. and Prieto, E. (2016). Big data in the finance and insurance sectors. In *New Horizons for a Data-Driven Economy*, pages 209–223. Springer, Cham.

[Jensen, 2012] Jensen, K. (2012). Crisp-dm process diagram. Consultado: 2020-05-01.

[Jones and Fox, 2009] Jones, S. and Fox, S. (2009). Generations online in 2009. data memo. pew internet and american life project, washington, dc.

[Joulin et al., 2016] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

[Juárez, 2018] Juárez, E. (2018). Bancos del g7, de importancia sistémica. Consultado: 2020-05-01.

[Kenter et al., 2016] Kenter, T., Borisov, A., and De Rijke, M. (2016). Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*.

[Larsen and Aone, 1999] Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–22.

[Levy and Goldberg, 2014] Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.

[Liu et al., 2015] Liu, Y., Liu, Z., Chua, T.-S., and Sun, M. (2015). Topical word embeddings. In *Twenty-ninth AAAI conference on artificial intelligence*. Citeseer.

[Lula and Wójcik, 2011] Lula, P. and Wójcik, K. (2011). Sentiment analysis of consumer opinions written in polish,„. *Economics and Management*, 16(1):1286–1291.

[Madhulatha, 2012] Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*.

[Manning et al., 2008] Manning, C. D., Schütze, H., and Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge university press.

[Mikolov et al., 2013] Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

[Miranda and Guzmán, 2017] Miranda, C. H. and Guzmán, J. (2017). A review of sentiment analysis in spanish. *Tecciencia*, 12(22):35–48.

[Orozco, 2011] Orozco, M. E. V. (2011). La bancarización en méxico.

[Pak and Paroubek, 2010] Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

[Restrepo and Pacheco, 2019] Restrepo, H. F. M. and Pacheco, G. A. F. (2019). Smart center for asset management: Transformation to intelligent maintenance through digitalization. In *2019 FISE-IEEE/CIGRE Conference-Living the energy Transition (FISE/CIGRE)*, pages 1–6. IEEE.

[Rokach and Maimon, 2005] Rokach, L. and Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer.

[Rowley, 2007] Rowley, J. (2007). The wisdom hierarchy: representations of the dikw hierarchy. *Journal of information science*, 33(2):163–180.

[Salton, 1989] Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*, 169.

[Santander, 2018] Santander (2018). Estado financiero: 2do triemestre 2018. Consultado: 2020-05-01.

[Scotiabank, 2020] Scotiabank (2020). Estado financiero: 1er triemestre 2020. Consultado: 2020-05-01.

[Shearer, 2000] Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

[Shiha and Ayvaz, 2017] Shiha, M. and Ayvaz, S. (2017). The effects of emoji in sentiment analysis. *Int. J. Comput. Electr. Eng.(IJCEE.)*, 9(1):360–369.

[Socher et al., 2011] Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 151–161.

[Tellez et al., 2017] Tellez, E. S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O. S., and Villaseñor, E. A. (2017). A case study of spanish text transformations for twitter sentiment analysis. *Expert Systems with Applications*, 81:457–471.

[Thelwall et al., 2011] Thelwall, M., Buckley, K., and Paltoglou, G. (2011). Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

[Turian et al., 2010] Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394.

[Uszkoreit and Brants, 2008] Uszkoreit, J. and Brants, T. (2008). Distributed word clustering for large scale class-based language modeling in machine translation. In *Proceedings of ACL-08: HLT*, pages 755–762.

[Vidya et al., 2015] Vidya, N. A., Fanany, M. I., and Budi, I. (2015). Twitter sentiment to analyze net brand reputation of mobile phone providers. *Procedia Computer Science*, 72:519–526.

[Ward Jr, 1963] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

[Wilbur and Sirotkin, 1992] Wilbur, W. J. and Sirotkin, K. (1992). The automatic identification of stop words. *Journal of information science*, 18(1):45–55.

[Yan et al., 2009] Yan, H., Ding, S., and Suel, T. (2009). Inverted index compression and query processing with optimized document ordering. In *Proceedings of the 18th international conference on World wide web*, pages 401–410.

[Zapatero et al., 2013] Zapatero, M. C., Brändle, G., and San-Román, J. R. (2013). Comunicación interpersonal en la web 2.0. las relaciones de los jóvenes con desconocidos. *Revista Latina de Comunicación Social*, (68):436–456.

ANEXOS

Dendogramas por Tópicos

Este anexo está dedicado a presentar los dendogramas que ayudan a identificar los subtópicos encontrados en cada tema.

Cajero

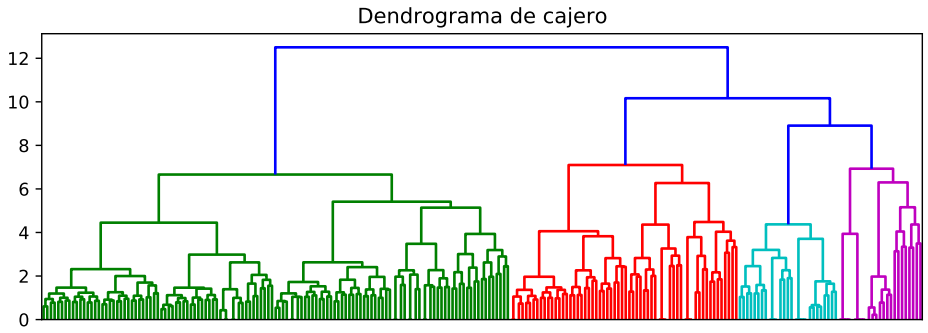


Figura 36: Dendograma obtenido al aplicar *HCA divisivo* al tópico **Cajero**. Elaboración propia.

En la Figura 36 se puede apreciar cuatro subclústers para el tópico **Cajero.** Al analizar los comentarios de cada subclúster se determinó que las quejas están relacionadas a:

- Problemas con depósitos o retiros en cajeros automáticos.
- El cajero automático no funciona.
- El cajero (personal de ventanilla) me robó.
- El cajero automático se quedó con el dinero.

Se puede notar que un tema sensible en términos de Cajeros es el dinero de los cuentahabientes.

Llamadas

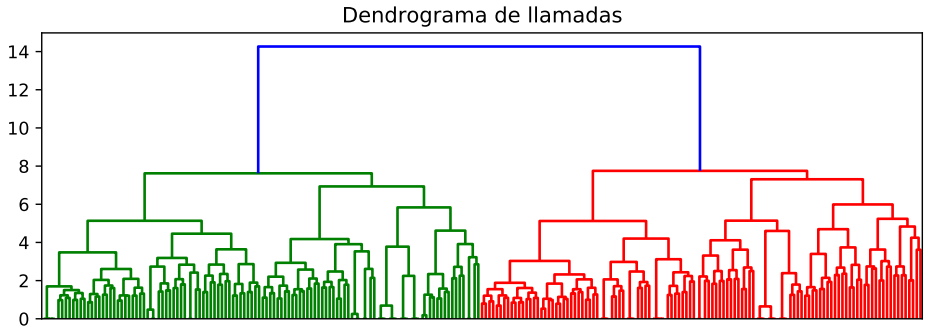


Figura 37: Dendograma obtenido al aplicar *HCA divisivo* al tópico **Llamadas**. Elaboración propia.

En la Figura 37 se observan dos subclústers para el tópico **Llamadas.** Al analizar los comentarios de cada subclúster se determinó que las quejas están relacionadas a:

- Recibir llamadas para ofrecer productos y servicios.
- Recibir llamadas fraudulentas.

Con lo anterior se puede observar que existe un riesgo latente hacia los usuarios de recibir llamadas fraudulentas.

Servicio al Cliente

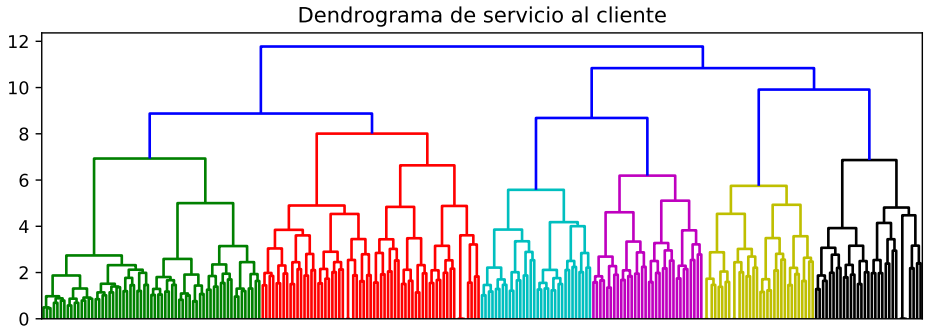


Figura 38: Dendograma obtenido al aplicar *HCA divisivo* al tópico **Servicio al Cliente**. Elaboración propia.

En la Figura 38 se muestran seis subclústers para el tópico **Servicio al cliente.** Al analizar los comentarios de cada subclúster se determinó que las quejas están relacionadas a:

- Mal servicio por parte de la institución.
- El banco no soluciona los problemas.
- La atención que se recibe es lenta en las sucursales bancarias.
- Es demasiado el tiempo de espera en Call Center.
- El banco no resuelve los problemas.

En este punto es importante mencionar que los subclústers *el banco no soluciona los problemas* y *el banco no resuelve los problemas* son matemáticamente diferentes pues las expresiones que se ocupan para uno y otro son diferentes.

Servicios Digitales

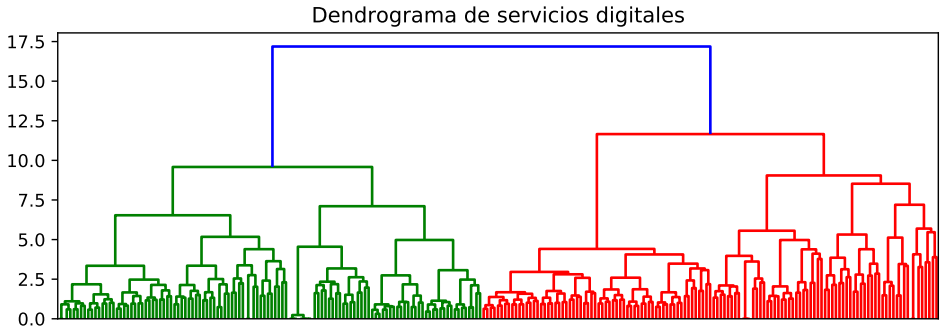


Figura 39: Dendograma obtenido al aplicar *HCA divisivo* al tópico **Servicios Digitales**. Elaboración propia.

En la Figura 39 se muestran dos subclústers para el tópico **Servicio digitales.** Al analizar los comentarios de cada subclúster se determinó que las quejas están relacionadas a:

- Problemas al realizar transferencias interbancarias a través de servicios digitales.
- Indisponibilidad de los servicios digitales.

Una vez más se puede notar que los problemas relacionados al dinero de los cuentahabientes es un tema crucial.

Fraudes

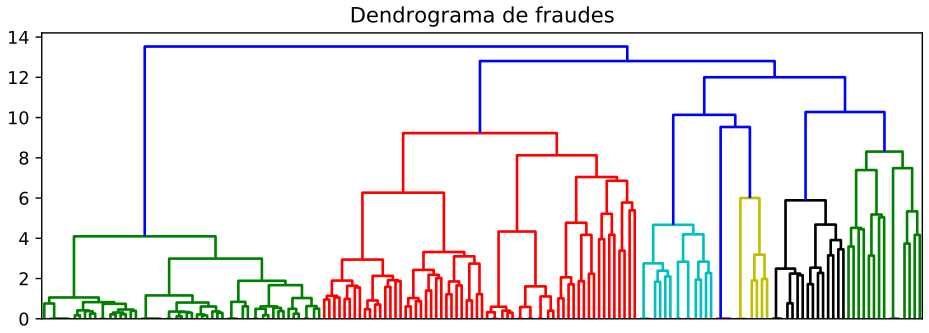


Figura 40: Dendograma obtenido al aplicar *HCA divisivo* al tópico **Fraudes**. Elaboración propia.

En la Figura 40 se muestran siete subclústers para el tópico **Fraudes.** Al analizar los comentarios de cada subclúster se determinó que los subclúster están relacionadas a:

- Seguimiento en línea (en *Twitter*) de algún reporte.
- Acusaciones de fraude hacia algún miembro del banco o terceros.
- Reporte de recibir mensajes fraudulentos.
- Respuestas por parte de la CONDUSEF en los *tweets* que fueron etiquetados.
- Reportes de robo de identidad.
- Reporte de fallas en el sistema de Santander.
- Tips de BBVA para prevenir robo a casa habitación.

Este es el tópico que generó más subclústers, lo cual muestra la polémica alrededor de este tema.

Reputación

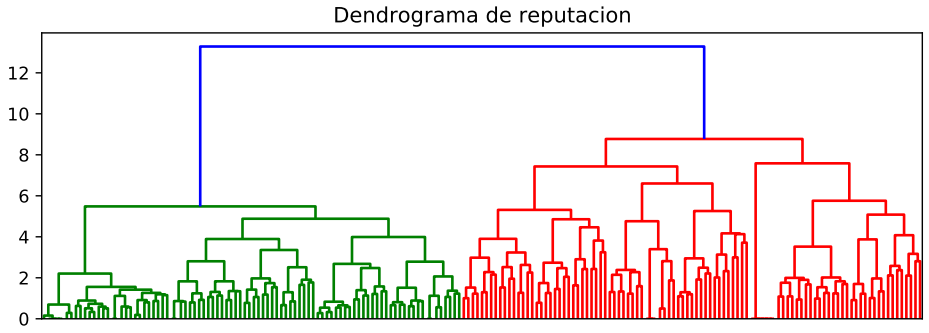


Figura 41: Dendograma obtenido al aplicar *HCA divisivo* al tópico **Reputación**. Elaboración propia.

En la Figura 41 se muestran seis subclústers para el tópico **Reputación.** Al analizar los comentarios de cada subclúster se determinó que las quejas están relacionadas a:

- Clientes otorgando el título de «peor banco» a alguna institución bancaria.
- Quejas relacionadas con el servicio lento recibido de forma general por parte del banco.

Es interesante observar como existe un subclúster completo dedicado a atacar de forma directa a las instituciones, pues en varios *tweets* de dicho clúster se ocupa un lenguaje soez.

Tarjetas

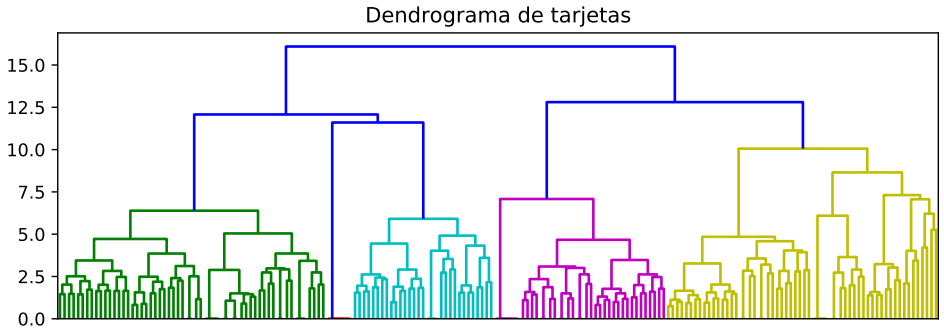


Figura 42: Dendograma obtenido al aplicar *HCA divisivo* al tópico **Tarjetas**. Elaboración propia.

En la Figura 42 se muestran cinco subclústers para el tópico **Tarjetas.** Al analizar los comentarios de cada subclúster se determinó que las quejas están relacionadas a:

- Cargos no reconocidos en tarjetas de crédito y tarjetas de débito.
- Incumplimiento de las promociones de tarjetas de crédito o de débito.
- Tiempo de atención para realizar aclaraciones relacionadas con tarjeta de crédito o de débito.
- Problemas al realizar el pago de la tarjeta de crédito.
- Comisiones en la tarjeta de débito.

Cabe resaltar que el último subclúster se compone de una serie de *retweets* donde el usuario se queja de una comisión que tuvo que pagar a BBVA por recibir un depósito en euros.

Resultados por Institución Financiera

Este apéndice esta dedicado a presentar los resultados temporales y espaciales por institución financiera perteneciente al G7.

Resultados de BBVA

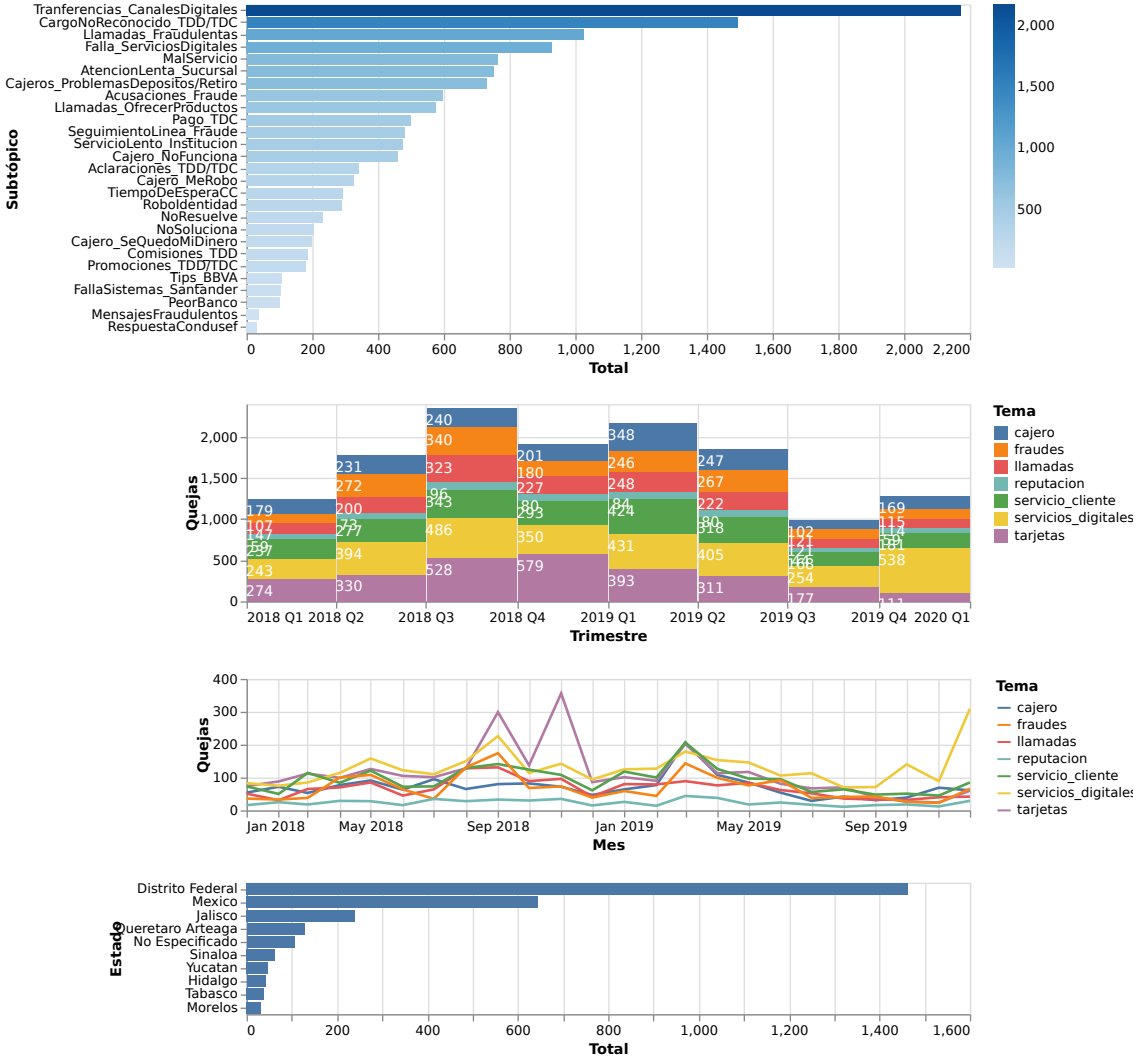


Figura 43: Análisis temporal y espacial por tipo de queja encontrada en *Twitter* para BBVA. Elaboración propia.

De la Figura 43 se puede observar que la queja principal para BBVA se detona al realizar o recibir transferencias interbancarias mediante el uso de canales digitales, seguido por problemas derivados de cargos no reconocidos en las tarjetas. En general, las problemáticas relacionadas con **Canales Digitales** han aumentado su concurrencia durante el último trimestre de 2019.

Resultados de Citibanamex

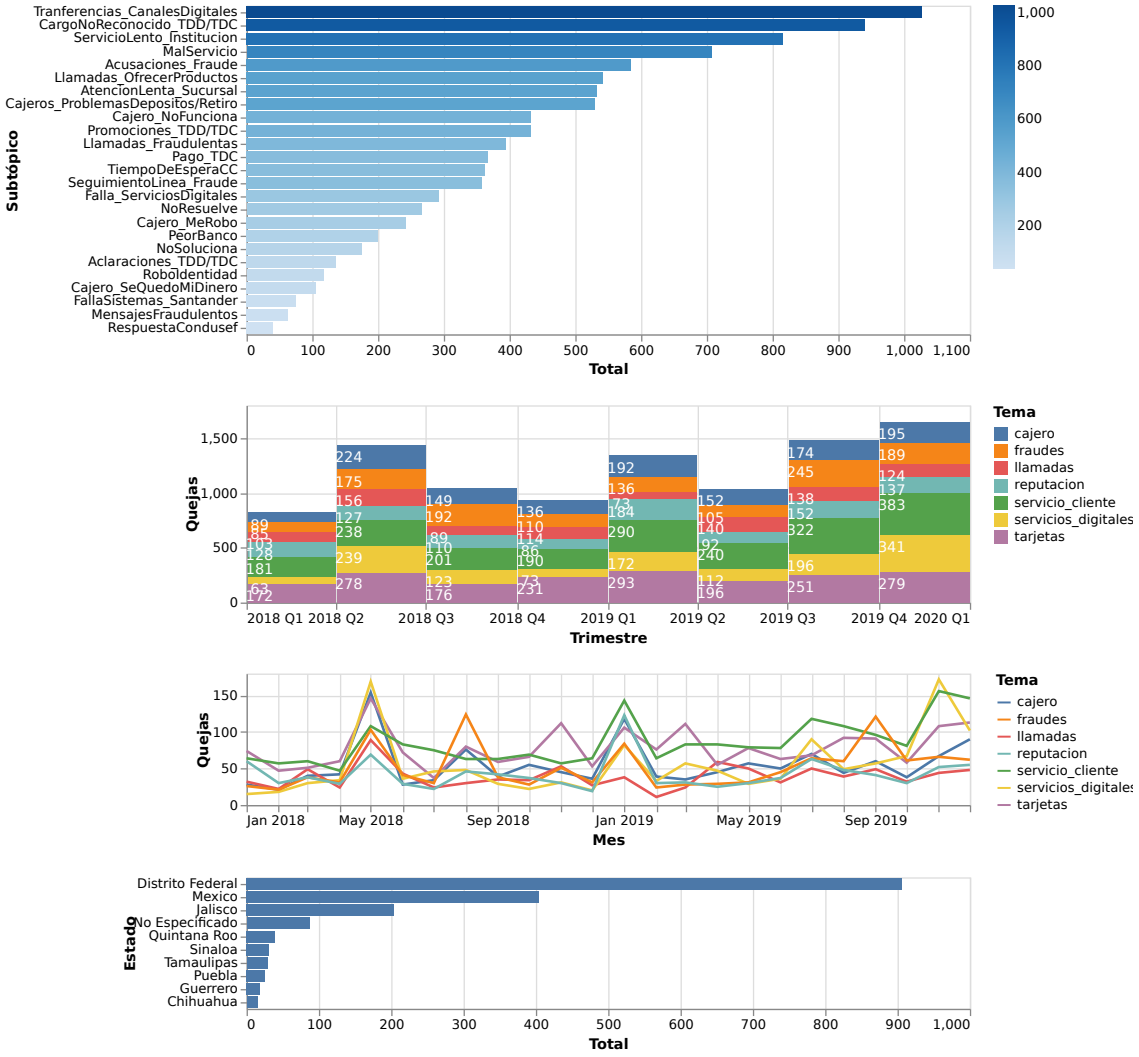


Figura 44: Análisis temporal y espacial por tipo de queja encontrada en *Twitter* para Citibanamex. Elaboración propia.

En la Figura 44 es importante observar que para Citibanamex no existe una diferencia importante entre las tres principales quejas de los clientes: **enviar o recibir transferencias en canales digitales, cargos no reconocidos en tarjetas** y **atención lenta por parte de la institución**. Además, a lo largo del periodo estudiado se presenta un **crecimiento en el número de quejas** dirigidas a este banco.

Resultados de Scotiabank

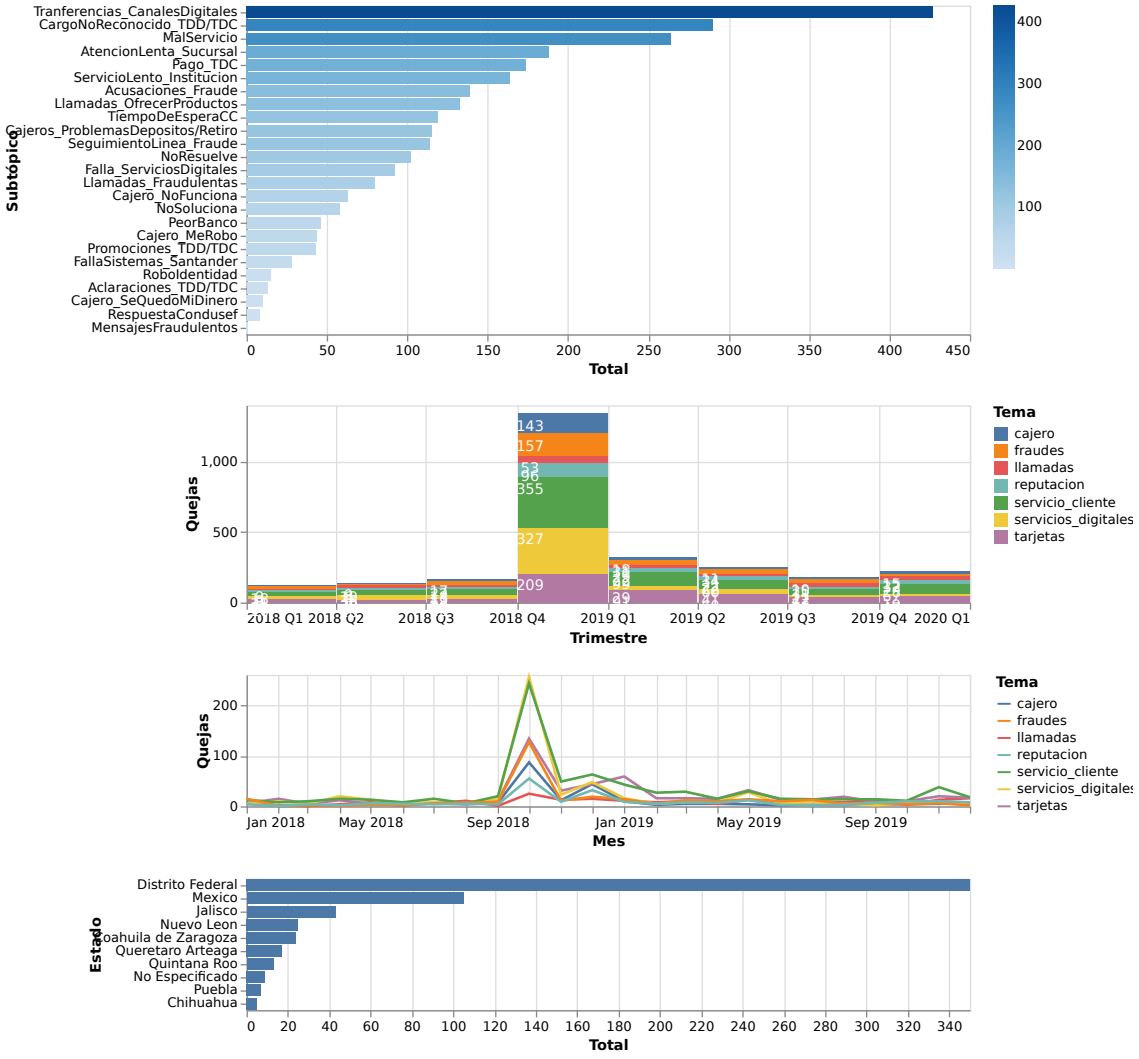


Figura 45: Análisis temporal y espacial por tipo de queja encontrada en *Twitter* para Scotiabank. Elaboración propia.

En la Figura 45 es posible observar que la queja principal de los clientes de Scotiabank está relacionada con problemas al realizar o recibir transferencias mediante el uso de canales digitales. Por otro lado, se nota un comportamiento atípico en las quejas pues se nota un crecimiento muy abrupto en octubre 2018, mismo que desaparece en noviembre del mismo año.

Resultados de Banorte

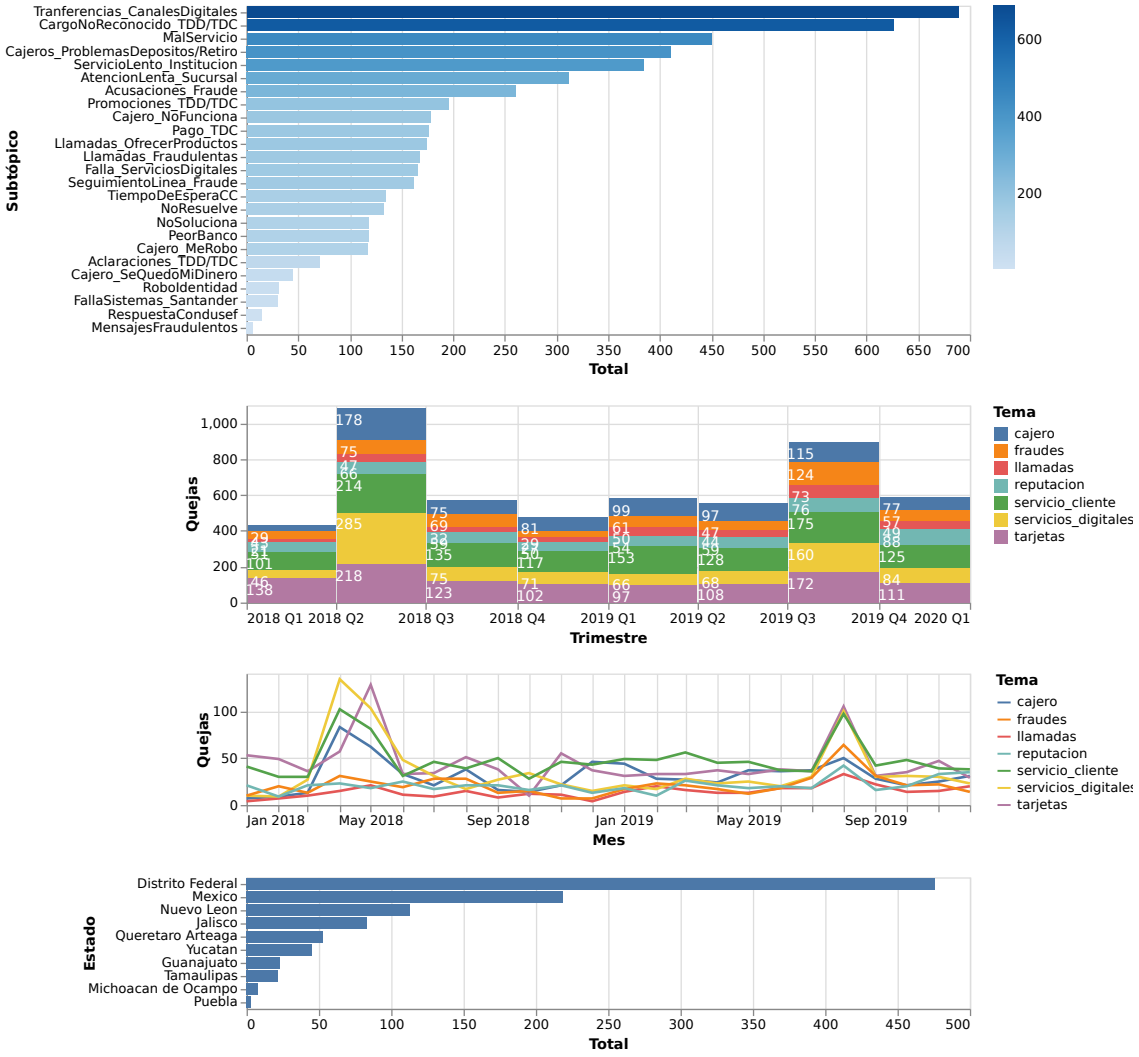


Figura 46: Análisis temporal y espacial por tipo de queja encontrada en *Twitter* para Banorte. Elaboración propia.

Con ayuda de la Figura 46 se observa que las dos principales quejas de Banorte son relacionadas a: problemas al enviar y recibir transferencias usando los canales digitales y cargos no reconocidos en tarjetas. En general, las áreas de oportunidad están enfocadas en tarjetas, servicios digitales y servicio al cliente.

Resultados de Inbursa

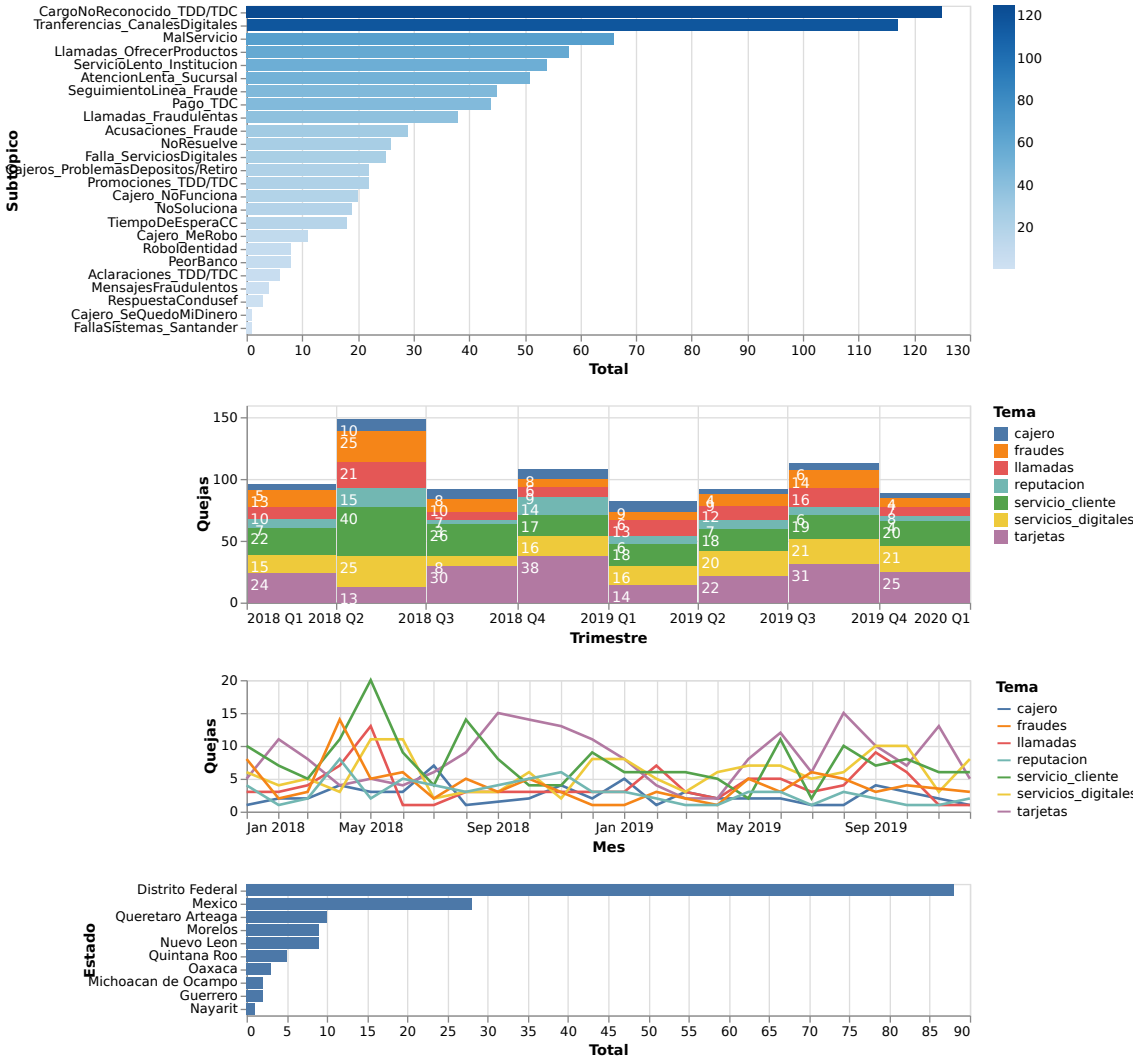


Figura 47: Análisis temporal y espacial por tipo de queja encontrada en *Twitter* para Inbursa. Elaboración propia.

La Figura 47 muestra los resultados obtenidos para Inbursa, de donde es importante observar que la principal queja de este banco es relacionada a **cargo no reconocido** seguida por problemas en **transferencias al usar servicios digitales**.

Resultados de HSBC

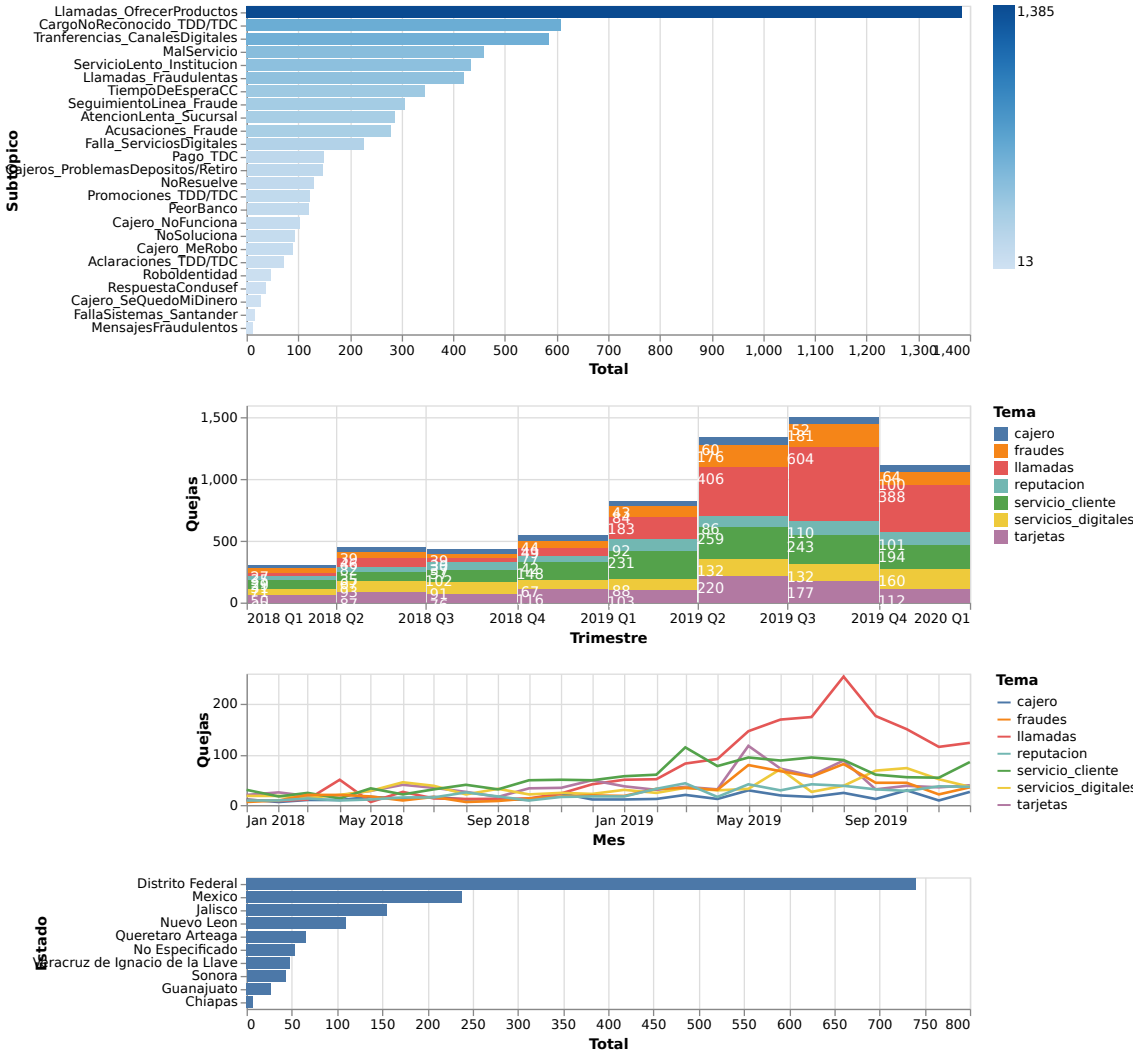


Figura 48: Análisis temporal y espacial por tipo de queja encontrada en *Twitter* para HSBC. Elaboración propia.

En la Figura 48 se observa un comportamiento muy particular, pues la principal queja de los clientes de HSBC son las constantes llamadas de este banco para ofrecer sus productos y servicios. Se puede observar un crecimiento acelerado en las quejas de 2018 a 2019 siendo particularmente alarmante aquellas relacionadas con llamadas y servicio al cliente. Las quejas recibidas son principalmente de **CDMX**.

Resultados de Santander

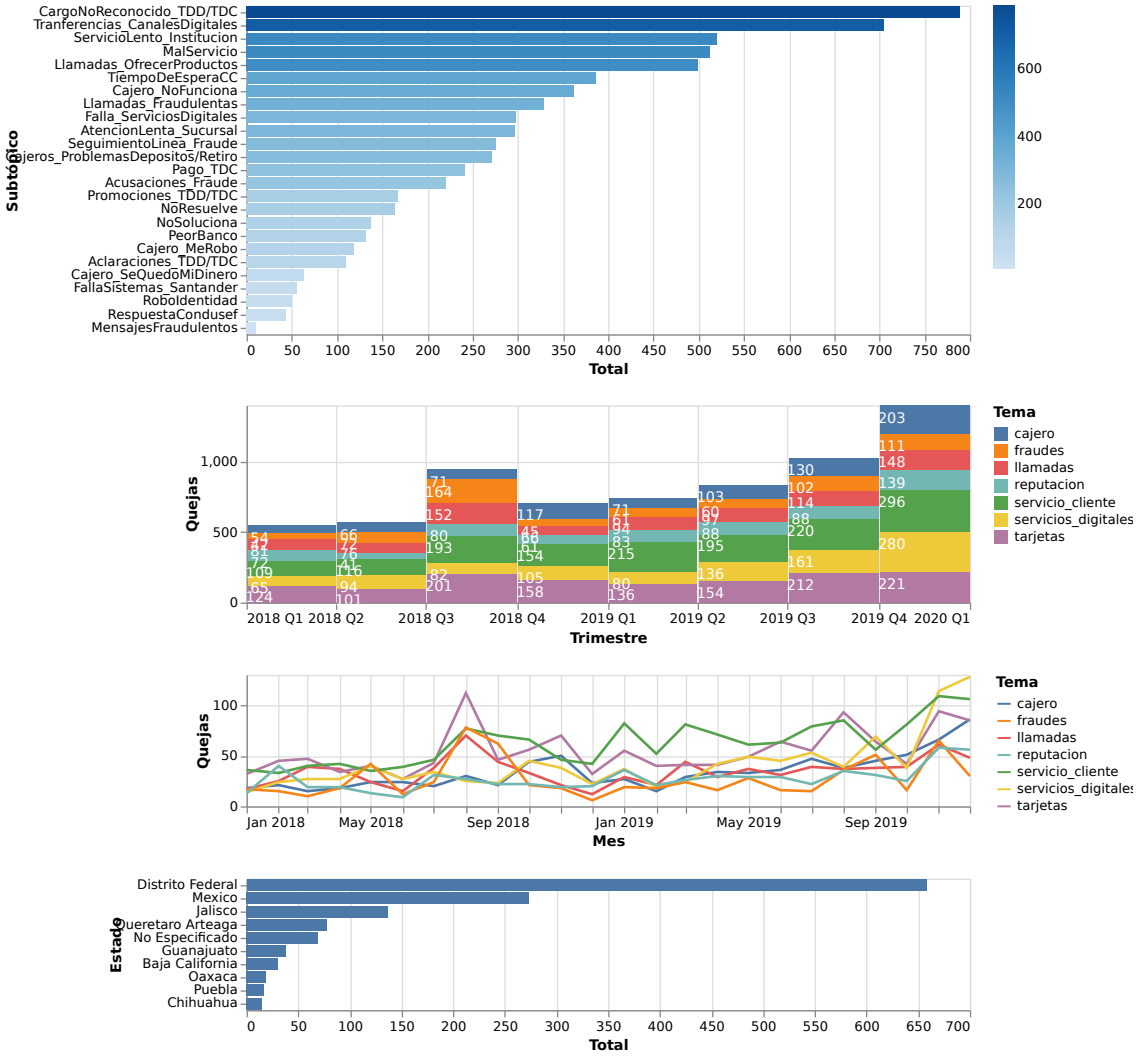


Figura 49: Análisis temporal y espacial por tipo de queja encontrada en *Twitter* para Santander. Elaboración propia.

En la Figura 49 se puede apreciar que las principales quejas de los clientes de Santander son relacionada **cargos no reconocidos en tarjetas** y **fallas en servicios digitales para enviar o recibir transferencias interbancarias**. Se puede observar un comportamiento creciente en el número de quejas recibidas durante 2018 y 2019.