UNIVERSIDAD RICARDO PALMA
ESCUELA DE POSGRADO

MAESTRÍA EN CIENCIA DE LOS DATOS

Tesis para optar el Grado Académico de Maestro en Ciencias de los Datos

"Modelo Predictivo del Índice NPS Basado en Información Textual de Percepción del Servicio al Cliente"

Autor: Bach. Tristán Gomez, Ludgardo Eder

Asesor: Mg. Roque Paredes, Ofelia

LIMA - PERÚ

2019

Los miembros del jurado examinador para la evaluación de la sustentación de la tesis, está integrado por:

1.	Mg. Enver Gerald Tarazona Vargas	Presidente
2.	Mg. Jesús Salinas Flores	Miembro
3.	Mg. Luiggi Dávila Rivera	Miembro
4.	PhD Edwyn Javier Aldana Bobadilla	Asesor Externo
5.	Mg. Ofelia Roque Paredes	Asesor Interno
6.	Dra. Reina Zúñiga de Acleto	Directora EPG

Dedicatoria

El presente trabajo está dedicado a mis padres Ludgardo y Gladys quienes con su amor, paciencia y esfuerzo me han permitido llegar a cumplir hoy un sueño más.

A mis hermanos por su cariño y apoyo incondicional, durante todo este proceso, por estar conmigo en todo momento gracias. A toda mi familia porque con sus consejos y palabras de aliento hicieron de mí una mejor persona y de una u otra forma me acompañan en todos mis sueños y metas.

Finalmente quiero dedicar esta tesis al Dr.Erwin Kraneu y todos mis amigos, por apoyarme cuando más las necesito, por extender su mano en momentos difíciles.

Agradecimiento

Le agradezco a la universidad, gracias por haberme permitido fórmame en ella, gracias a todas las personas que fueron participes de este proceso, ya sea de manera directa o indirecta, gracias a todos ustedes en especial al fundador de la maestría de ciencia de los datos al Dr. Erwin Kraneu que nos abrió el camino, gracias a la coordinadora de la maestría a la Mg. Ofelia Roque quien con su constante apoyo fue la gran artífice este aporte, que el día de hoy se vería reflejado en la culminación de mi paso por la universidad. Gracias a mi asesor que a la distancia se tomó el tiempo de poder guiarme en la elaboración de este trabajo. Gracias a mis padres y a mi pareja, que fueron mis mayores promotores durante este proceso.

Este es un momento muy especial que espero, perdure en el tiempo, no solo en la mente de las personas a quienes agradecí, sino también a quienes invirtieron su tiempo para echarle una mirada a mi borrador de tesis, a ellos asimismo les agradezco.

Índice de Contenido

Lista de tablas y figuras

Lista de tablas

Lista de Figuras

RESUMEN

En la actualidad la administradora del fondo de pensiones (AFP) del mercado peruano viene realizando encuestas telefónicas para medir el nivel de servicio que ofrece a sus clientes, se hace uso del Net Promoter Score una metodología que clasifica a los clientes en función a la calificación que brinda a la pregunta: "¿Qué tan probable es que recomiende nuestro producto / servicio / empresa a un amigo o colega?". La escala de la calificación va del 0 hasta el 10.

En este trabajo de tesis se planteó abordar un enfoque no estructurado es decir analizar los comentarios de los clientes con la finalidad de hallar insights que permitan generar los resultados que se buscan, en este trabajo se verán los pasos que se realizaron para identificar los motivos de no recomendación y finalmente pasar por la etapa de revisión de resultados y el modelado, en este caso clasificar los comentarios en función a los tipos de clientes: promotores, neutros y detractores.

Entre los principales hallazgos referidos a los motivos de no recomendación se encontró para el grupo de los NEUTROS los conceptos de difundir más la información y mejorar la rentabilidad mientras que para el grupo de los DETRACTORES se tienen los conceptos de mejorar la atención al cliente y reducción de comisiones, por el lado del modelo de clasificación, el que arrojo mejores resultados fue Naive Bayes los siguientes resultados de desempeñando (F_1 Score) fueron: para el grupo de Detractores 34.7%, para el grupo de Neutros 57.2% y para el grupo de Promotores 86.6%.

Palabras Clave: AFP, Net Promoter Score, comentarios, motivos, promotores, neutros, detractores, Naive Bayes, F_1 Score.

ABSTRACT

Currently the manager of the pension fund (AFP) of the Peruvian market is conducting telephone surveys to measure the level of service offered to its customers, using the Net Promoter Score a methodology that classifies customers according to the rating It gives the question: "How likely is it that you recommend our product / service / company to a friend or colleague?" The scale of the rating goes from 0 to 10.

In this thesis work, an unstructured approach was discussed, that is to analyze the comments of the clients in order to find insights that allow generating the results that are sought, in this work we will see the steps that were taken to identify the reasons for no recommendation and finally go through the stage of review of results and modeling, in this case classify the comments according to the types of clients: promoters, neutrals and detractors.

Among the main findings regarding the reasons for not recommending the concepts of disseminating information and improving profitability was found for the PASSIVES group, while for the group of DETRACTORS the concepts of improving customer service and reducing of commissions, on the side of the classification model, the one that produced the best results was Naive Bayes. The following performance results (F1 Score) were: for the group of Detractors 34.7%, for the passive group 57.2% and for the group of Promoters 86.6%.

Keywords: AFP, Net Promoter Score, comments, motives, promoters, neutrals, detractors, Naive Bayes, F1 Score.

INTRODUCCIÓN

La presente tesis se refiere al análisis de la retroalimentación (comentarios) que un cliente brinda cuando se le pregunta si recomendaría a una administradora de fondo de pensiones del mercado peruano. Además, de cómo la elaboración de un modelo de clasificación ayudará en el análisis de dichos comentarios con la finalidad de conocer y encontrar diferencias entre uno u otro comentario.

El desarrollo de esta tesis se realizó por el interés de identificar y conocer cuáles son las razones por las que un cliente recomendaría o no a la administradora de fondo de pensiones, este tipo de análisis va más allá de solo seguir los resultados obtenidos en las encuestas telefónicas (Net Promoter Score). Identificar estas razones permitirá tomar acción sobre ellas para poder mostrar mejorías en el indicador de recomendación.

La investigación se realizó tomando como fuente información las encuestas mensuales que se realizan a los clientes que tienen interacción con los distintos canales de la administradora de fondo de pensiones, dentro de estas encuestas que poseían distintos ítems, está la pregunta de recomendación la cual posee los comentarios que se han analizado, cabe mencionar que las encuestas se realizan de manera telefónica y es dirigida a una muestra probabilística.

Durante el desarrollo de la investigación, uno de los obstáculos que se tuvo fue la redacción, en muchos de los comentarios no era adecuada (los comentarios de los clientes eran transcritos durante la encuesta telefónica) la idea no era del todo clara lo que obligó a una corrección de la misma para poder hacer un correcto análisis, lo cual tomó un tiempo importante dentro del desarrollo de la investigación.

El objetivo de esta tesis es identificar las diferencias de los motivos de no recomendación a través de la retroalimentación de los clientes, además de ello establecer un modelo de clasificación que permita saber si los comentarios corresponden a un cliente detractor, neutro o promotor para finalmente identificar acciones clave en los procesos de atención para la mejora del índice de recomendación.

CAPÍTULO 1: PLANTEAMIENTO DEL ESTUDIO

1.1 Descripción del Problema

Desde hace ya unos años la experiencia del cliente se viene midiendo con la finalidad de evaluar la calidad del servicio que se le brinda, en este caso en particular se mide en el rubro de servicios financieros para conocer la situación actual y sobre todo mejorar, más aun con el auge de la transformación digital donde ya los canales convencionales de atención van perdiendo terreno, es muy importante escuchar a los clientes ya que son ellos los protagonistas de las atenciones. Las empresas que articulen sus procesos en base a la retroalimentación de sus clientes son en promedio 5% más eficientes y 6% más rentables (McAfee & Brynjolfsson, 2012).

Existen distintas metodologías para evaluar y sobre todo conocer la calidad de servicio, para este caso se usará el indicador Net Promoter Score (NPS), esta métrica fue desarrollada en el año 2003 por el consultor de gestión Fred Reichheld de Bain & Company en colaboración con la empresa Satmetrix, el NPS es usado por muchas empresas hoy en día debido a su fácil implementación además de proporcionar una retroalimentación de los clientes en base a la experiencia vivida durante la interacción con la empresa (SATMETRIX, 2014), para calcular el NPS es necesario hacer pregunta simple pero directa : "¿Qué tan probable es que recomiende nuestro producto / servicio / empresa a un amigo o colega?", basada en la respuesta brindada (escala del 0 al 10) un cliente puede clasificarse de 3 maneras:

- Si el resultado fue 9 o 10, serán considerados promotores
- Si el resultado fue 7 u 8, serán considerados neutros

- Si el resultado es menor o igual a 6, serán considerados detractores

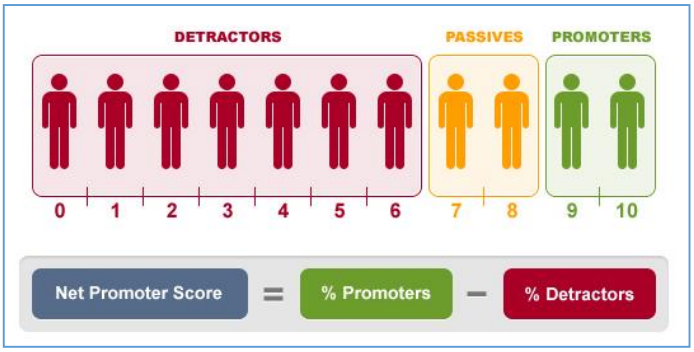

Figura 1: Metodología del cálculo NPS

Fuente: Customer Experience – Daniel Segarra

En la Figura 1 se puede ver que el Net Promoter Score (NPS) se calculan restando el porcentaje de los promotores y detractores, se deber tener en cuenta que es posible que el NPS puede llegar a ser negativo es decir que exista más detractores que promotores (EFQM, 2013).

1.2 Formulación del Problema

1.2.1 Problema general

En base a lo expuesto anteriormente, se formuló la pregunta de esta investigación: ¿En qué se diferencian las razones por las que los clientes detractores y neutros no recomiendan una administradora de fondo de pensiones privada en el mercado peruano?

Asimismo, se tuvieron los 2 siguientes problemas específicos:

- ¿Cómo implementar y desarrollar un modelo de clasificación de la información textual de los clientes que participan de la medición de calidad?
- ¿Qué descubrimientos y acciones clave se identifican en los procesos y protocolos de atención con la finalidad de mejorar el índice de recomendación?

1.3 Importancia y Justificación del Estudio

Actualmente en el Perú, el mercado de administradoras de fondo de pensiones maneja más de 154 mil millones de soles y cuenta con más de 6.5 millones de clientes de los cuales cerca de 3 millones son cotizantes (SBS Noviembre 2017), a raiz de los últimos cambios en la legislación peruana el modelo de servicio para este mercado ha cambiado radicalmente donde ya no es el foco el número de jubilaciones, ahora además de la rentabilidad que se ofrece, el servicio que se brinda es un factor fundamental para un cliente al momento de decidir quedarse o traspasarse.

Para ello la administradora de fondo de pensiones (AFP) tiene montado un sistema de medición de calidad a traves de sus distintas plataformas de atención, Los resultados de la medición por encuesta telefonica se despliegan sólo a las áreas que tienen contacto con el cliente (front office), la AFP ha visto a lo largo de los meses la evolución de su NPS, creciendo poco a poco en base a los resultados obtenidos, pero en la actualidad tiene varios problemas para lograr un incremento, el primero y quizás el más importante es "¿Qué hacemos para mejorar?".

El NPS en los últimos 6 meses se ha visto estancado y ya no basta solo con informar que el NPS en el último mes es 30% mientras que en el mes anterior fue de 31% (Profuturo AFP, 2017), otro problema a tener en cuenta es que el indicador NPS concentra toda la experiencia de servicio que vive el cliente, para él es indiferente la cantidad de procesos que se necesitan para cumplir con la atención que solicita, que pueden influir positiva o negativamente. Hay que tener en cuenta que la calificación obtenida es para todas y cada una de las áreas involucradas para realizar la atención que el cliente necesita, recordar que la pregunta es básicamente "¿Usted recomienda a la empresa?".

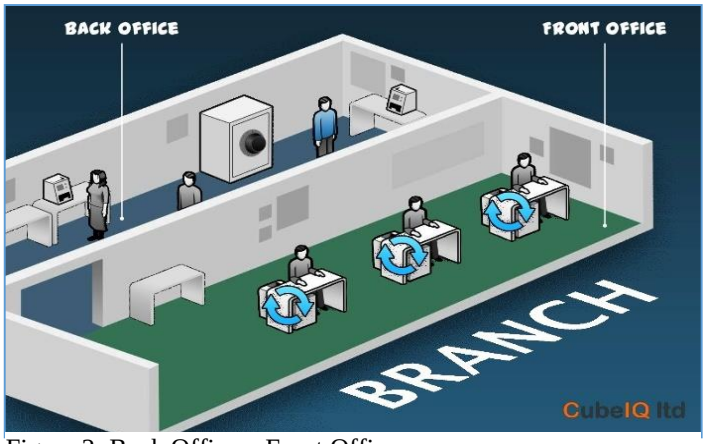

Figura 2: Back Office y Front Office

Fuente: Cubel Q – Andreas Papadedes

En la figura 2 se puede apreciar cómo operan las empresas que trabajan con los resultados que miden la experiencia del cliente, detrás del Front Office (cara visible al cliente) se encuentra el Back Office quienes son los que diseñan los procesos, como se mencionó anteriormente el índice NPS va para toda la empresas y todas las áreas que están de alguna manera relacionadas con los servicios que recibe el cliente.

Las empresas se han acostumbrado a trabajar con información estructurada ya que es fácil de recolectar, medir y sobretodo analizar, pero esta solo proporciona una imagen superficial de la situación (Caemmerer & Wilson, 2010), la nueva tendencia implica que se tome la información cualitativa con el fin de poder capturar la retroalimentación del cliente en formato contextual y no estructurada (Witell, 2011).

El cliente es quien tiene un papel principal redefiniendo los procesos que la empresa maneja, podría decirse entonces que este tipo de información tiene mayor relevancia que el enfoque estructurado porque su contenido refleja la motivación del cliente (Belkahla & Triki, 2011).

A través de un análisis de texto se podrán obtener las razones de recomendaciones de los grupos de clientes, aquellas empresas que articulen sus procesos en base a la retroalimentación de sus clientes son en promedio 5% más eficientes y 6% más rentables (McAfee & Brynjolfsson, 2012), basados en esta premisa si se llega a identificar dichas razones y se toman acciones sobre estas, la administradora de pensiones verá cambios que necesita para seguir mejorando.

En las últimas publicaciones de la consultora Satmetrix (SATMETRIX, 2014) el mejor NPS dentro de la industria financiera en el mundo es de 81%, mientras que el promedio es de 34%, este es un buen medidor para poder evaluar en qué situación se encuentra la empresa que hace uso de este indicador y metodología.

Para el caso de la AFP se sugiere iniciar con un análisis de texto no lingüístico como por ejemplo una nube de palabras para luego ya complementar con una análisis de texto lingüístico que implica la lematización además de capturar el contexto para luego poder hacer una diferenciación y clasificación de motivos de no recomendación, con ese conocimiento obtenido la idea sería implementar un grupo de trabajo que involucre a todas las áreas y donde ellos puedan conocer a mayor detalle los hallazgos encontrados en este proyecto, por ejemplo: mejorar el contenido de las comunicaciones, cumplir con las promesas o de reducir tiempo de espera para atención.

Como ya se había mencionado anteriormente, se tiene que conocer a profundidad las razones y establecer acciones para mostrar una diferencia, una mejoría en los procesos y claro en la atención, que el cliente perciba el cambio, esto implica evaluar completamente la experiencia de atención de los clientes a través de los canales de atención tanto presenciales como no presenciales y se propone abarcar dicha problemática con minería de texto en un nivel más avanzado para poder dar esos resultados.

1.4 Delimitación del Estudio

El presente estudio está delimitado para una Administradora de fondo de pensiones (AFP) dentro del mercado peruano del sistema privado de pensiones, está comprendido entre los periodos de Octubre 2016 y Mayo 2018. Como delimitación teórica de este estudio se tiene a los comentarios de los clientes que fueron encuetados en el periodo señalado.

1.5 Objetivos de la Investigación

1.5.1 Objetivo General

- Identificar a través de la retroalimentación de los clientes detractores y neutros las diferencias de sus razones de no recomendación usando técnicas lingüísticas y no lingüísticas de minería de texto.

1.5.2 Objetivos Específicos

- Implementar un modelo de clasificación de texto que permita predecir el tipo de cliente según sus comentarios.
- Identificar acciones clave que permitan redefinir y mejorar procesos, protocolos de atención, etc. con la finalidad de reducir en número a los grupos de clientes NEUTROS y DETRACTORES.

CAPÍTULO 2: MARCO TEORICO

2.1 Marco Histórico

El análisis de texto describe un conjunto de técnicas lingüísticas, estadísticas y de aprendizaje automático que modelan y estructuran el contenido de información de fuentes textuales para la inteligencia comercial, el análisis de datos exploratorios, la investigación entre otros.

El análisis de texto surgió por primera vez a fines de la década de 1990 como "minería de datos de texto" o simplemente como "minería de texto". Los enfoques iniciales consideraban una fuente de texto como una "bolsa de palabras". Evolucionaron para usar lingüística básica y superficial para manejar formas de palabras, variantes como como abreviaturas, plurales y conjugaciones, así como términos de varias palabras.

El término es usualmente relacionado a minería de textos, Ronen Feldman modificó una descripción del término cerca al año 2000 (Feldman, y otros, 1998) para cambiarla por Análisis de texto en el 2004 (Ronen Feldman Y. A.-Y., 2003).

Los primeros usuarios de minería de texto eran analistas de inteligencia e investigadores biomédicos que buscaban alguna relacion que podía ser detectada a través de un patrón de acciones y asociaciones una proteína cuya presencia activa o inhibe una vía genética que conduce a un cáncer que podría conocerse mediante la extracción de literatura biomédica. Estas son aplicaciones muy importantes, y este estilo de análisis se ha adoptado desde entonces para diversas aplicaciones

Ronen Feldman, un pionero en el campo habló en un panel de la Asociación de Maquinaria de Computación de 2006 sobre sistemas de minería de textos que podrán aprobar pruebas de comprensión de lectura estándar como SAT, GRE, GMAT, etc. Según Feldman, dicho sistema debe lograr un excelente reconocimiento de entidades y extracción de relaciones, con muy alta precisión (relevancia de la información recuperada) y recuperación (capacidad de encontrar toda la información que sea relevante). Según Feldman, estos sistemas deberían funcionar en cualquier dominio, operando de forma totalmente autónoma sin intervención humana, y deberían analizar enormes corpúsculos (conjuntos de documentos) y llegar a "hallazgos verdaderamente interesantes.

Es sabido que el 80% de la información relevante que se genera desde cualquier sector se origina en forma no estructurada, principalmente en texto. Estas técnicas y procesos descubren y presentan el conocimiento (hechos, reglas y relaciones) que de otro modo se bloquea en forma textual. El análisis de texto es una respuesta al problema de "datos no estructurados", y que es un problema que ha sido reconocido por décadas. De hecho, la primera definición de Business Intelligence (BI) en sí, en un artículo de IBM Journal se describe un sistema de BI que: "utilizará máquinas de procesamiento de datos para el auto-resumen y auto-codificación de documentos y para crear perfiles de interés para cada uno de los 'puntos de acción' en una organización. Tanto los documentos recibidos como los generados internamente se abstraen automáticamente, se caracterizan por un patrón de palabras y se envían automáticamente a los puntos de acción y apropiados (Luhn, 1958).

El análisis de texto se convierte en otro activo en el grupo de herramientas de análisis integrado. Tener la flexibilidad de adoptar cualquiera de los diversos enfoques de integración, eligiendo el enfoque que mejor se adapte al carácter de los datos y las necesidades de una empresa. Poder extraer las características de los almacenes de datos estructurados, donde la información de origen de texto se puede analizar junto con los datos numéricos generados por las aplicaciones operativas. La retroalimentación de los clientes en cualquier empresa del mundo de hoy debe ser considerada un aspecto crítico, ya que es invaluable para determinar qué es lo que le gusta y no le gusta al cliente (Raja, Sukhwani, & Khots, 2017), la minería de texto apoyada en el análisis de sentimiento de los comentarios de la encuesta ayudan a identificar los motivos por los cuales los clientes dan una nota, el uso de la minería de textos está creciendo rápidamente a medida que las organizaciones están dando valor a los datos no estructurados.

Con el avance de las soluciones de software para el análisis de texto se puede estar escuchando al mismo tiempo a todos con la finalidad de hacer las inversiones adecuadas para responder y actuar sobre esas retroalimentaciones. Las empresas que monten un programa de voz del cliente deben hacerlo en tiempo real en todas sus formas posibles con la finalidad de gestionar todos los comentarios en una plataforma, para que tenga sentido y entregar lo que realmente se necesita: ideas necesarias para tomar decisiones de negocios distribuidas a las personas adecuadas en el momento adecuado, se habla de 4 grandes componentes: escuchar la voz del cliente, interpretar los datos de los clientes para extraer información significativa, ofrecer información actualizable e informes personalizados, medir y rastrear tendencias a lo largo del tiempo para mejorar continuamente el programa.

Los casos de éxito de empresas que hoy en día que tienen todo un programa de experiencia para el cliente se han percatado que ya no son suficientes las fases de escuchar y responder a los clientes (Confirmit everywhere), el siguiente desafío para los profesionales de la experiencia del cliente es encontrar una diferenciación, para lograrlo debe haber un área que debe resolver estos problemas de manera eficiente y sistemática: como tratar los datos no estructurados y analizar el texto de diversas fuentes. El aumento exponencial en el volumen de este tipo de datos ya sean los verbatims de las encuestas, foros de clientes o sitios de medios sociales conduce a la conclusión inevitable de que ya no se puede ignorar si es que las empresas quieren permanecer en la cima. El método primario usado para registrar pensamiento y sentimiento es el texto, es importante saber y ahora más que nunca que es lo que están pensando los clientes.

2.2 Investigaciones relacionadas con el tema

Uno de los primeros investigadores de hablar sobre las reseñas y comentarios de un servicio que recibían los clientes fue Peter Turner, en su artículo “Thumbs Uo or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews” se presenta un algoritmo de aprendizaje que revisa las calificaciones que se le brinda a un cliente respecto a un servicio y/o producto, en este caso en el sector automotriz, menciona que la clasificacion de una revision es predicha por la orientacion semantica de las frases que contienen adjetivos y/o adverbios.

Turney menciona que una frase puede tener una orientacion positiva cuando existen buenas asociaciones pero tambien malas asociaciones como por ejemplo “muy arrogante”, menciona que un comentario se clasifica como recomendado si la orientacion semantica promedio de sus frases

es positiva, si está considerando unas vacaciones en Mexico, es posible que haga uso de un motor de busqueda e ingrese una consulta en este caso “Revision de viaje de Mexico”, sin embargo, en este caso google informa cerca de 5000 resultados, el detalle está en saber que cuales resultados recomiendan el destino, esto se puede realizar empleando un algoritmo que revise dichos comentarios y que los califique como aprobatorios o desaprobatorios, es más que hasta muestre estadisticas de resumen (Turney, 2002).

Jürgen Broß (2013), en su tesis doctoral “Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques” menciona que las opiniones y experiencias de otras personas constituyen una importante fuente de información en nuestra vida cotidiana. Por ejemplo, se pregunta a nuestros amigos qué dentista, restaurante o teléfono inteligente nos lo recomendaría hoy en día, las reseñas de clientes en línea se han convertido en un recurso invaluable para responder a tales preguntas. Además de ayudar a los consumidores a tomar decisiones de compra más informadas, los comentarios en línea también son de gran valor para los proveedores, ya que representan clientes genuinos a los que no se les ha solicitado retroalimentación, que está disponible prácticamente sin costo. Sin embargo, para productos populares a menudo existen varios miles de revisiones y puede que el análisis manual no sea una opción. En esta tesis, se ofrece un estudio exhaustivo de cómo modelar y analizar automáticamente la información rica en opiniones contenida en las reseñas de los clientes. En particular, se considera los aspectos orientados al análisis de los sentimientos. Dado un conjunto de reseñas, el objetivo de la tarea es detectar el individuo. Los revisores de los aspectos de los productos han comentado y deciden si los comentarios son más bien positivos o negativos. El desarrollo de sistemas de análisis de texto a menudo implica un trabajo tedioso y costoso, por ejemplo, etiquetar un corpus de capacitación para métodos de aprendizaje automático o construir bases de conocimiento de propósito especial. Como un tema general de la tesis, se examina la utilidad de las técnicas de supervisión a distancia para reducir la cantidad de supervisión humana requerida.

Francisco Villaroel, J. B. (2014), en su investigacion “Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach.” propone un marco que ofrece un enfoque holístico para analizar la retroalimentación de los clientes mediante componentes del proceso de creación de valor: actividades, recursos y contexto. Se plantea que la retroalimentación textual del cliente no solo puede clasificarse como positivo o negativo, sino que también se le puede asignar

una cadena de actividades y recursos que describen como se crea el valor en un contexto particular. El marco ARC se aparta del simple análisis basado en resultados por ejemplo frecuencia de palabras o análisis de sentimiento, este ofrece un nuevo enfoque de procesos e interacciones que puede ser aplicado con la minería lingüística de textos, que captura los elementos clave del servicio en la retroalimentación, el marco ARC guía la utilización de funciones de minería de texto basadas en la lingüística, en el desarrollo de un modelo de minería de texto para el análisis de retroalimentación de clientes. Este proceso propone un desarrollo de un modelo que captura el contexto del cliente tanto personal como situacional, actividades y recursos además del sentimiento asociado, es por lo tanto un cuadro más completo y holístico de los clientes, el caso de estudio en donde se aplicaron dichas técnicas fue en un aeropuerto de Gran Bretaña, específicamente en el servicio de estacionamiento y transferencia. Entre los principales hallazgos se encontró que muchos de los comentarios de elogios se centraron en la facilidad para estacionar sus vehículos mientras que los comentarios de quejas hacian referencia a aspectos como la señalizacion, el personal y otros recursos del estacionamiento.

Pribil, J. (2014), en su investigacion “Text Mining in Business Practice” menciona que el análisis de texto es un elemento clave en el proceso de negocio, las empresas no solo compiten con sus competidores de mercado sino también con sus clientes ya que con sus opiniones dan forma a una buena o mala impresión sobre la compañía, sus productos y servicios en otras palabras identificar el sentimiento del mercado y la posibilidad de responder a estos eventos.

Crijns,Tanja (2016), en su tesis “Classifying events to ugenda calendar genres” hace uso de una técnica de Text mining que es la clasificación de textos con la finalidad de clasificar todos y cada uno de los eventos de la pagina web de Ugenda en base a informacion como: el titulo, la descripcion, el lugar, el horario y demas variables que pudieran ayudar a la clasificacion de estos eventos. En la tesis se menciona que se logró automatizar la clasificación de futuros eventos que iban publicándose en la pagina web que de hacerlos de manera manual iban a incurrir en un gasto de tiempo y equipo.

2.3 Estructura teórica y científica que sustenta el estudio

2.3.1 ¿Qué es retroalimentación?

Para Villaroel, la naturaleza de los comentarios de los clientes se puede clasificar como explícita o implícita dependiendo de si los clientes, consciente o inconscientemente, proporcionan información a un tercero sobre sus experiencias. Las empresas tradicionalmente han recopilado comentarios explícitos a través de plataformas (encuestas, correo electrónico, revisiones en línea, buzón de sugerencias, redes sociales etc.), en las cuales solicitan directamente información de los clientes sobre sus experiencias en cualquier interacción que tenga con la empresa.

Los clientes dan retroalimentación implícita a través de acciones determinadas, sin que la empresa solicite información (por ejemplo, seguimiento de ojos, tiempo de lectura, cantidad de desplazamiento, cantidad de clics obtenidos en una red). Ambos tipos de comentarios contribuyen a un continuo proceso de aprendizaje sobre los clientes. Que es lo que finalmente se desea adecuar: la empresa al cliente y no lo contrario, eso implica que muchos de los procesos o protocolos que maneja la empresa sean redefinidos y/o adaptados de modo que el cliente perciba un cambio en su experiencia en cualquier interacción.

Muchas empresas recopilan comentarios explícitos utilizando métodos cuantitativos debido a la simplicidad en el análisis de información estructurada. Por ejemplo, encuestas (es decir usando escalas de Likert de 5 puntos o más) este tipo de métodos ayudan a las empresas a analizar atributos predeterminados de calidad de productos / servicios. A pesar de la importancia de estos datos, la evaluación de un servicio utilizando los atributos predeterminados dará como resultado una comprensión incompleta de la experiencia del cliente. Clasificar atributos en dimensiones de calidad predefinidas que luego se utilizan para recopilar retroalimentación estructurada proporciona a las empresas solo información superficial sobre la experiencia completa del cliente y no puede capturar todos los recursos y actividades involucradas o el contexto. La nueva tendencia implica que se tome la información cualitativa con el fin de poder capturar la retroalimentación del cliente en forma contextual y no estructurada de tal manera que se tenga una vista mucho más completa de la experiencia del cliente con la empresa.

En contraste con estos enfoques cuantitativos que solicitan datos estructurados, los avances ofrecen a los clientes nuevos canales y plataformas a través de los cuales pueden proporcionar servicios solicitados y por ello la retroalimentación será cualitativa no solicitada en un formato textual no estructurado. Esta forma de comentarios incluye respuestas a preguntas abiertas, correos electrónicos, reseñas en línea y redes sociales conversaciones, etc. Aquí, el cliente toma el rol principal, definiendo activamente el proceso y tiempo de retroalimentación y el contexto en el que se proporciona la información. Podría decirse que este tipo de retroalimentación es de mayor relevancia que los enfoques estructurados porque su expresión y el contenido reflejan mejor la motivación del cliente ya que va más allá de un número en una escala.

Reconociendo que los clientes desempeñan un papel activo en la generación de información cualitativa relevante, ellos podrían proporcionar fuentes de información más valiosas y completas a las empresas, Sin embargo, el análisis de esta información exige un esfuerzo significativo en el tiempo requerido para generar conocimiento a partir de grandes cantidades de datos cualitativos. De hecho, comentarios de correos electrónicos, comentarios de clientes, mensajes cortos y redes sociales los medios están creciendo rápidamente, empujando a las organizaciones a desarrollar enfoques más eficientes para medir y comprender la información. Los métodos de minería de textos ofrecen un potencial una solución para lidiar con los grandes volúmenes de datos no estructurados, ya sean datos explícitos o implícitos y solicitados o no solicitados (Villaroel, Burton, Theodoulidis, Gruber, & Zaki, 2014).

2.3.1.1 ¿Cómo se relaciona la minería de texto con la retroalimentación del cliente?

La minería de texto es el proceso de analizar información textual en un intento de descubrir la estructura y significados implícitos "ocultos" dentro de los textos (Mykroyannidis & Theodoulidis, 2009). Es un desarrollo tecnológico relativamente reciente que aborda el problema de la gestión de la información mediante el uso de técnicas de extracción de datos, aprendizaje automático, procesamiento del lenguaje natural, recuperación de información y gestión del conocimiento (Chong Ho, Jannasch-Pennell, & DiGangi, 2011). Más específicamente, la minería de texto implica el procesamiento de una colección de documentos, o corpus, en los que los documentos se convierten en datos estructurados, de modo que cada documento se describe utilizando un conjunto

de características llamadas conceptos para proporcionar una perspectiva holística de información textual y no textual (Mykroyannidis & Theodoulidis, 2009).

La minería de texto es un desarrollo tecnológico con un potencial altamente comercial (Owens et al.2009). Por ejemplo, los estudios indican que el 80% de la información de la empresa está contenida en el texto (Ur-Rahman y Harding 2011). Los estudios también han demostrado los beneficios de automatizar el análisis de grandes cantidades de datos cualitativos de retroalimentación de los clientes relacionados con el turismo y servicios de fabricación, han demostrado cómo gestionar y convertir clientes de manera efectiva (Lau, Lee y Ho 2005; Ludwig et al., 2013).

Los enfoques descritos en la literatura se pueden agrupar en lingüísticos y no lingüísticos (Taboada et al. 2011). Las técnicas lingüísticas consideran las características del lenguaje natural del texto en los documentos (por ejemplo, sintaxis, gramática), mientras que las técnicas no lingüísticas ven documentos como una serie de personajes, palabras, oraciones, párrafos, etc. (Ur-Rahman y Harding 2011). Las técnicas no lingüísticas tratan cada documento como una lista de términos, contando el número de veces aparecen palabras específicas en un documento y calculan su proximidad a otros términos relacionados en el documento o en documentos relacionados (Zhong, Li y Wu 2012).

La minería de textos basada en la lingüística a menudo hace uso de recursos externos, como WordNet, la base de datos en línea más grande de términos lingüísticos en inglés que contiene palabras relacionadas con el significado (http://wordnet.princeton.edu/). Estos recursos pueden pertenecer a un lenguaje natural específico (por ejemplo, diccionario, tesauro). Sin embargo, los conceptos en un documento particular también podrían referirse a un área temática específica, llamada dominio (por ejemplo, servicios financieros, biología), en cuyo caso también podría ser apropiado para usar recursos específicos del dominio, como léxicos, taxonomías y ontologías. Previamente la investigación en esta área ha demostrado que los modelos de minería de textos basados en la lingüística pueden superar a la categorización manual (humana) de las reseñas de los clientes porque son más precisas para predecir la calificación asociada con las revisiones (Ghazvinian 2011).

Sobre esta base, la minería de textos se puede clasificar además como dominio independiente o dominio dependiente. La minería de texto de dominio independiente aún puede implicar el uso de recursos de lenguaje natural, pero su uso es independiente de cualquier cuerpo específico de conocimiento, teoría o dominio; por lo tanto, podría argumentarse que puede aplicarse a todos los documentos de un idioma. En general, la aplicación de minería de textos es más eficaz si se incorpora algún nivel de especificidad de dominio en el análisis (Bhuiyan, Xu y Josang 2009).

2.3.2 Clasificación de texto

Aprender a procesar y comprender el texto es uno de los primeros pasos en el camino hacia obtener información significativa a partir de datos textuales, este es un subtema del aprendizaje automático que cada vez se vuelve más importante debido a la gran cantidad de información en formato texto. Es importante entender cómo se estructura el lenguaje y los patrones de sintaxis de texto, pero esto no es suficiente para muchas de las empresas que desean extraer conocimiento y sacar el máximo provecho. Conocer el procesamiento del lenguaje apoyado con conceptos de análisis y aprendizaje automático es decir Machine Learning (ML) ayudan a la construcción de sistemas que pueden aprovechar los datos tipo texto que no requieran demasiada supervisión manual para resolver problemas prácticos del mundo real es ahí donde se ve el beneficio para las empresas, un caso muy conocido es la clasificación de correos spam.

La clasificación de texto es uno de los problemas más grandes y desafiantes en el mundo del text mining (Dipanjan, 2016), el concepto puede parecer no tan complejo es más cuanto tiene una pequeña cantidad de documentos puede mirar en cada uno de estos y obtener alguna idea de lo que trata este documento, o para fines del proyecto se puede hablar de un comentario o retroalimentación, según este conocimiento se puede agrupar documentos o comentarios en clases similares o categorías, pero que sucede cuando se está frente a miles o millones de documentos o comentarios, aquí es donde la técnica mencionada resulta útil. Hay que tener en cuenta que la clasificación no solo se limita a data estructurada sino también no estructurada de todo tipo además de texto, se aplica en música, imágenes, videos y demás medios.

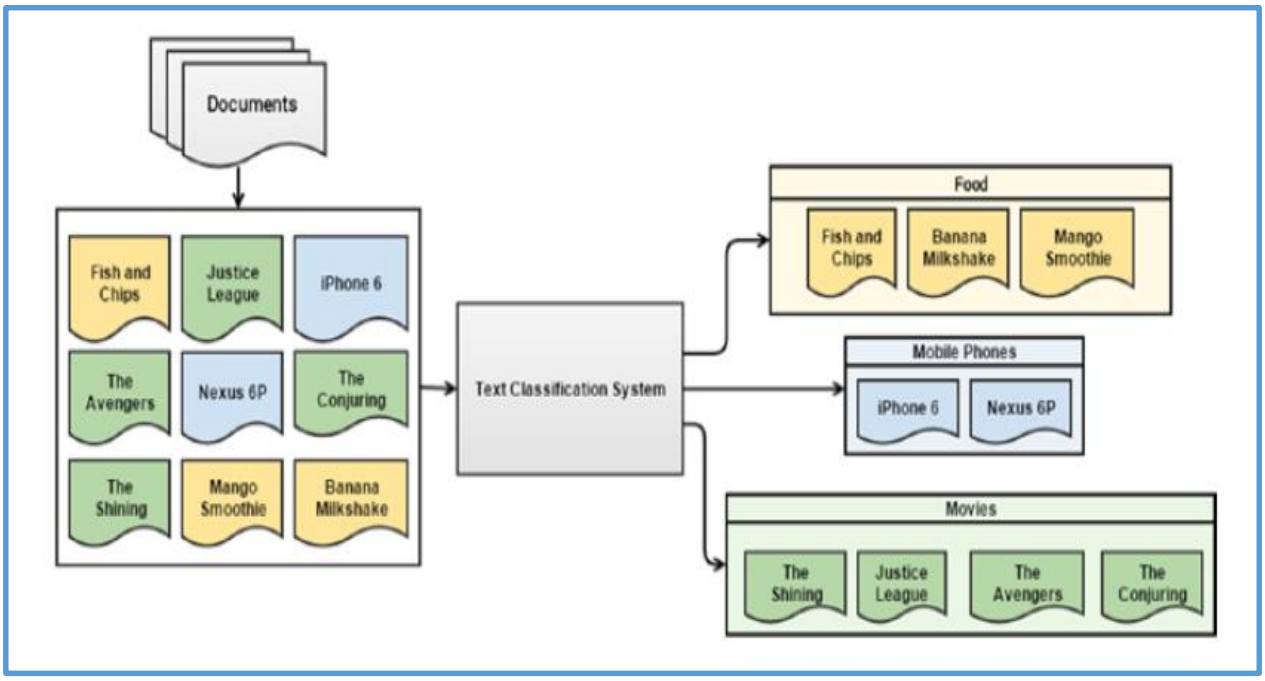

Figura 3: Descripción de clasificación de textos

Fuente: Text Analytics – Dipanjan Sarkar

En la Figura 3 se puede observar que hay múltiples documentos que representan productos que se pueden asignar a varias categorías de alimentos, teléfonos celulares y películas. En un inicio toda esta información se encontraba junta, al igual que un repositorio de documentos.

Una vez que estos datos pasan por un sistema de clasificación de texto, representado por una caja negra se puede notar que cada documento (información) se encuentra en una clase o categoría especifica que se había definido previamente. Para este ejemplo en particular se ha usado solo los

nombres de los documentos pero en otros casos pueden existir atributos adicionales como el género de una película, especificaciones del producto, componentes y muchas propiedades más que se pueden usar como características en el sistema de clasificación de texto que facilite la identificación y posterior clasificación de los documentos, en la actualidad existen diversos algoritmos de clasificación además de parámetros que pueden mejorar los resultados de dichos algoritmos que serán evaluados mediante sus indicadores de desempeño.

Al igual que cuando se trabaja con información estructurada (Data mining) existe una metodología de trabajo ya sea CRISP, KDD, etc. Para text mining también existe un flujo de trabajo para construir un clasificador de texto automático, este consistirá de pasos que se deben de seguir en la fase de entrenamiento y prueba. Para implementar un sistema de clasificación de texto se debe asegurar una fuente de datos que será la que alimentara al sistema, los siguientes pasos resumen un flujo de trabajo típico para un sistema de clasificación de texto sin importar el algoritmo o la herramienta a usar, este flujo parte del supuesto de que ya se posee un conjunto de datos ya descargado y listo para ser utilizado.

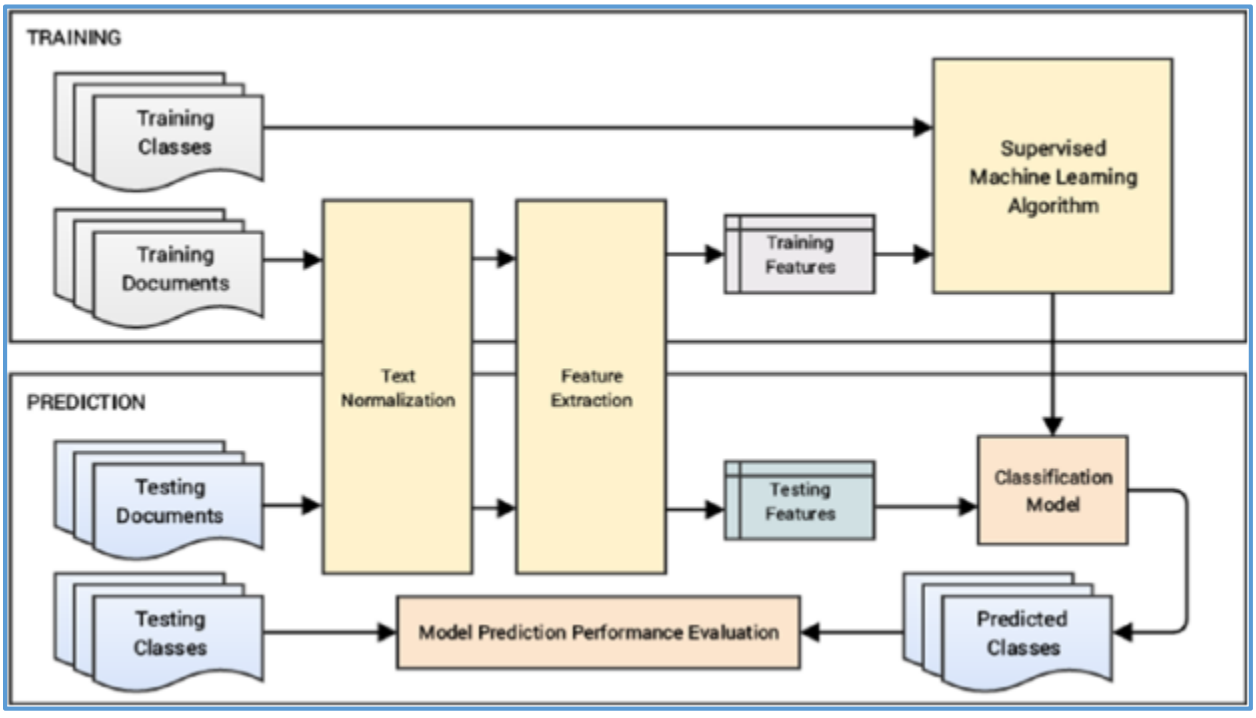

Figura 4: Plan para la construcción de un sistema automatizado de clasificación de texto

Fuente: Text Analytics – Dipanjan Sarkar

En la Figura 4 se puede observar que existen 2 grandes cuadros uno para el entrenamiento y otro la predicción, estos son los principales procesos en la construcción de un clasificador de texto (es muy usual tener estos 2 grupos en los proyecto de analítica), otros de los procesos importantes son la normalización de texto (eliminación de palabras sin contenido, extracción de la raíz de la palabra y demás proceso que permitan un trabajo adecuado con los textos) y la extracción de características (fase en donde se escogen las textos que son relevantes al momento de la construcción del modelo) , una característica que se puede notar para este flujo de trabajo es que tanto para el entrenamiento como la predicción los textos se pasaran por el proceso de normalización y extracción de características, estos procesos son siempre uniformes para poder garantizar que el sistema de clasificación se desempeñe de manera consistente.

2.3.3 Gestión de la experiencia del cliente

Gestionar la experiencia del cliente no es fácil, más aun en estos tiempos de entornos variables, clientes que ahora son más exigentes y que están dispuestos a comunicar sus necesidades y soluciones de manera más enfática e inmediata, casi instantáneas se podría decir a raíz de la revolución digital, dado este escenario actual se ha vuelto una necesidad aprender a identificar las necesidades de los clientes, traduciendo sus mensajes de cualquier fuente que vengan ya sea oral o escrito aplicando múltiples técnicas que ayuden a mejorar el conocimiento sobre la información entregada. Pero además de conocer esas necesidades es más importante aún analizar la experiencia que está viviendo el consumidor con cualquier interacción que ha tenido con cualquier empresa, con el fin de conocer desde su opinión personal de cómo lo está haciendo y desde esa evaluación mejorar la gestión de las experiencias que vive el cliente con la empresa.

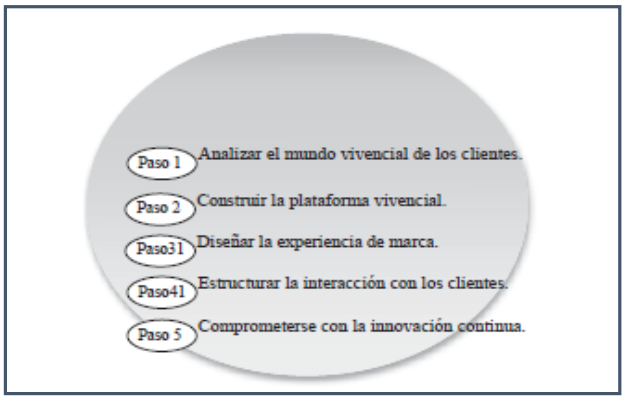

Figura 5: Los 5 pasos del marco de trabajo de CEM

Fuente: Proceso de gestión de experiencia - Schmitt

En la Figura 5 se observan los 5 pasos que tiene el modelo de la gestión de la experiencia del cliente (CEM por Customer Experience Management) de Schmitt (2003), el objetivo de CEM es mejorar el valor que recibe el cliente mediante la administración de la experiencia que este tiene. El autor señala que la experiencia es algo que va muy de la mano con el proceso, a diferencia de la satisfacción del cliente que en este caso se lo relaciona más con la funcionalidad de algún producto. Al comprender y administrar la experiencia del cliente en una compañía se puede identificar como añadir valor a los clientes y que estos lo perciban mediante su producto y quizás como consecuencia de ello los resultados de satisfacción se incrementen de manera natural. La gestión de experiencias del cliente es definido por Schmitt como un proceso de administración estratégica de la experiencia total de un cliente con un producto o una compañía. Tiene una visión que más allá del enfoque de gestión de relaciones con los clientes, dado que no solo es un registro de las transacciones sino que se enfoca en crear relaciones sustanciosas con los clientes, se espera que CEM fomente la lealtad de los clientes y por ende que agregue valor a la compañía que a su vez generará mayor valor financiero para las organizaciones.

2.4 Definición de términos básicos

- Analytics*:* Es el descubrimiento, la interpretación y la comunicación de patrones significativos en los datos.
- Modelo*:* Es una representación formal de una teoría.
- Call Center*:* Es un área donde agentes, asesores, supervisores o ejecutivos, especialmente entrenados, realizan llamadas (llamadas salientes o outbound) hacia clientes (externos o internos), socios comerciales, compañías asociadas u otros.
- Minería de texto: es el análisis de información no estructurada que se basa en el uso de técnicas lingüísticas, no lingüísticas, modelamiento estadístico y técnicas de aprendizaje automático para descubrir conocimiento.
- Procesamiento de Lenguaje Natural: es el campo que combina las tecnologías de la ciencia computacional como la inteligencia artificial, aprendizaje automático e inferencia estadística, apoyado con la lingüística aplicada con el objetivo de hacer posible la comprensión y el procesamiento asistido por un computador.
- Análisis de sentimiento: es el proceso de determinar el tono emocional que hay detrás de una serie de palabras que se utiliza para entender las actitudes, opiniones y emociones expresadas,

se basa en el procesamiento de lenguaje natural, análisis de texto y lingüística computacional con la finalidad de identificar y extraer información relevante de manera automática.

- Retroalimentación: son los mensajes o respuestas abiertas que dan los clientes cuando se les está aplicando una encuesta que medirá la experiencia de servicio.
- NPS (índice de recomendación): es un indicador que mide la recomendación de los clientes de una empresa.
- Verbatim es la reproducción exacta de una oración, frase, cita u otra secuencia de texto desde una fuente a otra. Las palabras aparecen en el mismo lugar, en el mismo orden, sin paráfrasis, sustitución o abreviación de cualquier tipo, sin realizar siquiera un cambio trivial que pueda alterar el significado.
- N-gram: en los campos de la lingüística computaciones y la probabilidad, un n-gram es un secuencia continua de n elementos de una muestra de datos de texto o voz, los elementos pueden ser fonemas, sílabas, letras, palabras. Los n-grams se recolectan de un cuerpo de texto o discurso, usando prefijos numéricos latinos se tiene que un n-gram de tamaño se denomina unigram, el de tamaño 2 se denomina bigrama y así sucesivamente.
- Recall: en el contexto de modelamiento, es la fracción de instancias de una clase que se predijo correctamente, por ejemplo para una búsqueda de texto en un conjunto de documentos, recall es el número de resultados correctos dividido por el número de resultados que deberían haberse devuelto.
- Precision: en el contexto de modelamiento, es la fracción de predicciones correctas para una determinada clase, por ejemplo para una búsqueda de texto en un conjunto de documentos, precision es el número de resultados correctos divididos por el número de todos los resultados devueltos.
- F_1 Score: en el contexto de modelamiento, es la medida que combina precision y recall usando la media armónica.

2.5 Hipótesis

2.5.1 Hipótesis General

Existen diferencias en las razones por las que los clientes detractores y neutros no recomiendan una administradora de fondo de pensiones en el mercado peruano.

2.5.2 Hipótesis Específicas

- Ayuda en la clasificación de comentarios el uso de un modelo de minería de texto
- Ayuda en la medición de calidad la gestión adecuada de los clientes detractores y neutros, en base a las acciones clave identificadas.

2.6 Variables

2.6.1 Verbatim

Es el comentario que el cliente brinda que durante la encuesta telefonica en el cual reconomienda o no recomienda los servicios de la administradora de pensiones, a partir de esta variable se desarrollo gran parte de esta tesis.

2.6.2 Marca de clase NPS

Durante la encuesta telefonica existe una pregunta donde se pide al cliente dar una puntuacion en la escala del 0 al 10, dependiendo de esta puntuacion el cliente puede ser
Promotor si el puntaje es 9 o 10, neutro si es 7 u 8 o detractor si es menor a 8, en base a esta variable se realizará la clasificación de textos.

CAPÍTULO 3: METODOLOGÍA DE INVESTIGACIÓN

3.1 Tipo, método y diseño de la investigación

La investigación según el objetivo es de tipo aplicativa, por los datos que se usaron es cualitativo, de método explicativo ya que busca las razones de recomendación y no recomendación, por el grado de manipulación de las variables es una de tipo no experimental ya que no se tiene un control sobre la variable de estudio, y de diseño transversal puesto que el estudio de la variable es en un solo periodo de tiempo (Hernandez Sampieri, Fernandez Collado, & Baptista Lucio, 2010).

3.1.1 Diseño

Para el desarrollo de la investigación de esta tesis:

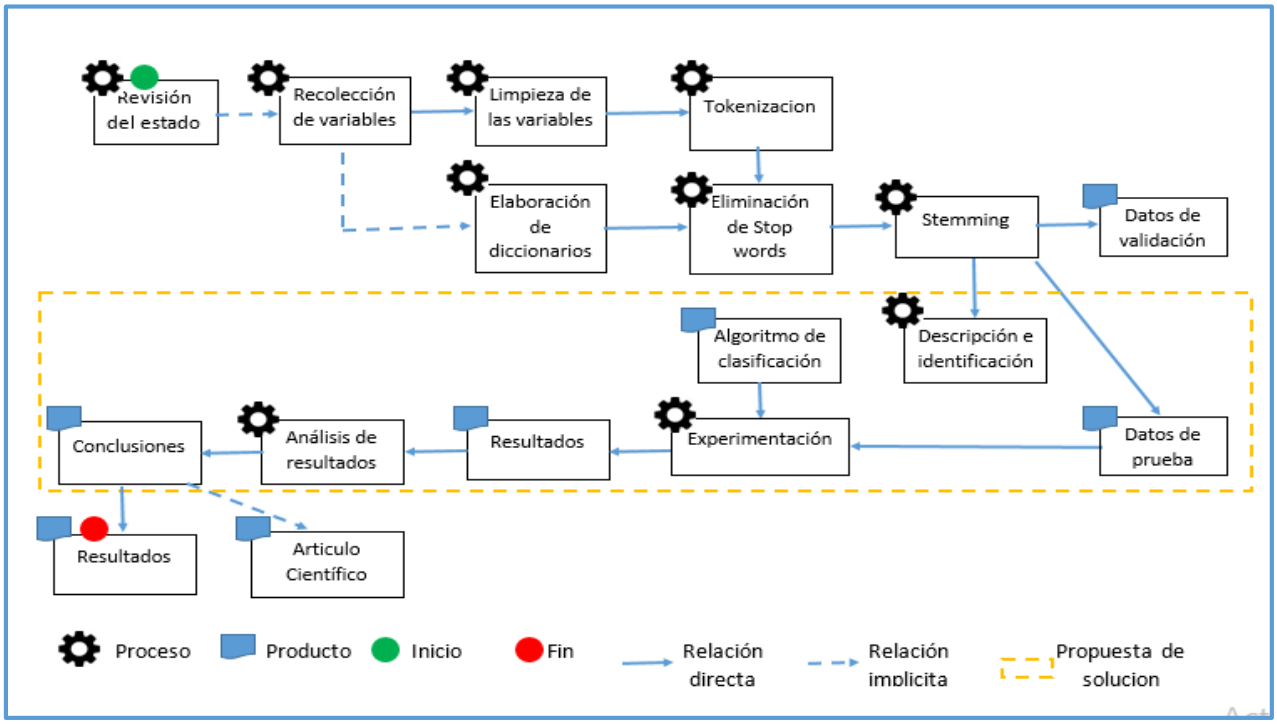

Figura 6: Flujo de trabajo para el proceso de clasificación de texto

Elaboración: Propia

De la Figura 6 se puede apreciar que la primera parte del diseño se basa en la recolección de las variables que provienen de la transcripción de las encuestas telefónicas que se realizan a los clientes con una frecuencia mensual, lo que sigue es la etapa del pre-procesamiento que consiste en la tokenizacion, limpieza de stop Words y finalmente la lematización, más adelante se detalla cada una de estas sub-etapas del pre-procesamiento, cabe mencionar que tanto que para la muestra de entrenamiento y de prueba se deberán aplicar dichas sub-etapas.

Luego de la limpieza y uniformización de los datos se abarcó la etapa descriptiva donde se profundizó en el contenido de los comentarios para identificar y diferenciar las principales ideas que se presentan en los 3 tipos de cliente según índice NPS: promotores, neutros y detractores, esta etapa tendrá como objetivo conocer un poco más sobre las razones que motivan la recomendación como la no recomendación.

En la siguiente etapa se vio la clasificación de texto donde a partir de los comentarios vertidos se establecieron reglas que determinarán si dicho comentario corresponde a un promotor, neutro o detractor, para esta tesis se plantea usar 2 algoritmos de clasificación: SVM y Naive Bayes en los cuales se mostraran los resultados de la clasificación así como las fortalezas y oportunidades de mejora en cada uno de los algoritmos para luego proceder a emitir conclusiones.

3.2 Población y muestra

3.2.1 Población

La población de estudio estuvo conformada por todos los clientes que han tenido alguna interacción con los canales de atención de la administradora de fondo de pensiones durante los periodos de Octubre 2016 hasta Mayo 2018.

3.2.2 Muestra

La muestra estuvo conformada por todos los clientes que han participado en las encuestas telefónicas durante los periodos de Octubre 2016 hasta Mayo 2018, esta muestra responde a un diseño estratificado por canal de atención (presencial, telefónico y web) asi como a un presupuesto anual previamente establecido.

3.3 Técnicas e instrumentos de recolección de datos

La recolección de los datos, en este caso los comentarios de los encuestados son recogidos luego del termino de la medición telefonica, a continuacion se mostrara un esquema que muestra el flujo de recoleccion mensual:

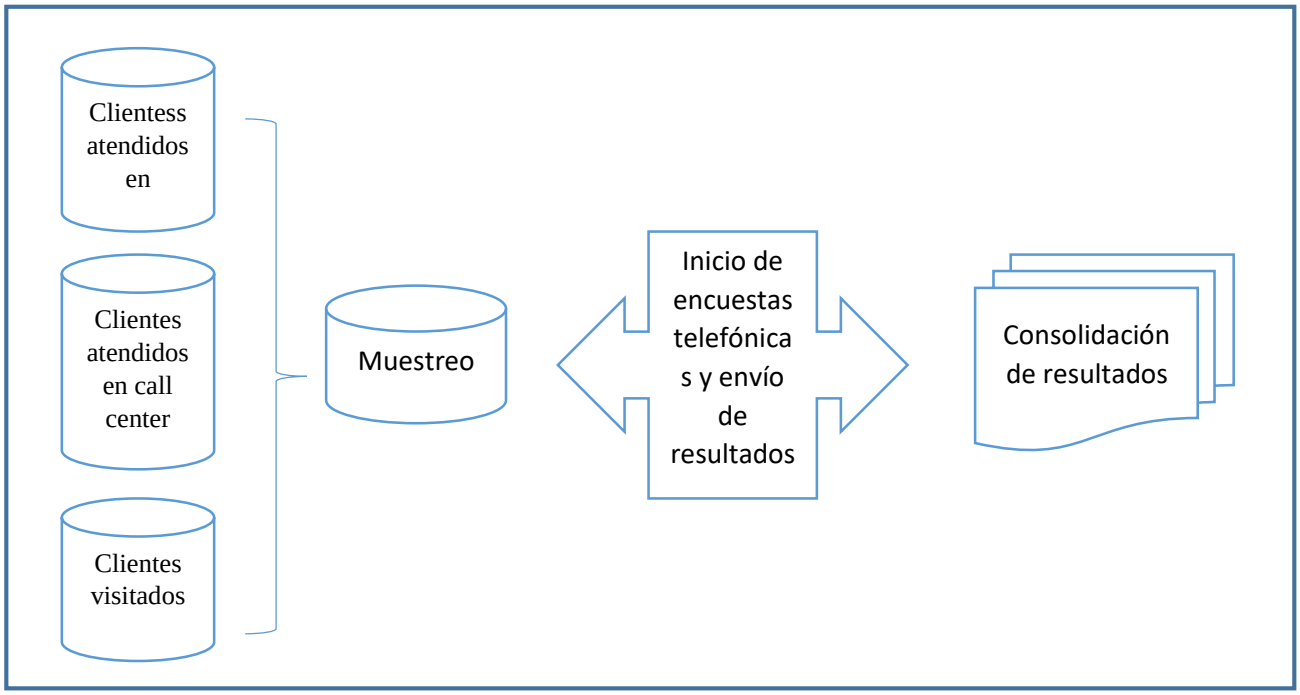

Figura 7: Flujo de recolección de datos

Elaboración: Propia

En la Figura 7 se observa que lo primero que se hace es consultar por base de datos todos aquellos clientes que han tenido alguna interacción con la administradora de fondo de pensiones a través de sus diferentes canales de atención: agencias presenciales, call center (centro de contacto), visitas de asesores. Luego de la consolidación se toma una muestra estratificada por canal de manera que se pueda tener un resultado de la situación de cada uno de estos, la base de datos se enriquece con datos importantes que facilitaran la encuesta como por ejemplo: nombre del cliente, edad, departamento, teléfonos y motivo de interacción.

La base es enviada al proveedor que carga dicha base en sus sistemas para realizar las encuestas telefónicas, al final del trabajo realizado el proveedor envía en un archivo Excel los resultados de cada una de las encuestas. Como se mencionó anteriormente los comentarios de los clientes son transcritos por los ejecutivos del proveedor quienes tienen la misión de capturar el comentario tal como lo dijo el cliente, ya con esta base consolidada se procede a recoger los campos necesarios como grupo (promotor, neutro y detractor) según su puntaje de recomendación y el comentario del cliente.

3.4 Descripción de procedimientos de Análisis

3.4.1 Técnicas de pre-procesamiento

3.4.1.1 Tokenizacion

Es el primer paso en la transformación del conjunto de datos, antes de que se pueda usar un texto para la clasificación, es necesario poder identificar las características separadas de un texto. Esto se puede hacer en diferentes formas. Por ejemplo, dividir el texto en un conjunto de palabras o dividirlo en un conjunto de secuencias de palabras. Estas palabras o secuencias de palabras serán referidas como tokens en adelante.

La división del texto en un conjunto de palabras se realiza mediante la búsqueda de separadores de fichas, dos ejemplos de separadores que son fáciles de determinar. Son espacios en blanco e indicadores de nueva línea como "\ n". Sin embargo, algunos separadores son ambiguos, como un punto. Puede ser parte de un token. Cuando está en un número como 35.0, pero también puede indicar el final de una sentencia, por ejemplo se tiene el siguiente comentario:

"La atención en el centro de contacto fue mala porque no me solucionaron mi problema"

La tokenizacion por palabras es:

"La" "atención" "en" "el" "centro" "de" "contacto" "fue" "mala" "porque" "no" "me" "solucionaron" "mi" "problema"

De la anterior oración se han obtenido 15 palabras, para este trabajo de tesis se realizó la tokenizacion por palabras, es decir que se deja de trabajar con una oración para trabajar con 15 palabras.

3.4.1.2 Lematización

La lematización reduce las palabras inflexionadas de manera adecuada, asegurando que la palabra raíz pertenezca al idioma. Toma en consideración el análisis morfológico de las palabras. Para hacerlo, es necesario tener un diccionario detallado para que el algoritmo pueda revisar para vincular el formulario a su lema. En Lematización la palabra raíz se llama Lema. Un lema es la

forma canónica, forma de diccionario o forma de cita de un conjunto de palabras, la clave de esta metodología es la lingüística.

Para este trabajo de tesis se construyó un diccionario específico que asegura una adecuada lematización, se tomó como base el diccionario que se encuentra en la siguiente dirección: https://github.com/snowballstem/snowball (el diccionario se encuentra en diferentes idiomas), se tiene por ejemplo:

Tabla 1

Tabla de Lemas

Lema	Palabra
Información	Info
Información	Informaciones
Información	Informa
Información	Informame
Información	Informándolo
Información	Informándome
Informador	Informadora
Informador	Informadores

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 1 se puede observar que este diccionario fue enriquecido en el contexto del fondo de pensiones privado y siendo más especificos en la administradora del cual se tiene la informacion, de todos los procesos que se realizaron este fue el mas tedioso y el mas complejo, con la finalidad de obtener resultados de acuerdo al contexto.

3.4.1.3 Eliminación de Stop Words

Luego de realizar la normalización de los comentarios es decir la division por palabras (tokens) y la lemmatizacion, sigue realizar la eliminación de palabras que no aportan valor o las que carecen de algun significado como por ejemplo articulos, pronombres o signos de puntuacion. Cabe mencionar que para este trabajo se enriqueció un diccionario ya existente de acuerdo al contexto del mercado privado de pensiones.

3.4.2 Representación vectorial

Uno de los principales intereses en text mining y el procesamiento de lenguaje natural (NLP), es cuantificar la temática de un texto, así como la importancia de cada término que lo forma. Una manera de medir la importancia de un término dentro de un documento es utilizando la frecuencia con la que aparece (*tf, term-frequency*). Esta aproximación, tiene la limitación de atribuir mucha importancia a aquellas palabras que aparecen muchas veces aunque no aporten información selectiva. Por ejemplo, si la palabra *matemáticas* aparece 5 veces en un documento y la palabra *página* aparece 50, la segunda tendrá 10 veces más peso a pesar de que no aporte tanta información. Para solucionar este problema se pueden ponderar los valores *tf* multiplicándolos por la inversa de la frecuencia con la que el término en cuestión aparece en el resto de documentos del corpus (*idf*). De esta forma, se consigue reducir el valor de aquellos términos que aparecen en muchos documentos y que, por lo tanto, no aportan información selectiva. El estadístico tf-idf mide la importancia de un término en un documento teniendo en cuenta la frecuencia con la que ese término aparece en otros documentos.

3.4.3 Algoritmos de clasificación

Un clasificador es un algoritmo que puede predecir las etiquetas de datos, Hay muchos tipos diferentes de clasificadores que son adecuados para diferentes problemas. Elegir el correcto es crucial para el rendimiento del programa. La elección depende principalmente de la forma y el tamaño del conjunto de datos.

Cuando se tiene un pequeño conjunto de datos, la varianza suele ser alta. En este caso, un clasificador que puede lidiar con una gran variación como Naive Bayes es una buena opción.

El alcance de la investigación es la clasificación de texto. Para lo cual existen diferentes clasificadores de texto y donde se encontró lo siguiente (Sebastiani., 2002):

1. Métodos ensamblados, máquinas de soporte vectorial (SVM), y métodos de regresión entregan un mejor desempeño.

2. Las redes neuronales y los clasificadores lineales funcionan muy bien, aunque un poco peor que los métodos mencionados anteriormente.
3. Clasificadores lineales por lotes y clasificadores de Naive Bayes suelen ser usados como línea base de los clasificadores.

Para este trabajo de tesis los algoritmos escogidos a desarrollar son SVM y Naive Bayes.

3.4.3.1 Máquinas de soporte vectorial (SVM)

Las máquinas soporte vectorial (SVM) son algoritmos de aprendizaje supervisados, utilizados para la clasificación, regresión o detección de valores atípicos. Considerando un problema de clasificación binaria, si se tiene datos de entrenamiento tales que cada punto de datos o la observación pertenece a una clase específica, el algoritmo SVM se puede entrenar en base a estos datos tales que pueden asignar puntos de datos futuros en una de las dos clases.

Este algoritmo representa las muestras de datos de entrenamiento como puntos en el espacio de manera que los puntos que pertenecen a cualquiera de las clases se pueden separar por un amplio espacio entre ellas, denominado hiperplano, y los nuevos puntos de datos que se pronostican se les asignan clases según el lado de este hiperplano en el que caen.

El algoritmo SVM toma un conjunto de puntos de datos de entrenamiento y construye un hiperplano de un conjunto de hiperplanos para un espacio de características de alta dimensión, cuanto mayores sean los márgenes del hiperplano mejor será la separación, por lo que esto conduce a una menor generalización.

La representación formal y matemática es la siguiente:

Se tiene un conjunto de entrenamiento con n puntos $(x_1, y_1),\ldots, (x_n,y_n)$ donde cada la variable de clase $y \in \{-1,1\}$, donde cada valor indica la clase correspondiente al punto x_i .Cada punto x_i es un vector de características, el objetivo de SVM es encontrar el hiperplano de margen máximo que separa el conjunto de puntos de datos que tienen la etiqueta de clase $y_i = 1$ del conjunto de puntos de datos que tienen la etiqueta $y_i = -1$, de tal manera que se maximice la distancia entre el hiperplano y los puntos de datos de muestra de cualquiera de las clases más cercanas a él. Estos puntos de datos de muestra se conocen como vectores de soporte.

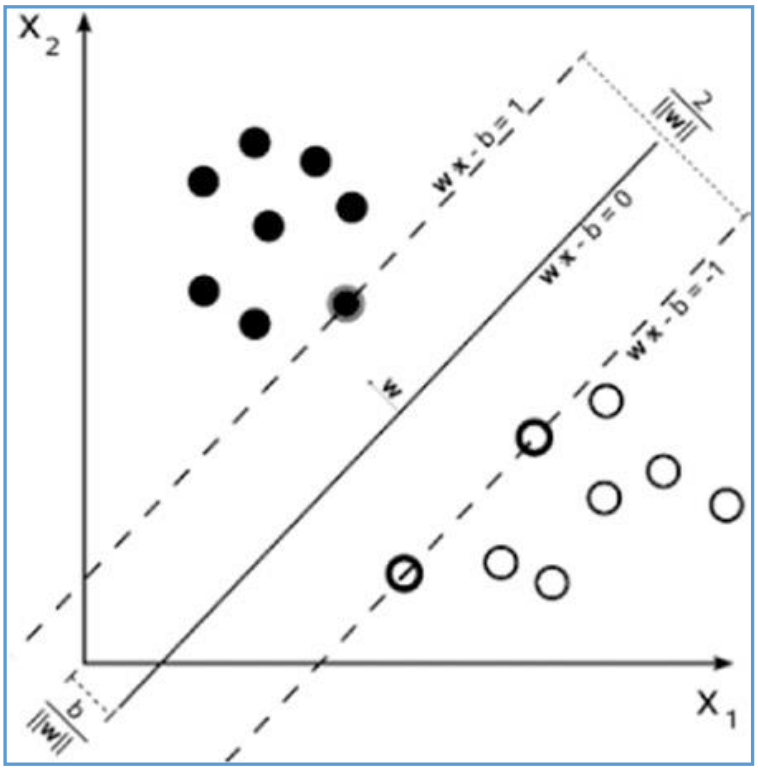

Figura 8: Support Vector Machine

Fuente: Text Analytics – Dipanjan Sarkar

En la figura 8, se pueden observar claramente el hiperplano y los vectores de soporte, el hiperplano se puede definir como el conjunto de puntos x que satisfacen $w * x + b = 0$, donde w es un vector normal al hiperplano, mientras que $\frac{b}{||w||}$ da el desplazamiento del hiperplano desde el origen hacia los vectores de soporte resaltados en la figura. Hay 2 tipos principales de márgenes que ayudan a separar los puntos de datos que pertenecen a las diferentes clases.

Cuando los datos son linealmente separables como en la figura 8, se pueden tener márgenes que estén representados básicamente por 2 hiperplanos paralelos representados por las líneas de puntos que ayudan a separar los puntos de datos que pertenecen a las dos clases diferentes. Esto se hace teniendo en cuenta que la distancia entre ellos es lo más grande posible. La región delimitada por estos dos hiperplanos forma el margen con el hiperplano de margen máximo en el centro. Estos hiperplanos se muestran en la figura que tiene estas ecuaciones $w * x + b = 1$ y $w * x + b = -1$.

Este algoritmo tiene muy buenos resultados para diversas tareas de procesamiento de lenguaje natural (NLP) como por ejemplo clasificación de textos, El algoritmo SVM representa el documento de texto como un vector donde la dimensión es el número de palabras distintas. Si el tamaño del documento es grande, entonces las dimensiones son enormes, si esto ocurre en la clasificación de texto se tendrá un alto coste computacional. Para lograr un mejor desempeño (un incremento entre el 1% a 5%), se deben evaluar en diferentes niveles los parámetros (Aliwy & Ameer, 2017).

3.4.3.2 Naive Bayes

El clasificador Naive Bayes es conocido como un grupo de simples clasificadores probabilísticos, es llamado ingenuo porque calcula las probabilidades condicionales de cada palabra por separado, como si fueran independientes una de otra, en el caso de documentos o conocida como la bolsa de palabras (bag of words) asume que la posición no importa.

Ecuación 1

$$P(x_1, x_2, x_3, \dots\dots\dots, x_n|c) = P(x_1|c) * P(x_2|c) * \dots\dots\dots * P(x_3|c)$$

Base Teórica Naive Bayes

Fuente: Analyticsvidhya – Sunil Ray

En la ecuación 1 se puede observar el principio de independencia, y como efectivamente el orden no importa.

Este modelo se basa en el teorema de Bayes:

Ecuación 2

$$p\,(c|d) = \frac{p(d|c)p(c)}{p(d)}$$

Base Teórica Naive Bayes

Fuente: Analyticsvidhya – Sunil Ray

En la Ecuación 2 se presenta a la base teórica del algoritmo donde d es un documento y c es una clase:

- p(c|d) es la probabilidad a posteriori de la clase (etiqueta) dado un indicador (atributo).
- p(d|c) es la probabilidad dada de la clase de un indicador
- p(c) es la probabilidad anterior de la clase
- p(d) es la probabilidad anterior de un indicador

El cálculo del clasificador Naive Bayes se determina:

Ecuación 3

$$c_{MAP} = argmax\ P(c|d)$$

En la ecuación 3 MAP es el máximo a posteriori es decir la clase más probable

Ecuación 4

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

En la ecuación 4 se observa cómo se aplica la regla de Bayes.

Ecuación 5

$$c_{MAP} = argmax\ P(d|c)P(c)$$

En la ecuación 5 se baja el denominador.

Ecuación 6

$$c_{MAP} = argmax\ P(x_1, x_2, x_3, \dots\dots\dots, x_n|c)\ P(c)$$

En la ecuación 6 se usa el principio se muestra como estaría compuesto el documento por una serie de términos o palabras (bag of word).

Ecuación 7

$$c_{NB} = argmax\ p(c_j) \prod P(x_i|c_j)$$

En la ecuación 7 se observa como la representación el clasificador Naive Bayes cuando se tienen múltiples clases, donde j es el número de clases e i son las características o palabras.

Ecuación 8

$$P(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$P(w_i|c_j) = \frac{count(w_i, c_j)}{\sum count(w, c_j)}$$

En la ecuación 8 se observa la fracción de veces que la palabra w_i aparece entre todas las palabras de los documentos del tema c_j, entonces se tendrá un gran documento para el tema j concatenando todos los documentos de cada tema.

Ecuación 9

$$P(w_i|c) = \frac{count(w_i, c) + 1}{\sum count(w, c) + |V|}$$

En la ecuación 9 se muestra la variante usando Laplace smoothing igual 1.

Para tener más clara la teoría, se tiene el siguiente ejemplo:

Tabla 2

Ejemplo de clasificación Naive Bayes

	Documento	Palabras	Clase
Entrenamiento	1	Chinese Beijing Chinese	C
Entrenamiento	2	Chinese Chinese Shangai	C
Entrenamiento	3	Chinese Macao	C
Entrenamiento	4	Tokyo Japan Chinese	J
Prueba	5	Chinese Chinese Chinese Tokyo Japan	xx

Fuente. Base teórica Naive Bayes
Elaboración: Propia

En la tabla 2 se tienen 5 documentos con las palabras que los componen y la respectiva clase a la que pertenecen, de las cuales 4 serán usados como entrenamiento y la última como prueba.

Ecuación 10

$$P\ (c) = \frac{N_c}{N}$$

En la ecuación 10 se observa cómo se construye la probabilidad de cada una de las clases, para la clase c seria $^3/_4$ mientras que para la clase j es $^1/_4$.
Las probabilidades condicionales usando Laplace smoothing igual a 1, según cada término por clase se dan de la siguiente manera:

P(Chinese|c) = $(5+1)/(8+6) = 6/14 = 3/7$
P(Tokyo|c) = $(0+1)/(8+6) = 1/14$
P(Japan|c) = $(0+1)/(8+6) = 1/14$
P(Chinese|j) = $(1+1)/(3+6) = 2/9$
P(Tokyo|j) = $(1+1)/(3+6) = 2/9$
P(Japan|j) = $(1+1)/(3+6) = 2/9$

Como se puede observan estas son las probabilidades de cada uno de los términos, basados en la regla de Naive Bayes.
Para determinar finalmente a que clase pertenece el documento 5 se tiene:
P(c|d5) = $3/4 * 3/7^3 * 1/14 * 1/14 = 0.0003$
P(j|d5) = $1/4 * 2/9^3 * 2/9 * 2/9 = 0.0001$

Siguiendo el principio del clasificador Naive Bayes, el documento 5 pertenece a la clase c.

Naive Bayes es rápido y fácil para implementar, será utilizado como una línea base en la clasificación de texto, Naive Bayes es lo suficientemente efectivo como para clasificar texto en muchos dominios, aunque es menos preciso que otros.

3.4.4 Buenas practicas para lograr un mejor rendimiento en la clasificacion

El rendimiento de un modelo de clasificación de texto depende en gran medida del tipo de palabras utilizadas en el corpus y del tipo de variables creadas, Shivam Bansal es un Cientifico de Datos de la Universidad Nacional de Singapur y brinda las siguientes recomendaciones para lograr un mejor rendimiento en la clasificacion de texto.

1. **Características específicas del dominio en el corpus**

Para un problema de clasificación, es importante elegir el corpus de prueba y entrenamiento con mucho cuidado. Para que una variedad de características actúen en el algoritmo de clasificación, el conocimiento del dominio juega una parte integral. Por ejemplo, si el problema es "Clasificación de sentimiento para datos de noticias", el corpus debe consistir en datos de fuentes de noticias. Esto se debe a que el vocabulario de un corpus varía con los dominios. Las redes sociales contienen muchos slangs y palabras clave inadecuadas como "awsum, lol, gooood", etc., que están ausentes en cualquiera de los corpus formales como noticias, blogs, etc.

2. **Uso de una adecuada lista de StopWords**

Los StopWords se definen como las palabras más utilizadas en un corpus. Las palabras de parada más utilizadas son “a, the, of, on,... etc”. Estas palabras se utilizan para definir la estructura de una oración. Pero, no sirven para definir el contexto. Tratar este tipo de palabras como palabras principales resultaría en un bajo rendimiento en la clasificación del texto. Estas palabras se pueden ignorar directamente desde el corpus para obtener un mejor rendimiento. Aparte de las palabras clave del lenguaje, también hay otras palabras de apoyo que son de menor importancia que cualquier otro término.

3. **Corpus sin ruido**

En la mayoría de los problemas de ciencia de datos, se recomienda realizar un algoritmo de clasificación en un corpus limpio en lugar de un corpus ruidoso. El corpus ruidoso se refiere a entidades sin importancia del texto, como marcas de puntuación, valores numéricos, enlaces y urls, etc. La eliminación de estas entidades del texto aumentaría la precisión, ya que el tamaño del espacio muestral de las posibles funciones establecido disminuye.

4. Eliminando palabras con frecuencia extremadamente baja

Las palabras clave que aparecen con menor frecuencia en el corpus generalmente no juegan un papel en la clasificación del texto. Uno puede deshacerse de estas características de baja frecuencia, lo que resulta en un mejor rendimiento del modelo. Sí se elige un umbral, todas las palabras clave con menos frecuencia pueden ignorarse, lo que resulta en una buena precisión.

5. Corpus Normalizado

Las palabras son la parte integral de cualquier técnica de clasificación. Sin embargo, estas palabras a menudo se usan con diferentes variaciones en el texto dependiendo de su gramática (verbo, adjetivo, sustantivo, etc.). Siempre es una buena práctica normalizar los términos a sus formas de raíz. Esta técnica es conocida como Lematización.

CAPÍTULO 4: ANÁLISIS DE RESULTADOS

Se debe tener en cuenta que los resultados que se muestran están sobre la base de comentarios que han pasado por el pre-procesamiento es decir eliminación de stop words, eliminación de caracteres sin significado y principalmente la lematización, mientras que para la construcción del modelo se realizó la representación vectorial con el estadístico tf-idf y para la clasificación se dividió la base de datos en base en 2 partes: base de entrenamiento que es la que sirvió para entrenar el modelo esta representa el 80% de la base total, la segunda partición es la base de prueba , a esta partición se le aplica el modelo elaborada con la finalidad de medir los resultados obtenidos (predicciones), esta representa el 20% de la base total.

4.1 Resultados descriptivos

Los resultados que se muestran corresponden a los comentarios ya procesados es decir normalizados, tokenizados y lematizados se han aplicado tanto las técnicas lingüísticas como no lingüísticas que se mencionaron anteriormente.

Primero se muestra una series de resultados de los 3 tipos de cliente según su calificación de recomendación (Promotor, Neutro y Detractor), que permite conocer a cada grupo así como sus similitudes y diferencias.

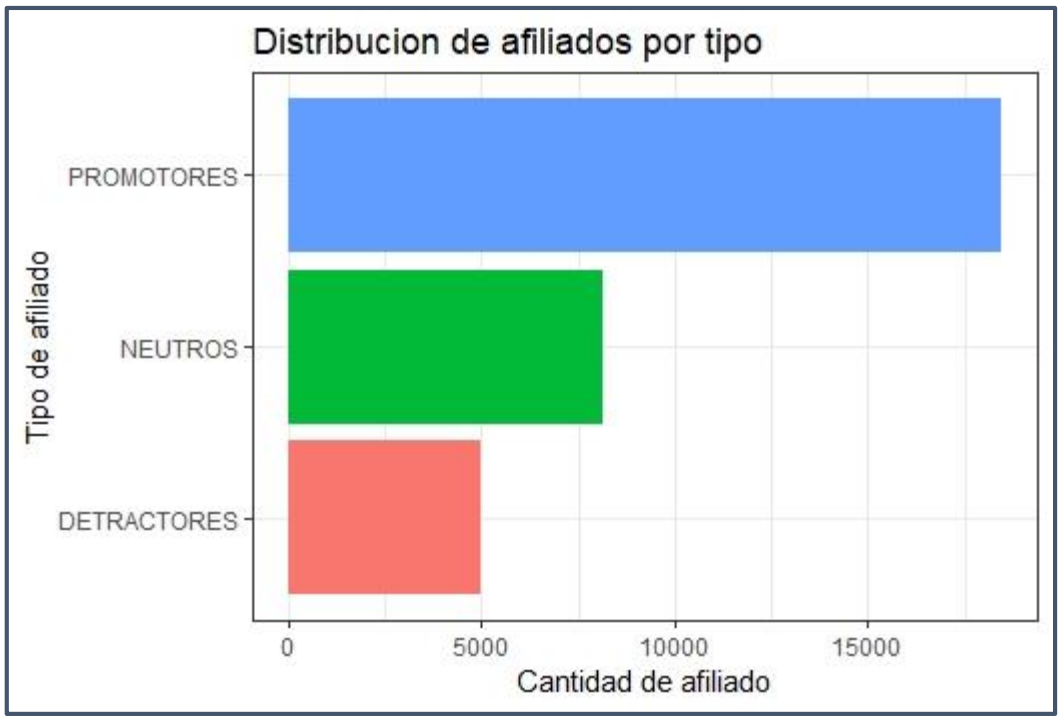

Figura 9: Distribución de clientes por tipo

Elaboración: Propia

En la Figura 9 se puede apreciar que el grupo más grande de clientes durante el periodo analizado corresponde a los promotores, mientras que el grupo de los neutros y detractores son la minoría, de acuerdo al cálculo del NPS se desprende que el resultado es positivo (porcentaje de promotores menos porcentaje de detractores), lo que se requiere para que el indicador del NPS suba es que tanto el grupo de los detractores como el de los neutros disminuya en ese orden.

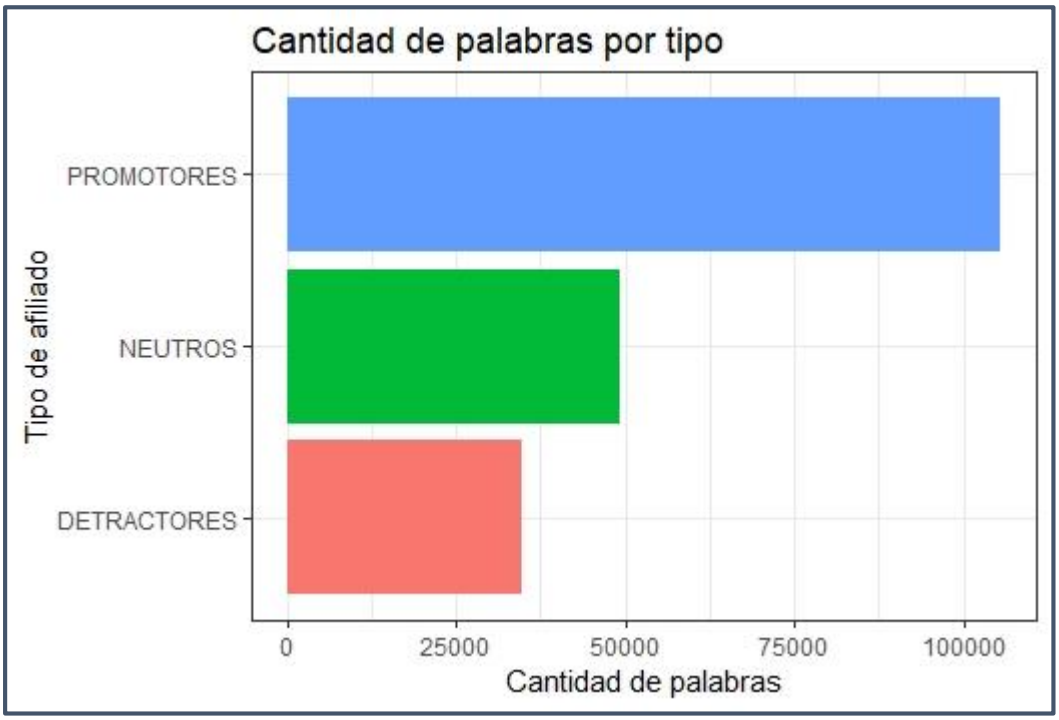

Figura 10: Cantidad de palabras por tipo

Elaboración: Propia

La Figura 10 muestra el número de palabras (después de la limpieza) que intervienen en cada uno de los grupos, se puede apreciar que la diferencia respecto al primer grafico es mucho mayor, mientras que la diferencia entre neutros y detractores se mantiene, como se observó en la Figura 8 el grupo de mayor tamaño corresponde a los promotores es por ello que las palabras son mayores, para poder establecer una adecuada comparación entre el número de palabras por tipo de cliente se mostraran frecuencias relativas.

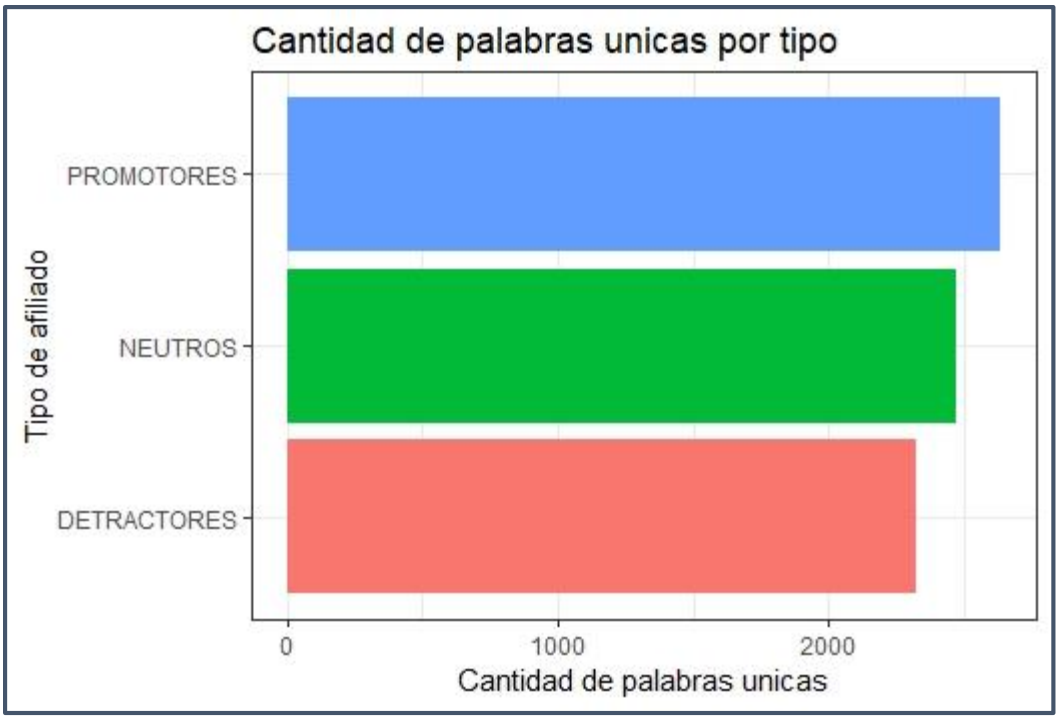

Figura 11: Cantidad de palabras únicas por tipo de cliente

Elaboración: Propia

En la Figura 11 se puede apreciar que el número de palabras distintas usadas en cada grupo es parejo están alrededor de 2500 palabras, tener en cuenta que solo se están considerando las palabras que aportan un significado relevante (limpieza de información), esta diversidad de palabras vendría a ser el diccionario que maneja cada uno de los grupos

Tabla 3

Detalle por Grupo, Cantidad, Palabras y Palabras Únicas

Grupo	Cantidad	Palabras usadas	Palabras únicas usadas
Promotores	18 466	105 579	2 636
Neutros	8 132	49 118	2 471
Detractores	5 005	34 798	2 325

Fuente. Resultados de medición NPS

Elaboración: Propia

La Tabla 3 muestra un resumen del cual se puede concluir que el grupo de los promotores usa un mayor número de palabras repetidas, en otras palabras el diccionario que manejan no es tan diverso, de las más de 100 mil palabras solo hay 2636 palabras diferentes se puede inferir que existen palabras con una gran prevalencia que definen a este grupo, situación que es contraria en cierta medida para el grupo de los detractores y neutros, el mejor ratio lo poseen los detractores con una menor cantidad de repeticiones.

Tabla 4

Longitud por Comentario

Grupo	Longitud promedio
Promotores	5.72
Neutros	6.04
Detractores	6.95

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 4 se puede apreciar que el grupo de los detractores usa más palabras en promedio que el resto de los otros grupos, este primer resultado es interesante ya que aquellos clientes que no recomiendan los servicios de la AFP tienen más que decir frente a los que sí lo recomiendan, se puede inferir que hay múltiples razones por las que se encuentran descontentos y es lo que se quiere analizar.

Tabla 5

Palabras y Frecuencia por Tipo de cliente

Palabra	Frecuencia Promot.	Frecuencia Neutros	Frecuencia Detracto.
Atender	9353	1544	1110
Información	5433	2558	1745
Brindar	3727	755	564
Rápido	2325	-	-
Servicio	1945	-	-
Rentabilidad	1864	1146	615
Cliente	1826	1398	1086
Amable	1794	-	-
Afp	1713	739	502
Cuenta	1539	-	-
Mejorar	-	1983	1423
Deber	1200		794
Comisión	-	669	-
Fondo	-	-	476

Fuente. Resultados de medición NPS
Elaboración: Propia

De la Tabla 5 se puede notar que para el grupo de los promotores las palabras más frecuentes son atender, información y brindar. Lo primero que se puede notar que el diccionario en general que se usa para los 3 grupos tiene similitudes pero en diferentes prioridades pero el que se mas se diferencia es el de los promotores, pero en el caso de los otros 2 grupos si existe una similitud más notoria incluso comparten las top 6 palabras, el desafío será encontrar diferencias en estos grupos

para poder elaborar acciones de mejora ya que si lo que se quiere es incrementar el IRN, se deben identificar las oportunidades para disminuir estos 2 últimos grupos.

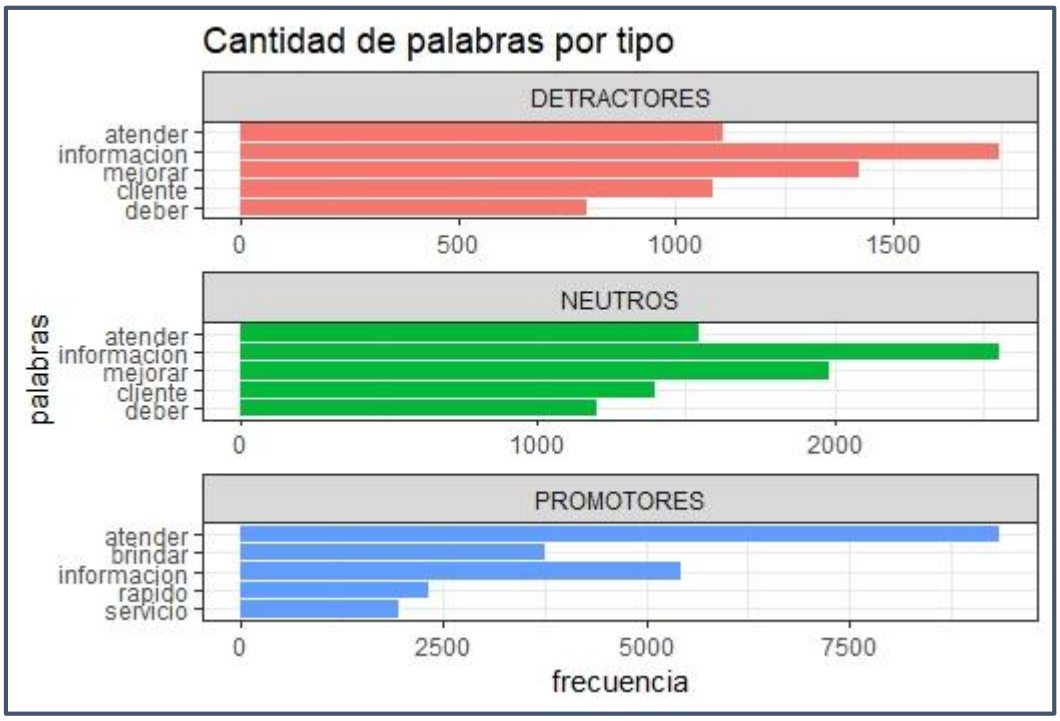

Figura 12: Top palabras por tipo de clientes

Elaboración: Propia

En la Figura 12 se puede observar algunos indicios de similitud y diferencia, por ejemplo se tiene que "información" y "mejorar" son palabras que están presentes en los primeros lugares tanto en el grupo de los clientes neutros con el grupo de clientes detractores, De igual forma las palabras "información" y "atender" están presentes en el grupo de clientes promotores y el grupo de clientes neutros.

Una manera más grafica de representar las palabras más usadas por grupo es usando las nubes de palabras (wordcloud) en text mining es muy frecuente su uso como primera etapa exploratoria:

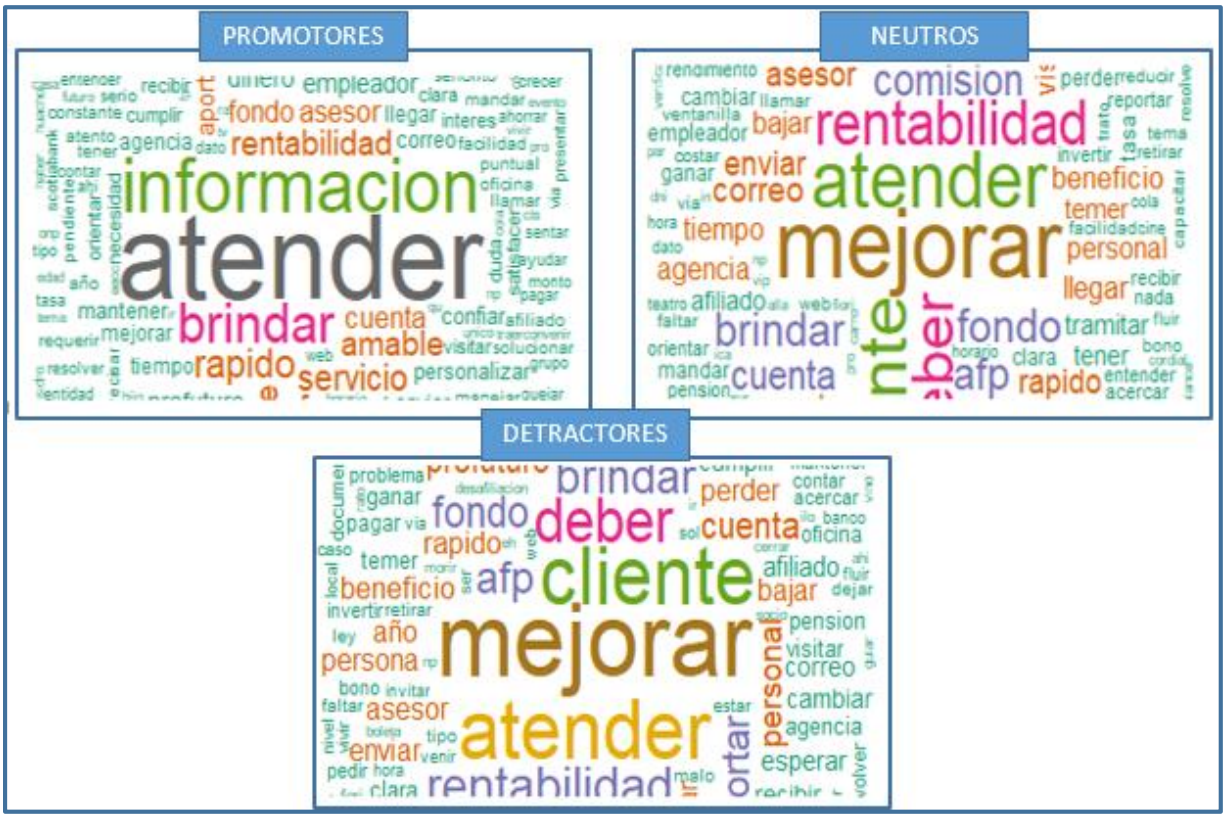

Figura 13: Nube de palabras por tipo de cliente.

Elaboración: Propia

En la Figura 13 se puede apreciar cuales son las palabras usadas para cada grupo, así como la frecuencia de estas representada por el tamaño de las palabras a más grande la palabra más frecuente su uso, se tendrá que medir la similitud o diferencia de cada uno de estos grupos. Ya que como se mencionó anteriormente a primera vista se puede observar que existe cierta similitud entre las palabras de los grupos DETRACTORES y NEUTROS.

Tabla 6

Palabras en Común por Grupo

Grupos	Promotores	Detractores	Neutros
Promotores		1416	1532
Detractores	1416		1439
Neutros	1532	1439	

Fuente. Resultados de medición NPS
Elaboración: Propia

De la Tabla 6 se puede observar que los grupos que comparten una mayor cantidad de palabras son los PROMOTORES y NEUTROS con 1532, teniendo en cuenta que el universo de palabras distintas de estos grupos son 2626 y 2471 respectivamente, la segunda relacion que se pueda observar es NEUTROS con DETRACTORES con 1439 palabras teniendo en cuenta que estos 2 grupos tienen un universo de palabras distintas de 2471 y 2325, en otras palabras se tiene que el 58% del diccionario de los PROMOTORES es compartido con los NEUTROS mientras que el 62% del diccionario de los DETRACTORES es compartido con el grupo de los NEUTROS, se confirma entonces que el diccionario (palabras diferentes) de los 3 grupos es similar, pero se tiene que tener en cuenta bajo que contexto están ya que se está hablando de clientes que están recomendando o no a la administradora.

Así como se tiene identificado el número de palabras comunes entre los grupos, también se tiene que identificar aquellas palabras que son propias de cada grupo (excluyentes), son estas palabras las que permitirán identificar las diferencias de los motivos de recomendación y no recomendación.

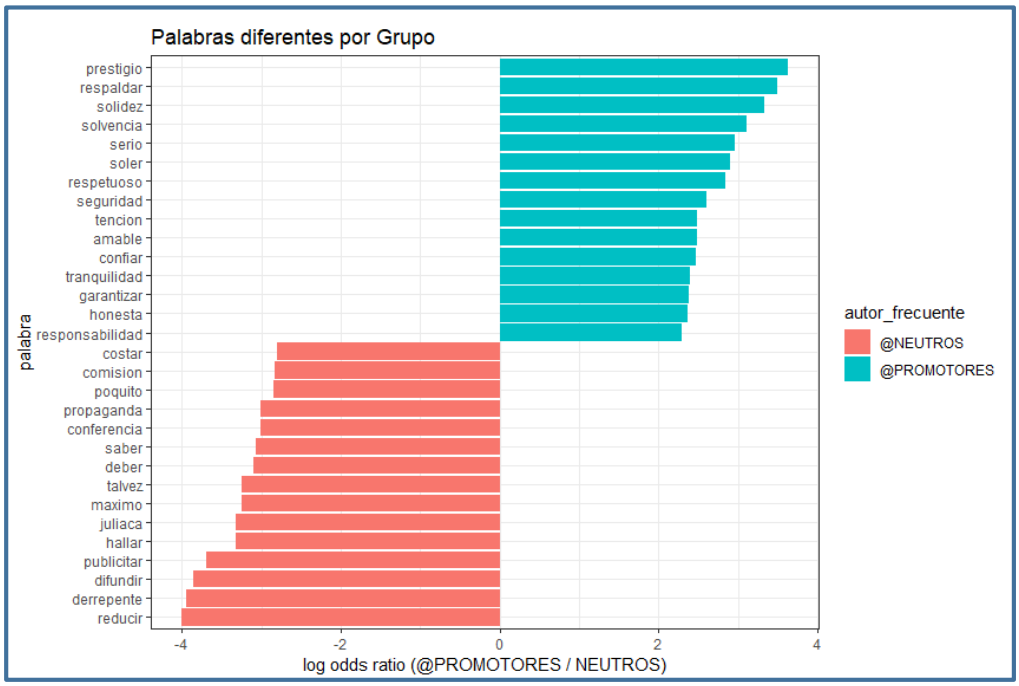

Figura 14: Palabras diferentes entre Neutros y Promotores

Elaboración: Propia

La Figura 14 muestra la comparación entre NEUTROS y PROMOTORES, para el primer grupo se tiene a palabras como reducir, difundir. Para poder realizar un análisis más completo estas palabras deben ser contextualizas ya que por sí solas no será sencillo saber a qué hacen referencia, por ejemplo la palabra reducir hace referencia a la reducción de comisiones (costo que se cobra al cliente por administrar su fondo) es importante saber que las comisiones por AFP no son las mismas y dependiendo del esquema de comisión estas varían, en este caso puntual de la AFP en un esquema de comisión es el más caro, mientras que para difundir se hace referencia a la difusión de información, beneficios y demás que tenga que ver con la AFP ya que el mercado no conoce mucho acerca de los beneficios y derechos que implica estar en el sistema privado de pensiones. Para grupo de los PROMOTORES se tiene prestigio, respaldo y solidez 3 palabras que hablan del prestigio que perciben que tiene la AFP, son probablemente clientes que ya llevan un tiempo importante como clientes, mientras que respaldo y solidez son conceptos percibidos ya que la AFP pertenece a un grupo transnacional que le da la sensación de respaldo al cliente.

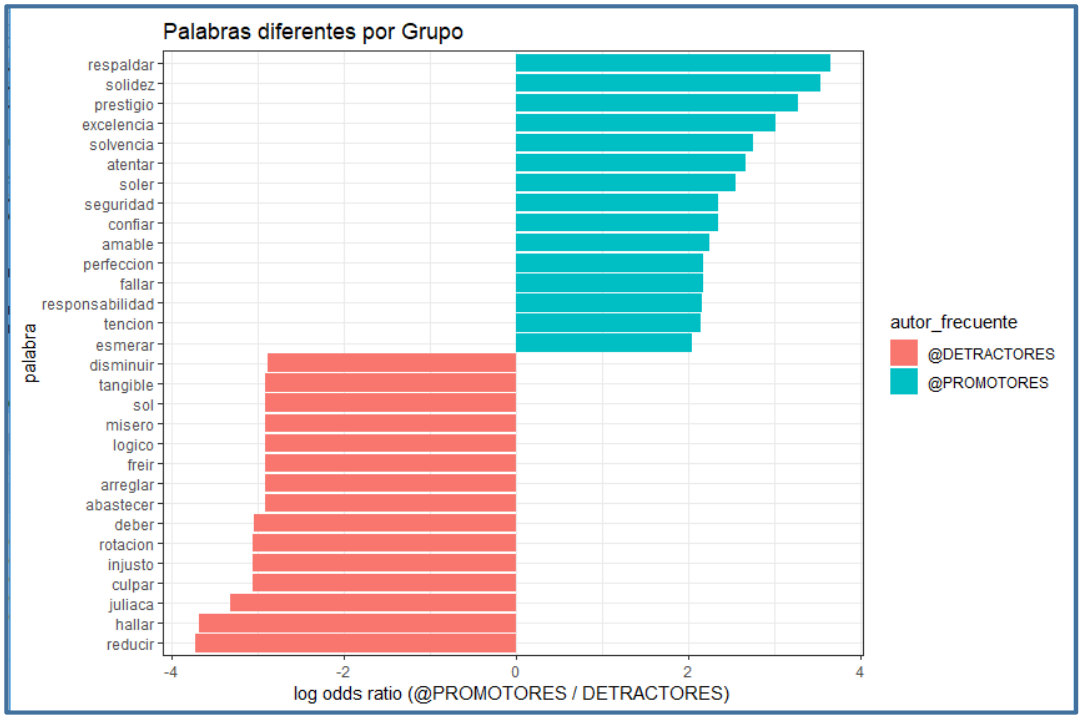

Figura 15: Palabras diferentes entre Detractores y Promotores

Elaboración: Propia

La Figura 15 está comparando a los grupos de PROMOTORES y DETRACTORES, para el primer grupo las palabras se mantienen es decir prestigio, respaldo y solidez cuyo concepto y contexto se explicó en la Figura 13, mientras que para el grupo de los detractores se tiene reducir, culpar e injusto hacen referencia al cobro de la comisión ya que los clientes consideran que esta se debe reducir, consideran que el cobro es injusto y en algunos casos culpan a la AFP por la reducción de su fondo (rentabilidad).

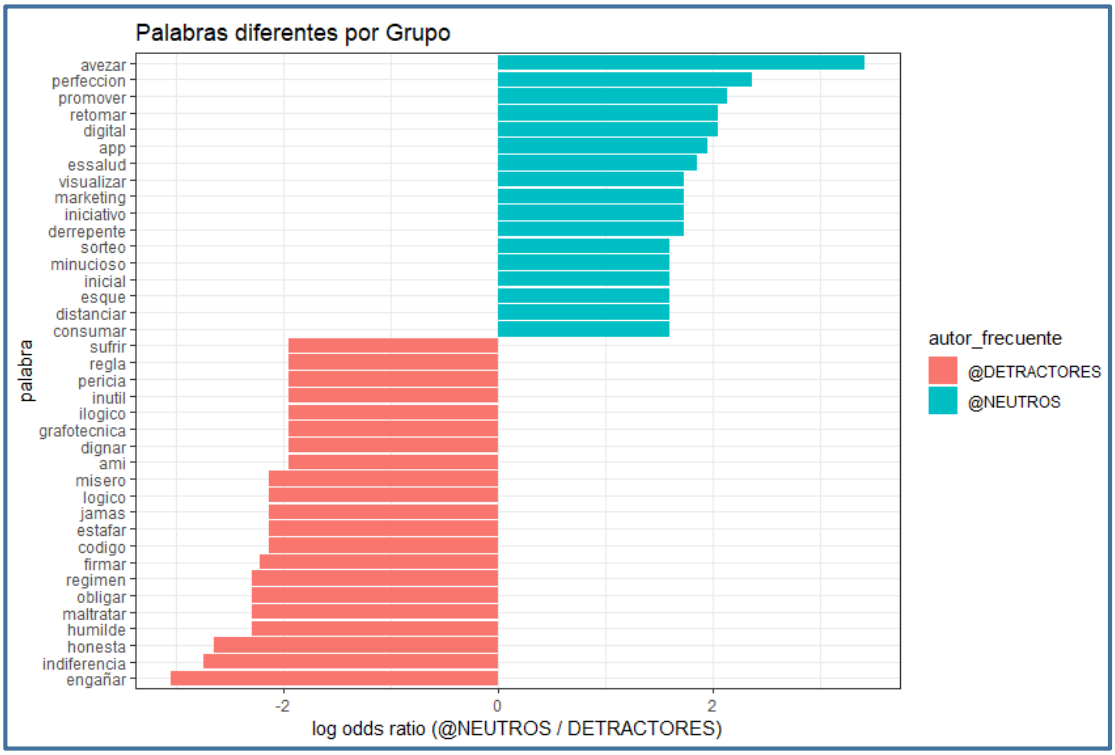

Figura 16: Palabras diferentes entre Detractores y Neutros

Elaboración: Propia

La Figura 16 muestra a los grupos de NEUTROS y DETRACTORES en los que para el primer grupo se tiene perfección, promover y retomar estas palabras hacen referencia a promover información sobre la AFP, en algunos casos los clientes simplemente no califican con las máximas notas debido a que solo lo hacen cuando ven perfección y consideran que no es el caso mientras que para el segundo grupo se tiene engañar, indiferencia y honesta palabras que hacen referencia a que los clientes sienten que se les engaña que la información que brindan no es clara, también se hace referencia a una mala atención y en algunos casos perciben indiferencia durante la atención que realizan.

Hasta el momento se ha realizado el análisis de identificación y diferenciación en base a palabras individuales (bolsa de palabras), pero si agregan niveles adicionales como la semántica y el contexto los resultados que se obtienen más interesantes, en este caso se hará uso de los Ngramas, es decir que se unirán palabras como una sola idea (tener en cuenta que para este análisis también se trabajara con la limpieza de la información), se trabajara con un Ngrama de tamaño 2 es decir bigramas.

Tabla 7

Bigramas de Promotores

Bigrama	Frecuencia
atender rápido	992
brindar información	758
atender brindar	660
mantener información	611
atender amable	550
atender personalizar	530
brindar atender	454
información brindar	454
rápido atender	437

Fuente: Resultados de medición NPS

Elaboración: Propia

La Tabla 7 muestra los bigramas más frecuentes para los PROMOTORES, manejando este concepto se puede ver ideas más claras como atender rápido, brindar información, atender amable.

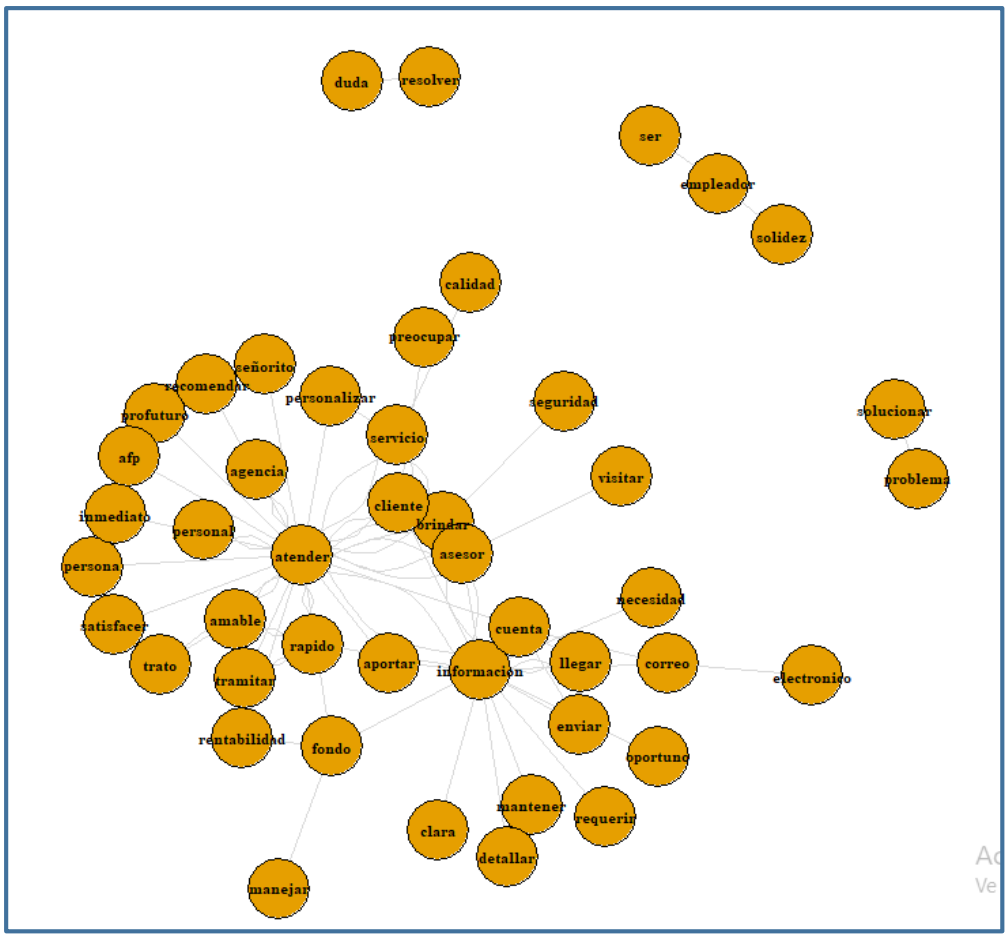

Figura 17: Grafos para promotores

Elaboración: Propia

La figura 17 muestra las relaciones más frecuentes entre las palabras del grupo de los PROMOTORES, se puede notar que las palabras más importantes son atender e información ya que se ve asociadas a ellas muchas más palabras que se forman como bigramas, estas van sujetas a conceptos como atención personalizada, atención amable, atención rápida , información clara, información de los fondos, existen otros conceptos adicionales como solucionar problema a diferencia del primer análisis aquí se le está dando un capa más de análisis como el contexto.

Ahora se mostrará el mismo análisis para los otros grupos, para los NEUTROS:

Tabla 8

Bigramas de Neutros

Bigrama	Frecuencia
mejorar rentabilidad	316
brindar información	304
deber mejorar	213
atender rápido	191
mejorar información	169
mejorar atender	158
bajar comisión	149
atender cliente	144
información cliente	131

Fuente: Resultados de medición NPS

Elaboración: Propia

En la Tabla 8 se observa que una diferencia marcada compara con el grupo de los promotores en cuanto a los bigramas, ahora se manejan conceptos como mejorar la rentabilidad, brindar información, debe mejorar o bajar comisión además de otros más.

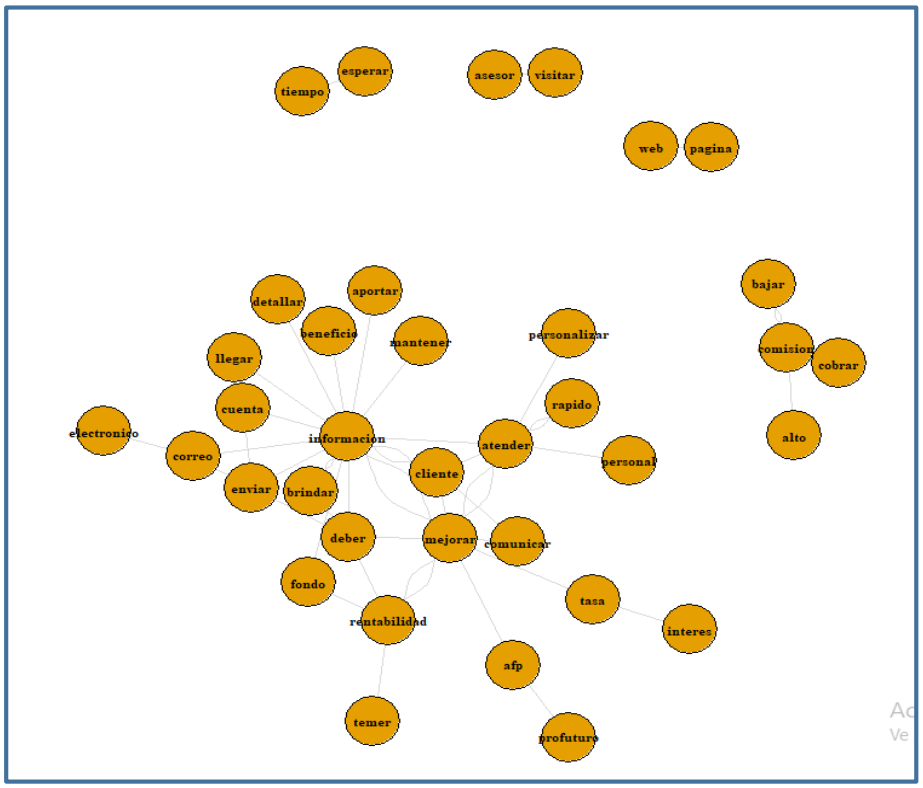

Figura 18: Grafos para Neutros

Elaboración: Propia

En la figura 18 se muestran las palabras más influyentes como información y mejorar, los conceptos asociados son múltiples como informar beneficios, informar al cliente, información aportes. Otros conceptos que se dejan ver son mejorar la rentabilidad, mejorar la comunicación, mejorar la atención y finalmente se ven unos conceptos asociados a comisión como comisión alta, bajar comisión y cobrar comisión.

Tabla 9

Bigramas de Detractores

Bigrama	Frecuencia
brindar información	250
mejorar atender	205
mejorar rentabilidad	172
deber mejorar	145
atender cliente	136
mejorar información	122
atender rápido	90
información cliente	90
bajar comisión	81

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 9 se muestra a los bigramas de mayor frecuencia que son brindar información, mejorar la atención, mejorar rentabilidad, mejorar información, así como atender rápido y bajar comisión.

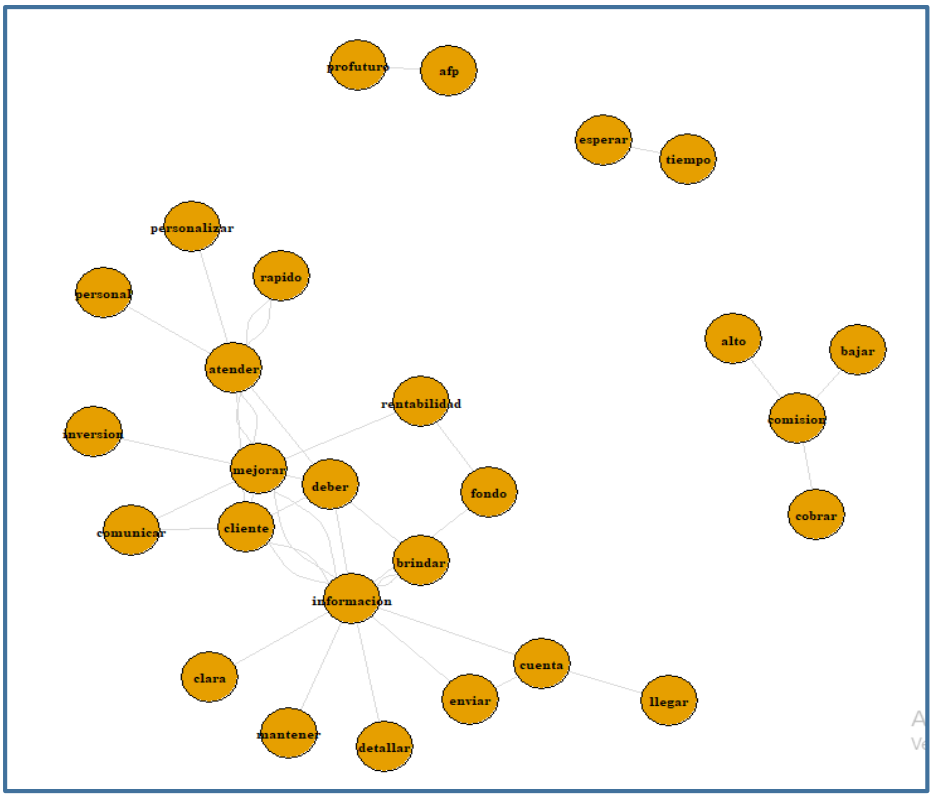

Figura 19: Grafos para detractores

Elaboración: Propia

En la Figura 19 se muestra que para el grupo de los DETRACTORES las palabras más influyentes son mejorar e información y van asociados a conceptos de mejorar la atención, debe mejorar, mejorar la comunicación, mejorar la rentabilidad así como detallar información, información clara, incluso enviar información al igual que el grupo anterior de los neutros existen algunos conceptos asociados a la comisión como bajar la comisión, cobro de comisiones altos.

4.2 Resultados de Predicción

La segunda parte del análisis corresponde a los modelos de clasificación, se van a construir modelos de clasificación donde en función a los comentarios de los clientes se determinara si estos son PROMOTORES, NEUTROS O DETRACTORES.

Los modelos que se desarrollan son SVM y Naive Bayes como modelo base, como resultado final se tendrá la matriz de confusión de cada uno de los modelos, como medida de desempeño de dichos

modelos se usara: Precision, Recall y F1 Score, Ciertamente existe por defecto el indicador Accuracy pero dada la naturaleza de los datos (desbalanceados) este no ya no es un indicador adecuado, mientras que los indicadores ya mencionados se ajustan mejor.

Recall se calcula como el número de veces que el clasificador asigna una etiqueta de una categoría especifica de todas las etiquetas existentes de esa categoría, Precision se calcula como el número de veces que una etiqueta de una categoría especifica se asigna correctamente y finalmente F1 Score es una compensación entre estas 2 últimas.

Tabla 10

Ejemplo tabla de clasificación

	Real	
Predicción	Relevante	Irrelevante
Relevante	A	B
Irrelevante	C	D

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 10 se puede observar un ejemplo de matriz de clasificación donde se tiene los datos reales como los datos predichos, en bases a estos resultados se establecen los indicadores de desempeño.

Ecuación 11

$$recall = \frac{A}{(A + C)}$$

Calculo recall

Fuente: Rdrr – Caret

La Ecuación 11 muestra el cálculo del indicador recall, mide los verdaderos positivos (A) sobre la condición positiva (A y C).

Ecuación 12

$$precision = \frac{A}{(A + B)}$$

Calculo precision

Fuente: Rdrr – Caret

La Ecuación 12 muestra el cálculo del indicador precision, mide los verdaderos positivos (A) sobre la predicción de la condición positiva (A y B).

Ecuación 13

$$F_1 = 2 * \frac{(prec * recall)}{(prec + recall)}$$

Calculo F_1 Score

Fuente: Rdrr – Caret

La Ecuación 13 muestra el cálculo del indicador F1 score, establece una media armónica con precision y recall.

Cuando se habla de clasificación de texto, accuracy no es la forma adecuada para evaluar un rendimiento, se tiene otras métricas diferentes como precision, recall y F_1 score (Panwar, 2019). Para Akinori Fujino, Hideki Isozaki y Jun Suzuki (2008), F_1 score es una métrica adecuada para medir el desempeño de un modelo de clasificación de texto, en su artículo mencionan lo importante de maximizar este indicador de diferentes formas.

Otro punto que se debe considerar en el modelamiento al momento de construir los conjuntos de datos tanto el de entrenamiento como prueba, es que los parámetros propios de los modelos de clasificación trabajados, no necesariamente darán los mismos resultados cuando se ejecuten en múltiples ocasiones. Es por ello que, para obtener mejores resultados en cuanto a la medición de la calidad de la predicción de los modelos, es mejor realizar una validación cruzada sobre los datos de entrada. La validación cruzada o k-fold cross validation consiste en tomar los datos originales y crear a partir de ellos dos conjuntos separados: un primer conjunto de entrenamiento y prueba, y un segundo conjunto de validación.

El conjunto de entrenamiento se va a dividir en k subconjuntos, se toma cada subconjunto como conjunto de prueba del modelo, mientras que el resto de los datos se tomará como conjunto de entrenamiento. Este proceso se repetirá k veces, una vez finalizadas las iteraciones, se calculan los indicadores de cada uno de los modelos construidos y finalmente se calcula el promedio de los indicadores de los k modelos entrenados.

Una vez que se tienen los indicadores promediados para el modelo, se puede repetir entonces el procedimiento de validación cruzada para todos los demás modelos de clasificación que se estén evaluando, se seleccionará aquel que produzca los mejores indicadores (resultados de predicción). Entonces se usa dicho modelo sobre el conjunto de validación generado en la primera parte, es este modelo el que mejor resultado en general ofreció durante la fase de entrenamiento.

4.2.1 **Resultados Naive Bayes**

Los resultados que se muestran corresponden al mejor valor del parámetro de Laplace Smoothing, en este caso es igual a 0 (es decir sin el suavizado de Laplace) ya que mostro una mejor clasificación del grupo de los DETRACTORES, cabe señalar que unos de los pasos previos para el modelamiento fue la proyección de los términos usados en cada uno de los conjuntos ya que si existían palabras que estaban en un conjunto y no en el otro la dimensión de las matrices sería diferente , para evitar ese problema se usó tf-idf usando el argumento dictionary, los resultados obtenidos fueron los siguientes:

Tabla 11

Resultados Matriz de Confusión Naive Bayes

	Real		
Predicción	Detractores	Neutros	Promotores
Detractores	302	233	139
Neutros	419	885	292
Promotores	304	497	3250

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 11 se muestra que como un primer indicador se tiene un Accuracy de 0.7019 (porcentaje de una correcta clasificación), es decir que en general de cada 10 casos clasifica bien 7, se ira revisando los demás indicadores para poder tener un panorama más claro del modelo elaborado.

Tabla 12

Indicadores de Desempeño del Modelo Naive Bayes

	DETRACTORES	NEUTROS	PROMOTORES
Sensitivity	0.295	0.548	0.883
Specificity	0.930	0.849	0.697
Pos Pred Value	0.448	0.555	0.802
Neg Pred Value	0.872	0.846	0.810
Prevalence	0.162	0.256	0.582
Detection Rate	0.048	0.140	0.514
Detection Prevalence	0.107	0.253	0.641
Balanced Accuracy	0.612	0.699	0.790

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 12 lo primero que se puede notar con el indicador Sensitivity es que para el grupo DETRACTORES el valor de una correcta clasificación es regular 30%, teniendo en cuenta que la base que se tiene es desbalanceada y la presencia de este grupo es del 16% aproximadamente como lo marca el indicador Prevalence, mientras que los grupos NEUTROS y PROMOTORES presentan un mejor valor de una correcta clasificación 55% y 88% respectivamente, adicional a ello se tiene el indicador Balanced Accuracy que es un indicador más adecuado de cara a la naturaleza desbalanceada de la base de datos este tiene en cuenta tanto la clasificación de la categoría positiva como la categoría negativa.

Los indicadores que se han descrito en el párrafo anterior permiten tener una referencia sobre desempeño del modelo pero como se había mencionado anteriormente para poder hacer la elección entre uno u otro modelo se ha realizó un paso adicional que es la validación cruzada y además de ellos se hará uso de los indicadores de desempeño como recall, precisión y F1 score.

Tabla 13

Indicadores de Desempeño del Modelo Naive Bayes

	DETRACTORES	NEUTROS	PROMOTORES
Recall	0.2717	0.5776	0.8841
Precision	0.4468	0.5676	0.8045
F1 score	0.3478	0.5725	0.8662

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 13 se muestran los resultados para el modelo Naive Bayes usando validación cruzada con un valor de k = 10 subconjuntos, además del parámetro Laplace smoothing igual a 0, se puede apreciar que el grupo de DETRACTORES no presenta un nivel de predicción aceptable pero se debe considerar que la prevalencia de este grupo naturalmente es baja, mientras que para el grupo de NEUTROS la predicción mejora y finalmente el grupo de PROMOTORES presenta mejores resultados de clasificación.

4.2.2 Resultados SVM

Los resultados que se mostraran corresponden al parámetro "cost" igual a 1.

Tabla 14

Matriz de Confusión SVM

	Real		
Predicción	Detractores	Neutros	Promotores
Detractores	93	58	32
Neutros	534	992	232
Promotores	398	565	3417

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 14 se observa un Accuracy de 0.7122, es decir que en general de cada 10 casos clasifica bien 7, se ira revisando los demás indicadores para poder tener un panorama más claro del modelo elaborado.

Tabla 15

Indicadores de Desempeño del Modelo SVM

	DETRACTORES	**NEUTROS**	**PROMOTORES**
Sensitivity	0.091	0.614	0.928
Specificity	0.983	0.837	0.635
Pos Pred Value	0.508	0.564	0.780
Neg Pred Value	0.848	0.864	0.864
Prevalence	0.162	0.256	0.582
Detection Rate	0.015	0.157	0.541
Detection Prevalence	0.029	0.278	0.693
Balanced Accuracy	0.537	0.726	0.782

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 15 se puede notar que el indicador Sensitivity (el valor de una correcta clasificación) para el grupo DETRACTORES es mala ya que solo tiene un 9%, teniendo en cuenta que la base que se tiene es desbalanceada y la presencia de este grupo es del 16% aproximadamente como lo marca el indicador Prevalence, mientras que los grupos NEUTROS y PROMOTORES presentan un mejor valor de una correcta clasificación 61% y 93% respectivamente, adicional a ello se tiene el indicador Balanced Accuracy que es un indicador más adecuado de cara a la naturaleza desbalanceada de la base de datos este tiene en cuenta tanto la clasificación de la categoría positiva como la categoría negativa.

Los indicadores que se han descrito en el párrafo anterior permiten tener una referencia sobre desempeño del modelo pero como se había mencionado anteriormente para este trabajo de tesis los indicadores de desempeño que se utilizarán son recall, precisión y F1 score.

Tabla 16

Niveles de Error por Parámetro	
Cost	Error
0.1	0.3148493
0.5	0.2863706
1	0.2847490
2.5	0.2857774
5	0.2861729

Fuente. Resultados de medición NPS

Elaboración: Propia

En la Tabla 16 se muestra los distintos niveles de error según el parámetro “Cost”, se observa que el parámetro más adecuado es que toma el valor de 1, cabe mencionar que para llegar a esta valor se realizó una validación cruzada con un valor de k = 10 subconjuntos (al igual que en Naive Bayes).

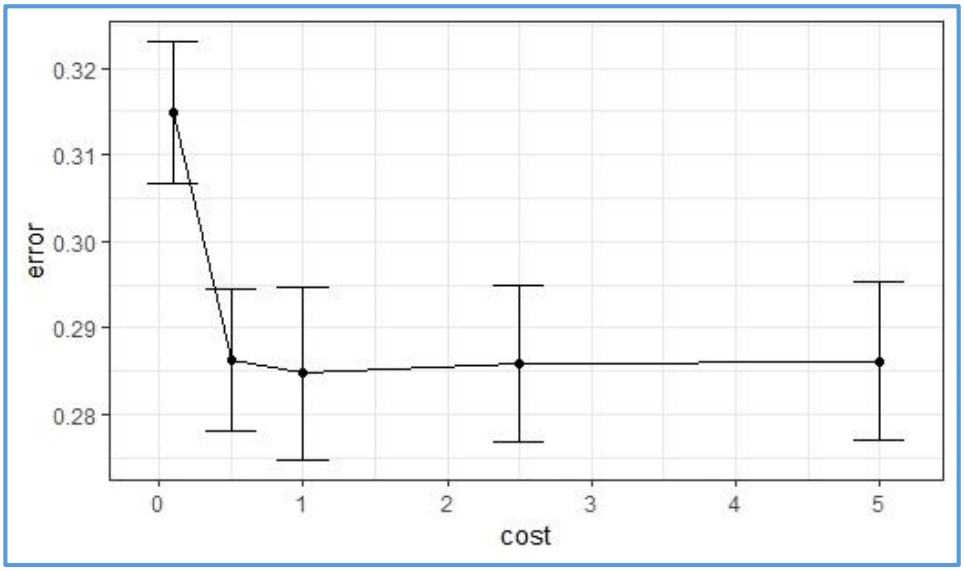

Figura 20: Parámetro cost SVM

Elaboración: Propia

En la Figura 20 se confirma que el valor a escoger es 1, se puede apreciar que el valor que peor desempeño muestra es el valor cost = 0, los demás valores arrojan errores similares.

Tabla 17

Indicadores de Desempeño del Modelo SVM

	DETRACTORES	NEUTROS	PROMOTORES
Recall	0.0884	0.6030	0.9242
Precision	0.4780	0.5597	0.7812
F1 score	0.1492	0.5805	0.8467

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 17 se muestran los resultados para el modelo SVM usando validación cruzada con un valor de k = 10 subconjuntos, además del parámetro Cost igual a 1, se puede apreciar que el grupo de DETRACTORES presenta un nivel de predicción bajo, mientras que para el grupo de NEUTROS la predicción mejora y finalmente el grupo de PROMOTORES presenta mejores resultados de clasificación

4.2.3 Comparación de resultados

Los resultados que se van a comparar corresponden a las mejores versiones halladas para los modelos SVM y Naive Bayes, usando los mejores parámetros donde para el modelo Naive Bayes el parámetro laplace tomo el valor de 0 mientras que para el modelo SVM el parámetro cost tomo el valor de 1, se debe tener en cuenta además que se aplicó validación cruzada (k=10 en todos los casos), los indicadores que servirán para la comparación serán recall, precisión y F1 score para poder elegir el modelo que arroja mejores resultados de predicción:

Tabla 18

Comparación de Resultados Entre Modelos Naive Bayes y SVM

	Recall			Precision			F1 score		
	Det	Neu	Prom	Det	Neu	Prom	Det	Deu	Prom
SVM	0.088	0.603	0.924	0.478	0.559	0.781	0.149	0.580	0.846
Naive Bayes	0.271	0.577	0.884	0.446	0.567	0.804	0.347	0.572	0.866

Fuente. Resultados de medición NPS
Elaboración: Propia

En la Tabla 18 se muestra una tabla donde se comparan los modelos propuestos, teniendo en cuenta que el modelo que serviría como base en este caso Naive Bayes arroja resultados similares de predicción a los grupos NEUTROS y PROMOTORES comparándolo con SVM, pero donde si destaca es en el grupo de los DETRACTORES otorgando un 20% más, motivo por el cual el termina siendo el modelo a escoger, aunque esto no signifique que no hay margen para mejorar ya sea con otros alternativas de modelos, manejo del desbalanceo y/o técnicas adicionales para el procesamiento de lenguaje natural.

CAPÍTULO 5: CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

Luego de haber realizado las etapas señaladas en el trabajo de tesis tanto la etapa descriptiva como la etapa de modelado (clasificación), es necesario recordar los objetivos que se habían señalado tanto el general como los específicos.

Como objetivo principal se tenía que identificar a través de los comentarios las diferencias de no recomendación entre los grupos de clientes, en la etapa descriptiva se pudo notar que si existen diferencias pero se evidencia que también existe similitud entre los grupos, como se señaló en un inicio la diversificación de palabras o el universo para los 3 grupos no es tan amplio y esto en cierta medida es bueno ya que la propuestas no serán tan amplias o específicas y hace más factible tener mejores resultados.

Las conclusiones son las siguientes:

1. En el grupo de los DETRACTORES existe un mayor uso de términos por comentario para ser más exactos en cada comentario recibido se usa alrededor de 7 términos (después de la depuración de termino sin significado), pero estos comparten más del 50% de palabras en común con el grupo de los NEUTROS, mientras que las diferencias que se identificaron

frente al grupo de NEUTROS son: Clientes que por razones de desinformación sobre el Sistema Previsional Peruano no conocen cuáles son sus derechos y/o beneficios y perciben que son engañados, estafados y hasta maltratados, lo que en otras palabras es difundir una mayor información, es este concepto el más importante de todos ya que en torno a ello giran otros conceptos.

2. El grupo de los NEUTROS maneja un número de términos similar al de los DETRACTORES, en este caso por cada comentario recibido se usa alrededor de 6 términos (después de la depuración de términos sin significado), entre las diferencias que se identificaron frente al grupo de DETRACTORES están: Clientes que simplemente por temas como que no suelen recomendar un servicio a menos que sea perfecto, también se encuentran conceptos como retomar ciertos beneficios de valor agregado que se dejaron de brindar además de la necesidad de contar con un canal digital que facilite sus requerimientos.

Estas fueron algunas de las diferencias halladas que responden al análisis realizado palabra por palabra, ahora se repasaran las diferencias usando bigramas (usando 2 términos):

1. Para el grupo de los NEUTROS se encontraron conceptos como informar beneficios, informar al cliente, información aportes. Otro concepto que se deja ver es el de mejorar como por ejemplo mejorar la rentabilidad, mejorar la información, mejorar la atención y finalmente se ven unos conceptos asociados a comisión como bajar comisión y cobrar una menor comisión mientras que para el grupo de los DETRACTORES se tiene mejorar la atención al cliente, además de reducir los tiempos de espera, brindar información clara, cobrar menos comisiones así como mejorar inversiones, etc.

2. En conclusión para el objetivo principal se han identificado y señalado las diferencias de los grupos, sin antes mencionar que cuentan también con semejanzas a tener en cuenta, lo que lleva entonces a responder la hipótesis principal del trabajo que es confirmar que se han identificado las diferencias entre las razones de recomendación entre los grupos NEUTROS y DETRACTORES.

En cuanto a los objetivos específicos de cara a la implementación de un modelo de clasificación de texto cabe señalar que a lo largo del desarrollo del trabajo de tesis se ha abordado cada uno de los pasos que comprenden el proceso para su elaboración hasta llegar finalmente a elaborar un modelo de clasificación, pasando desde la extracción, limpieza , el procesamiento hasta el modelado, en cada uno de los pasos se han encontrado temas interesantes que se han ido desarrollando y que también se pueden perfeccionar, por ejemplo en el procesamiento la identificación n-gramas mostró conceptos a tener cuenta y en la etapa del modelado se tiene abierto el abanico de aplicaciones de diferentes algoritmos así como enfoques, en el caso de la hipótesis especifica formulada donde se planteaba que el uso de un modelo de minería de texto ayudaría a identificar las diferencias entre los grupos, se puede decir:

1. Para elaborar un modelo se tiene que pasar por todo un esquema de trabajo (metodología) es ahí en las diferentes etapas donde se pudieron evidenciar las diferencias, para finalmente elaborar el modelo, en otras palabras el hecho de elaborar el modelo permitió encontrar las primeras diferencias en los pasos previos.

2. Ya en la elaboración del modelo en sí se ha podido brindar porcentajes de predicción que permiten poder clasificar a los clientes según los comentarios que brindan durante la encuesta que se les realiza, en este caso el modelo que dio mejores resultados fue Naive Bayes.

Finalmente ya de cara a la gestión de los resultados, se tenía como objetivo identificar acciones clave que permita aumentar el índice de recomendación es decir aumentar PROMOTORES o reducir a los grupos NEUTROS y DETRACTORES, en este caso si se han podido identificar acciones que si se encuentran en manos de la AFP y existan algunas que no:

1. En manos de la AFP si esta mejorar la comunicación hacia los clientes, se debe elaborar un mejor programa de comunicación, mejorar los tiempos de atención, mejorar la rentabilidad de los fondos aunque esta última es compartida ya que dependiendo de los indicadores macroeconómicos y/o factores externos esta pudiera disminuir o bajar, otra acción que ayudaría seria mejorar el trato al cliente pero antes de ello se debe evaluar de manera dicha atención que actualmente se mide pero no se incide sobre ello.

2. Existe un concepto no menos importante que pero por temas de regulación del mercado no está en manos directamente de la AFP y es la reducción de las comisiones, no es algo que pueda realizar como en un banco, pero claro esto no tiene que ser conocido por el cliente o quizás sí, pero lo que si se debe dar a conocer son las comisiones de las todas las AFP ya que para ojos del cliente creen en muchos casos que la AFP de estudio es la más cara cuando no es así realmente.

3. Estas acciones que parecen sencillas son muy importantes y si se dan mejorarían el índice de recomendación pero implican en cambio en la forma de trabajar de la AFP, se deben estructuras planes de mejora en muchas áreas para poder lograr estas acciones con lo que sí se puede afirmar la otra hipótesis de que si ayuda en la medición de calidad contar con una gestión adecuada de los clientes DETRACTORES y NEUTROS.

5.2 Recomendaciones

5.2.1 Captura de información

Algo que se evidencio durante el trabajo de limpieza de datos, es que en muchos de los casos los comentarios capturados no eran buenos, existían problemas de redacción, ortografía y semántica y hasta comentarios repetidos, lo que tuvo que ser corregido y como se comentó en un inicio es esta etapa del trabajo que tomo más tiempo, pero también fue enriquecedora para conocer mucho más el contexto, una de las recomendaciones que se dan es mejorar el proceso de captura de información, que esta sea lo más fiel al pensamiento del cliente, se pueden manejar algunas alternativas que pudieran ser costosas como por ejemplo las grabaciones de las llamadas para su posterior transcripción o más baratas como mediciones mediante correo electrónico donde le daría mucho más libertad al cliente, cabe mencionar actualmente que en algunas empresas esta modalidad ya se viene aplicando.

5.2.2 Modelamiento

Como se mencionó en etapas anteriores, la unidad medida que uso para el modelamiento fueron tokenizados palabra por palabra, pero en la etapa descriptiva se observaron temas interesantes usando bigramas, lo que deja como trabajo futuro el modelar desde un enfoque de n-gramas, donde

se podrían o no ver mejores resultados, algo que también se señalo es que se estaba antes una base desbalanceada y antes ello se usaron únicamente indicadores de rendimiento ajustado que permiten evaluar de manera eficiente a los modelos presentes, queda también como trabajo futuro introducir métodos que se usen justamente en estos como oversampling, undersumpling o algún otro método que en lo particular exploraría métodos alternativos sin la necesidad de omitir información o replicarla ya que la base procesada es lo que se ajusta a la realidad de la AFP, una medida que pudiera funcionar en este caso para efectos de poder mejorar modelo clasificador estaría en manejar 2 clases PROMOTORES y NO PROMOTORES ya que finalmente lo que se busca es incrementar el índice de recomendación y es justamente que para poder hacerlo se debe reducir este último grupo, y viendo los resultados se pudieran mejorar manejando estas 2 clases.

5.2.3 Gestión

Finalmente de cara a la gestión y a la mejora del índice de recomendación, los hallazgos encontrados no deberían quedarse en este documento, lo que tocaría debería ser poder en ejecución las recomendaciones e involucrar a todas a las áreas ya que como menciono inicialmente la pregunta es ¿Usted recomendaría a la administradora de fondo de pensiones?, se debe tener presente y más en la AFP que esta pregunta es una evaluación a toda la compañía y tocan hacer cambios solo por mencionar uno, en el caso de mejorar la comunicación se habla de campañas de marketing, de envíos masivo por correo electrónico, cambios en los estados de cuenta, presencia en redes sociales y muchas más acciones que impliquen ello, queda como trabajo futuro poner en producción el modelo y ponerlo a trabajar para empezar en las redes sociales ya que es ahí donde se tiene de primera mano los comentarios del cliente, y se podrán identificar fácilmente sus motivos y saber si este es un DETRACTOR, NEUTROS o PROMOTOR.

Pero se debe tener en cuenta que la industria del sistema privado de pensiones se ve influenciada en algunos casos por los cambios en la legislación peruana, es por ello que estos hallazgos no son perpetuos se deben ajustar periódicamente y el montar este modelo en los diferentes canales de atención empezando por los digitales permitirá poder tomar temperatura de lo que sucede de manera más oportuna y eficiente lo que se traduce en tomar acciones de manera más rápida.

REFERENCIAS BILIOGRAFICAS

Referencias bibliográficas

Aggarwal. (2014). Frequent Pattern Mining.

Alarcon, C. N. (2014). Indicadores clave de gestion sobre la experiencia del cliente: un estudio basado en fuzzy text mining. Barcelona, España.

Aliwy, A. H., & Ameer, E. H. (2017). Comparative Study of Five Text Classification Algorithms with their. International Journal of Applied Engineering Research.

Arcila Calderon, C., Barboza, E., & Cabezuelo, F. (2012). Tecnicas de Big Data:Analisis de texto a gran escala para la investigacion cientifica y periodistica.

Bansal, S. (29 de Octubre de 2015). Analytics Vidhya. Obtenido de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/10/6-practices-enhance-performance-text-classification-model/

Belkahla, W., & Triki, A. (2011). Customer knowledge enabled innovation capability: proposing a measurement scale. Tunisia.

Berzal, F. (2010). Introduccion al data mining.

Berzal, F. (s.f.). Patrones secuenciales. Granada.

Brill, E. (1994). A Simple Rule-Based Part of Speech Tagger . Pennsylvania.

Broß, J. (2013). Aspect-Oriented Sentiment Analysis. Berlin, Alemania.

Caemmerer, B., & Wilson, A. (2010). Customer feedback mechanisms and organisational learning in service operations. Glasgow.

Chong Ho, Y., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. The Qualitative Report, 730-744.

Confirmit everywhere. (s.f.). The power of text Analytics in customer experience programs. Horizons.

Crijns, T. (2016). Classifing events to ugenda calendar genres. Holanda.

Cruzado, J. G. (s.f.). Analisis de sentimientos para la toma de decisiones.

Dipanjan, S. (2016). Text Analytics with python. apress.

EFQM. (2013). Net Promotor Score. Bruselas.

El-Beltagy, D. S. (2017). RPubs. Obtenido de RPubs: https://rpubs.com/moka92/demo3

Feldman, R., Fresko, M., Hirsh, H., Auman, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. PAKM.

Fujino, A., Isozaki, H., & Suzuki, J. (2008). Multi-label Text Categorization with Model Combination. NTT Communication Science Laboratories.

Grimes, S. (30 de Octubre de 2007). http://www.b-eye-network.com. Obtenido de http://www.b-eye-network.com/view/6311

Gutierrez, J. M. (s.f.). Data Mining extraccion del conocimiento en las bases de los datos. Santander.

Hernandez Sampieri, R., Fernandez Collado, C., & Baptista Lucio, M. (2010). Metodologia de la investigacion. Mexico: McGrawHill.

Luhn, H. P. (1958). A business intelligence system. IBM Journal of Research and Development.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management. Spotlight, 60-70.

Mykroyannidis, A., & Theodoulidis, B. (2009). Ontology managment and evolution for business Intelligence. International Journal of Information Management, 559-566.

Nguifo, V. -F.-N. (2012). CMRules: Mining Sequential Rules Common to Several Sequences. Montreal.

Panwar, N. (2019). Text Classification using machine learning with code. Delhi.

Perez, A., Cardoso, A., & Bini, A. (s.f.). Mineria de textos: sistema de busqueda de respuestas.

Pribil, J. (2014). Text Mining in Business Practice. San Francisco.

Profuturo AFP. (2017). Resultados IRN. Lima.

Raja, V. K., Sukhwani, S., & Khots, D. (2017). Enhancing Customer Experience through Text Analysis of Survey Comments. Dallas.

Robles, F. (2015). lifeder. Obtenido de lifeder: https://www.lifeder.com/tipos-investigacion-cientifica/

Rodrigo, J. A. (Diciembre de 2017). RPubs. Obtenido de RPubs: https://rpubs.com/Joaquin_AR/334526

Rojas, B. J. (s.f.). Reglas de asociacion y secuencias. Costa Rica.

Ronen Feldman, Y. A.-Y. (2003). Determining trends using text mining.

SATMETRIX. (2014). NET PROMOTER CONSUMER BENCHMARKS.

Sebastiani., F. (2002). Machine learning in automated text categorization. ACM computing surveys.

Segarra, D. (15 de Febrero de 2015). DanielSegarra. Obtenido de DanielSegarra: www.danielsegarra.com/enps-nps-del-empleado-importancia

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Canada.

Vega, J. B. (Abril de 2018). Planetachatbot. Obtenido de Planetachatbot: https://planetachatbot.com/na%C3%AFve-bayes-con-r-para-clasificacion-de-texto-56ef90cc3aed

Villaroel, F., Burton, J., Theodoulidis, B., Gruber, T., & Zaki, M. (2014). Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach.

Villena Roman, J., Martinez Camara, E., Garcia Morera, J., & Jimenez Zafra, M. (2014). Aspect-Oriented Sentiment Analysis. Barcelona, España.

Witell, L. (2011). Idea generation: customer co-creation versus traditional market research techniques. Karlstad.

ANEXOS

ANEXO 1: Declaración de Autenticidad

Anexo 2: Autorización de consentimiento para realizar la investigación

Universidad Ricardo Palma

Escuela de Posgrado

AUTORIZACIÓN DE CONSENTIMIENTO PARA REALIZAR LA INVESTIGACIÓN

DECLARACIÓN DEL RESPONSABLE DEL AREA O DEPENDENCIA DONDE SE REALIZARA LA INVESTIGACIÓN

Dejo constancia que el área o dependencia que dirijo, ha tomado conocimiento del proyecto de tesis titulado:

Un Modelo Predictivo del Indice NPS Basado en Información Textual de Percepción del Servicio al Cliente para Identificar Motivos de No Recomendación

el mismo que es realizado por el Sr./Srta. Estudiante (Apellidos y nombres):

TRISTAN GOMEZ LUDGARDO EDER

, en condición de estudiante - investigador del Programa de:

MAESTRIA EN CIENCIA DE LOS DATOS

Así mismo señalamos, que según nuestra normativa interna procederemos con el apoyo al desarrollo del proyecto de investigación, dando las facilidades del caso para aplicación de los instrumentos de recolección de datos.

En razón de lo expresado doy mi consentimiento para el uso de la información y/o la aplicación de los instrumentos de recolección de datos:

Nombre de la empresa:	Autorización para el uso del nombre de la Empresa en el Informe Final	SI / NO
PROFUTURO AFP		NO

Apellidos y Nombres del Jefe/Responsable del área:	Cargo del Jefe/Responsable del área:
OSORES ARAGONEZ WALTER	GERENTE DE PLANEAMIENTO Y CONTROL DE GESTIÓN

Teléfono fijo (incluyendo anexo) y/o celular:	Correo electrónico de la empresa:
215 2800 – 4280	WOSORES@PROFUTURO.COM.PE

Firma

27/03/2019

Fecha

Anexo 3: Matriz de consistencia

PROBLEMA	OBJETIVO	HIPOTESIS	VARIABLE	DIMENSIÓN
¿En qué se diferencian las razones por las que los clientes detractores y neutros no recomiendan una administradora de fondo de pensiones privada en el mercado peruano?	GENERAL Identificar a través de la retroalimentación de los clientes detractores y neutros las diferencias de sus razones de no recomendación usando técnicas lingüísticas y no lingüísticas de minería de texto.	GENERAL Existen diferencias en las razones por las que los clientes detractores y neutros no recomiendan una administradora de fondo de pensiones en el mercado peruano.	INDEPENDIENTE Motivos (razones de no recomendación) .	Indicadores de desempeño del modelo.
ESPECIFICOS 1. ¿Cómo implementar y desarrollar un modelo de clasificación de la información textual de los clientes que participan de la medición de calidad? 2. ¿Qué descubrimientos y acciones clave se identifican en los procesos y protocolos de atención con la finalidad de mejorar el índice de recomendación?	ESPECIFICOS 1. Implementar un modelo de clasificación de texto que permita predecir el tipo de cliente según sus comentarios. 2. Identificar acciones clave que permitan redefinir y mejorar procesos, protocolos de atención, etc. con la finalidad de atacar a los grupo de clientes NEUTROS y DETRACTORES.	ESPECIFICOS 1. Ayuda en la clasificación de comentarios el uso de un modelo de minería de texto 2. Ayuda en la medición de calidad la gestión adecuada de los clientes detractores y neutros, en base a las acciones clave identificadas.		

Anexo 4: Código de Procesamiento

```
##############################Carga Inicial####################

library(stringi)
library(wordcloud)
library(wordcloud2)
library(tm)
txt <- readLines("E:/TMT/neu_prue2.txt",encoding="UTF-8")
txt = iconv(txt, to="ASCII//TRANSLIT")
txt = text_tokens(txt, stemmer = stem_liste)
corpus <- Corpus(VectorSource(txt))
d  <- tm_map(corpus, content_transformer(tolower))
d  <- tm_map(d, stripWhitespace)
d  <- tm_map(d, removeWords,"c")
d <- tm_map(d, removePunctuation)
d <- tm_map(d,removeNumbers)
sw <- readLines("E:/TMT/stopwordses.txt",encoding="UTF-7")
sw = iconv(sw, to="ASCII//TRANSLIT")
d <- tm_map(d, removeWords, stopwords("spanish"))
d <- tm_map(d, removeWords, sw)
tdm <- TermDocumentMatrix(d)
m <- as.matrix(tdm)
v <- sort(rowSums(m),decreasing=TRUE)
df <- data.frame(word = names(v),freq=v)
#write.table(df,"E:/TMT/entrena_neu8.txt")
write.table(m,"E:/TMT/prue_neu2.txt")
#mt=t(m)
#write.table(mt,"E:/TMT/prueba_dtrmt.txt")

#############stemming y lemmantizacion########

library(SnowballC)
library(corpus)
tabes=read.table("E:/TMT/lemmatization-es.txt", header = FALSE, sep = "",
                 stringsAsFactors = FALSE,encoding = "UTF-7")
names(tabes) <- c("stem", "term")


stem_liste <- function(term) {
  i <- match(term, tabes$term)
  if (is.na(i)) {
    stem <- term
  } else {
    stem <- tabes$stem[[i]]
  }
  stem
}
```

```
limpiar_tokenizar <- function(texto){
  # El orden de la limpieza no es arbitrario
  # Se convierte todo el texto a minúsculas
  nuevo_texto <- tolower(texto)
  # Eliminación de páginas web (palabras que empiezan por "http." seguidas
  # de cualquier cosa que no sea un espacio)
  nuevo_texto <- str_replace_all(nuevo_texto,"http\\S*", "")
  # Eliminación de signos de puntuación
  nuevo_texto <- str_replace_all(nuevo_texto,"[[:punct:]]", " ")
  # Eliminación de números
  nuevo_texto <- str_replace_all(nuevo_texto,"[[:digit:]]", " ")
  # Eliminación de espacios en blanco múltiples
  nuevo_texto <- str_replace_all(nuevo_texto,"[\\s]+", " ")
  # Tokenización por palabras individuales
  nuevo_texto <- str_split(nuevo_texto, " ")[[1]]
  # Eliminación de tokens con una longitud < 2
  nuevo_texto <- keep(.x = nuevo_texto, .p = function(x){str_length(x) > 1})
  return(nuevo_texto)
}



######################Procesamiento descriptivo###############

library(rtweet)
library(tidytext)
library(tidyverse)
library(wordcloud2)
library(wordcloud)
library(RColorBrewer)
library(knitr)
library(igraph)
library(ggraph)
library(dplyr)
######descriptivos######
realtextl=read.csv(file="carga_real_lema.csv",header=TRUE,sep=";")
##conteo de tipo label##
realtextl %>% group_by(label) %>% summarise(numero_label = n())
realtextl %>%  ggplot(aes(x = label,fill=label)) + geom_bar() + coord_flip() +
  labs(x="Tipo de afiliado")+labs(y="Cantidad de afiliados")+
  ggtitle("Distribucion de afiliados por tipo")+theme_bw()+theme(legend.position = "none")
##tokenizacion
realtextl <- realtextl %>% mutate(texto_tokenizado = map(.x = texto,
```

```
                                        .f = limpiar_tokenizar))
realtextl %>% select(texto_tokenizado) %>% head()
realtextl %>% slice(1) %>% select(texto_tokenizado) %>% pull()

realtextl_tidy <- realtextl %>% select(-texto) %>% unnest()
realtextl_tidy <- realtextl_tidy %>% rename(token = texto_tokenizado)
head(realtextl_tidy)
##conteo de palabras por label##
realtextl_tidy %>% group_by(label) %>% summarise(n = n())
realtextl_tidy %>%  ggplot(aes(x = label,fill=label)) + geom_bar() + coord_flip() +
  labs(x="Tipo de afiliado")+labs(y="Cantidad de palabras")+
  ggtitle("Cantidad de palabras por tipo")+theme_bw()+theme(legend.position = "none")
##conteo de palabras distintas por label
realtextl_tidy %>% select(label, token) %>% distinct() %>%  group_by(label) %>%
  summarise(palabras_distintas = n())
realtextl_tidy %>% select(label, token) %>% distinct() %>%
  ggplot(aes(x = label,fill=label)) + geom_bar() + coord_flip() +
  labs(x="Tipo de afiliado")+labs(y="Cantidad de palabras unicas")+
  ggtitle("Cantidad de palabras unicas por tipo")+theme_bw()+theme(legend.position = "none")
##longitud media por usuario
realtextl_tidy %>% group_by(label, id) %>% summarise(longitud = n()) %>%
  group_by(label) %>% summarise(media_longitud = mean(longitud),sd_longitud =
sd(longitud))

realtextl_tidy %>% group_by(label, id) %>% summarise(longitud = n()) %>%
  group_by(label) %>%
  summarise(media_longitud = mean(longitud),
            sd_longitud = sd(longitud)) %>%
  ggplot(aes(x = label, y = media_longitud)) +
  geom_col() +
  geom_errorbar(aes(ymin = media_longitud - sd_longitud,
                    ymax = media_longitud + sd_longitud)) +
  coord_flip() + theme_bw()
###palabras mas usadas por grupo
realtextl_tidy %>% group_by(label, token) %>% count(token) %>% group_by(label) %>%
  top_n(10, n) %>% arrange(label, desc(n)) %>% print(n=30)

realtextl_tidy %>% group_by(label, token) %>% count(token) %>% group_by(label) %>%
  top_n(5, n) %>% arrange(label, desc(n)) %>%
  ggplot(aes(x = reorder(token,n), y = n, fill = label)) +
  geom_col() +
  theme_bw() +
  ggtitle("Cantidad de palabras por tipo")+
  labs(y = "frecuencia", x = "palabras") +
  theme(legend.position = "none") +
  coord_flip() +
```

```
  facet_wrap(~label,scales = "free", ncol = 1, drop = TRUE)
####nube de palabras###
df_grouped <- realtextl_tidy %>% group_by(label, token) %>% count(token) %>%
  group_by(label)

dfp <- filter(df_grouped, label=="PROMOTORES")
dfn <- filter(df_grouped, label=="NEUTROS")
dfd <- filter(df_grouped, label=="DETRACTORES")

wordcloud(words = dfp$token, freq = dfp$n,
          min.words = 100, random.order = FALSE,
          colors = brewer.pal(8, "Dark2"))

wordcloud(words = dfn$token, freq = dfn$n,
          min.words = 100, random.order = FALSE,
          colors = brewer.pal(8, "Dark2"))

wordcloud(words = dfd$token, freq = dfd$n,
         min.words = 100, random.order = FALSE,
         colors = brewer.pal(8, "Dark2"))
#####palabras comunes###

palabras_comunes <- dplyr::intersect(realtextl_tidy %>% filter(label=="PROMOTORES") %>%
                            select(token), realtextl_tidy %>% filter(label=="NEUTROS") %>%
                            select(token)) %>% nrow()
paste("Número de palabras comunes entre PROMOTORES y NEUTROS", palabras_comunes)

palabras_comunes <- dplyr::intersect(realtextl_tidy %>% filter(label=="PROMOTORES") %>%
                            select(token), realtextl_tidy %>% filter(label=="DETRACTORES")
%>%
                            select(token)) %>% nrow()
paste("Número de palabras comunes entre PROMOTORES y DETRACTORES",
palabras_comunes)

palabras_comunes <- dplyr::intersect(realtextl_tidy %>% filter(label=="DETRACTORES")
%>%
                            select(token), realtextl_tidy %>% filter(label=="NEUTROS") %>%
                            select(token)) %>% nrow()
paste("Número de palabras comunes entre DETRACTORES y NEUTROS", palabras_comunes)
#####palabras mas diferenciadas######
###PROMOTORES VS NEUTROS
# Pivotaje y despivotaje
realtextl_tidy$label=as.character(realtextl_tidy$label)

realtextl_spread <- realtextl_tidy %>% group_by(label, token) %>% count(token) %>%
  spread(key = label, value = n, fill = 0, drop = TRUE)
```

```
realtextl_unpivot <- realtextl_spread %>% gather(key = "label", value = "n", -token)

# Selección de los autores elonmusk y mayoredlee
realtextl_unpivot <- realtextl_unpivot %>% filter(label %in% c("PROMOTORES",
                                                  "NEUTROS"))
# Se añade el total de palabras de cada autor
realtextl_unpivot <- realtextl_unpivot %>% left_join(realtextl_tidy %>%
                                       group_by(label) %>%
                                       summarise(N = n()),
                                     by = "label")

# Cálculo de odds y log of odds de cada palabra
realtextl_logOdds <- realtextl_unpivot %>%  mutate(odds = (n + 1) / (N + 1))
realtextl_logOdds <- realtextl_logOdds %>% select(label, token, odds) %>%
  spread(key = label, value = odds)
realtextl_logOdds <- realtextl_logOdds %>%  mutate(log_odds =
log(PROMOTORES/NEUTROS),
                                       abs_log_odds = abs(log_odds))
# Si el logaritmo de odds es mayor que cero, significa que es una palabra con
# mayor probabilidad de ser de Elon Musk. Esto es así porque el ratio sea ha
# calculado como elonmusk/mayoredlee.
realtextl_logOdds <- realtextl_logOdds %>%
  mutate(autor_frecuente = if_else(log_odds > 0,
                                   "@PROMOTORES",
                                   "@NEUTROS"))
realtextl_logOdds %>% arrange(desc(abs_log_odds)) %>% head()

realtextl_logOdds %>% group_by(autor_frecuente) %>% top_n(15, abs_log_odds) %>%
  ggplot(aes(x = reorder(token, log_odds), y = log_odds, fill = autor_frecuente)) +
  geom_col() +
  labs(x = "palabra", y = "log odds ratio (@PROMOTORES / NEUTROS)") +
  coord_flip() +
  ggtitle("Palabras diferentes por Grupo")+
  theme_bw()

###PROMOTORES VS DETRACTORES
# Pivotaje y despivotaje
realtextl_tidy$label=as.character(realtextl_tidy$label)

realtextl_spread <- realtextl_tidy %>% group_by(label, token) %>% count(token) %>%
  spread(key = label, value = n, fill = 0, drop = TRUE)
realtextl_unpivot <- realtextl_spread %>% gather(key = "label", value = "n", -token)

# Selección de los autores elonmusk y mayoredlee
realtextl_unpivot <- realtextl_unpivot %>% filter(label %in% c("PROMOTORES",
                                                  "DETRACTORES"))
```

```
# Se añade el total de palabras de cada autor
realtextl_unpivot <- realtextl_unpivot %>% left_join(realtextl_tidy %>%
                                          group_by(label) %>%
                                          summarise(N = n()),
                                        by = "label")

# Cálculo de odds y log of odds de cada palabra
realtextl_logOdds <- realtextl_unpivot %>%  mutate(odds = (n + 1) / (N + 1))
realtextl_logOdds <- realtextl_logOdds %>% select(label, token, odds) %>%
  spread(key = label, value = odds)
realtextl_logOdds <- realtextl_logOdds %>%  mutate(log_odds =
log(PROMOTORES/DETRACTORES),
                                       abs_log_odds = abs(log_odds))
# Si el logaritmo de odds es mayor que cero, significa que es una palabra con
# mayor probabilidad de ser de Elon Musk. Esto es así porque el ratio sea ha
# calculado como elonmusk/mayoredlee.
realtextl_logOdds <- realtextl_logOdds %>%
  mutate(autor_frecuente = if_else(log_odds > 0,
                             "@PROMOTORES",
                             "@DETRACTORES"))
realtextl_logOdds %>% arrange(desc(abs_log_odds)) %>% head()

realtextl_logOdds %>% group_by(autor_frecuente) %>% top_n(15, abs_log_odds) %>%
  ggplot(aes(x = reorder(token, log_odds), y = log_odds, fill = autor_frecuente)) +
  geom_col() +
  labs(x = "palabra", y = "log odds ratio (@PROMOTORES / DETRACTORES)") +
  coord_flip() +
  ggtitle("Palabras diferentes por Grupo")+
  theme_bw()

###NEUTROS VS DETRACTORES
# Pivotaje y despivotaje
realtextl_tidy$label=as.character(realtextl_tidy$label)

realtextl_spread <- realtextl_tidy %>% group_by(label, token) %>% count(token) %>%
  spread(key = label, value = n, fill = 0, drop = TRUE)
realtextl_unpivot <- realtextl_spread %>% gather(key = "label", value = "n", -token)

# Selección
realtextl_unpivot <- realtextl_unpivot %>% filter(label %in% c("NEUTROS",
                                                 "DETRACTORES"))
# Se añade el total de palabras de cada autor
realtextl_unpivot <- realtextl_unpivot %>% left_join(realtextl_tidy %>%
                                          group_by(label) %>%
                                          summarise(N = n()),
                                        by = "label")
```

```
# Cálculo de odds y log of odds de cada palabra
realtextl_logOdds <- realtextl_unpivot %>%  mutate(odds = (n + 1) / (N + 1))
realtextl_logOdds <- realtextl_logOdds %>% select(label, token, odds) %>%
  spread(key = label, value = odds)
realtextl_logOdds <- realtextl_logOdds %>%  mutate(log_odds =
log(NEUTROS/DETRACTORES),
                                       abs_log_odds = abs(log_odds))
# Si el logaritmo de odds es mayor que cero, significa que es una palabra con
# mayor probabilidad de ser de Elon Musk. Esto es así porque el ratio sea ha
# calculado como elonmusk/mayoredlee.
realtextl_logOdds <- realtextl_logOdds %>%
  mutate(autor_frecuente = if_else(log_odds > 0,
                                   "@NEUTROS",
                                   "@DETRACTORES"))
realtextl_logOdds %>% arrange(desc(abs_log_odds)) %>% head()

realtextl_logOdds %>% group_by(autor_frecuente) %>% top_n(15, abs_log_odds) %>%
  ggplot(aes(x = reorder(token, log_odds), y = log_odds, fill = autor_frecuente)) +
  geom_col() +
  labs(x = "palabra", y = "log odds ratio (@NEUTROS / DETRACTORES)") +
  coord_flip() +
  ggtitle("Palabras diferentes por Grupo")+
  theme_bw()

###bigramas####
###PROMOTORES##
realtextlp <- filter(realtextl, label=="PROMOTORES")

bigramas <- realtextlp %>% mutate(texto = texto) %>%
  select(texto) %>%
  unnest_tokens(input = texto, output = "bigrama",
                token = "ngrams",n = 2, drop = TRUE)

# Contaje de ocurrencias de cada bigrama
bigramas  %>% count(bigrama, sort = TRUE)


graph <- bigramas %>%
  separate(bigrama, c("palabra1", "palabra2"), sep = " ") %>%
  count(palabra1, palabra2, sort = TRUE) %>%
  filter(n > 90) %>% graph_from_data_frame(directed = FALSE)
set.seed(123)

plot(graph, vertex.label.font = 2,
     vertex.label.color = "black",
```

```
    vertex.label.cex = 0.7, edge.color = "gray85")

###NEUTROS##
realtextln <- filter(realtextl, label=="NEUTROS")

bigramas <- realtextln %>% mutate(texto = texto) %>%
  select(texto) %>%
  unnest_tokens(input = texto, output = "bigrama",
                token = "ngrams",n = 2, drop = TRUE)

# Contaje de ocurrencias de cada bigrama
bigramas  %>% count(bigrama, sort = TRUE)


graph <- bigramas %>%
  separate(bigrama, c("palabra1", "palabra2"), sep = " ") %>%
  count(palabra1, palabra2, sort = TRUE) %>%
  filter(n > 50) %>% graph_from_data_frame(directed = FALSE)
set.seed(123)

plot(graph, vertex.label.font = 2,
     vertex.label.color = "black",
     vertex.label.cex = 0.7, edge.color = "gray85")

###DETRACTORES##
realtextld <- filter(realtextl, label=="DETRACTORES")

bigramas <- realtextld %>% mutate(texto = texto) %>%
  select(texto) %>%
  unnest_tokens(input = texto, output = "bigrama",
                token = "ngrams",n = 2, drop = TRUE)

# Contaje de ocurrencias de cada bigrama
bigramas  %>% count(bigrama, sort = TRUE)


graph <- bigramas %>%
  separate(bigrama, c("palabra1", "palabra2"), sep = " ") %>%
  count(palabra1, palabra2, sort = TRUE) %>%
  filter(n > 40) %>% graph_from_data_frame(directed = FALSE)
set.seed(123)

plot(graph, vertex.label.font = 2,
     vertex.label.color = "black",
     vertex.label.cex = 0.7, edge.color = "gray85")
```

```
####ensayo###
bigramasd <- realtextld %>% mutate(texto = texto) %>%
  select(texto) %>%
  unnest_tokens(input = texto, output = "bigramad",
                token = "ngrams",n = 2, drop = TRUE)

cbd= bigramasd  %>% count(bigramad, sort = TRUE)

bigramasn <- realtextln %>% mutate(texto = texto) %>%
  select(texto) %>%
  unnest_tokens(input = texto, output = "bigraman",
                token = "ngrams",n = 2, drop = TRUE)

cbn =bigramasn  %>% count(bigraman, sort = TRUE)

bigramasp <- realtextlp %>% mutate(texto = texto) %>%
  select(texto) %>%
  unnest_tokens(input = texto, output = "bigramap",
                token = "ngrams",n = 2, drop = TRUE)

cbp =bigramasp  %>% count(bigramap, sort = TRUE)



###############MODELAMIENTO#########################

###############BAYE SIN TF-IF###############

library(tm)
realtextl2=read.csv(file="carga_real_lema.csv",header=TRUE,sep=";")
realtextl2$texto=as.character(realtextl2$texto)

str(realtextl2)

table(realtextl2$label)

realtextl2_corpus <- Corpus(VectorSource(realtextl2$texto))

print(realtextl2_corpus)

inspect(realtextl2_corpus[1:3])

realtextl2_clean <- tm_map(realtextl2_corpus, tolower)
realtextl2_clean <- tm_map(realtextl2_clean, removeNumbers)
realtextl2_clean <- tm_map(realtextl2_clean, removeWords, stopwords())
realtextl2_clean <- tm_map(realtextl2_clean, removePunctuation)
realtextl2_clean <- tm_map(realtextl2_clean, stripWhitespace)
```

```
inspect(realtextl2_clean[1:3])

realtextl2_dtm <- DocumentTermMatrix(realtextl2_clean)
realtextl2_dtm

realtextl2_dtmif = weightTfIdf(realtextl2_dtm, normalize = TRUE)
realtextl2_dtmif

set.seed(123)
trainl <- sample(x = 1:nrow(realtextl2_dtm), size = 0.8 * nrow(realtextl2_dtm))

realtextl2_train <- realtextl2[trainl, ]
realtextl2_test  <- realtextl2[-trainl, ]

realtextl2_corpus_train <- realtextl2_clean[trainl ]
realtextl2_corpus_test  <- realtextl2_clean[-trainl ]

realtextl2_dtm_train <- realtextl2_dtm[trainl, ]
realtextl2_dtm_test  <- realtextl2_dtm[-trainl, ]

prop.table(table(realtextl2_train$label))
prop.table(table(realtextl2_test$label))

realtextl2_dict <- findFreqTerms(realtextl2_dtm_train, 5)
#sms_dict <- Dictionary(findFreqTerms(sms_dtm_train, 5))
realtextl2_md_train <- DocumentTermMatrix(realtextl2_corpus_train, list(dictionary =
realtextl2_dict))
realtextl2_md_test  <- DocumentTermMatrix(realtextl2_corpus_test, list(dictionary =
realtextl2_dict))

convert_counts <- function(x) {
  x <- ifelse(x > 0, 1, 0)
  x <- factor(x, levels = c(0, 1), labels = c("No", "Yes"))
}

memory.limit(3977)

realtextl2_md_train <- apply(realtextl2_md_train, MARGIN = 2, convert_counts)
realtextl2_md_test  <- apply(realtextl2_md_test, MARGIN = 2, convert_counts)

library(e1071)
library(gmodels)
realtextl2_classifier <- naiveBayes(realtextl2_md_train, realtextl2_train$label)
realtextl2_test_pred <- predict(realtextl2_classifier, realtextl2_md_test)
```

```
CrossTable(realtextl2_test_pred, realtextl2_test$label,
        prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
        dnn = c('predicted', 'actual'))

realtextl2_classifier2 <- naiveBayes(realtextl2_md_train, realtextl2_train$label, laplace = 0.1)
realtextl2_test_pred2 <- predict(realtextl2_classifier2, realtextl2_md_test)

CrossTable(realtextl2_test_pred2, realtextl2_test$label,
        prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
        dnn = c('predicted', 'actual'))

realtextl2_classifier3 <- naiveBayes(realtextl2_md_train, realtextl2_train$label, laplace = 1)
realtextl2_test_pred3 <- predict(realtextl2_classifier3, realtextl2_md_test)

CrossTable(realtextl2_test_pred3, realtextl2_test$label,
        prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
        dnn = c('predicted', 'actual'))

library(caret)
confusionMatrix(realtextl2_test_pred, realtextl2_test$label)
confusionMatrix(realtextl2_test_pred2, realtextl2_test$label)
confusionMatrix(realtextl2_test_pred3, realtextl2_test$label)


library(klaR)
model =
train(realtextl2_md_train,realtextl2_train$label,'nb',trControl=trainControl(method='cv',number=
10))

model$finalModel

realtextl2_testcv_pred <- predict(model$finalModel, realtextl2_md_test)

CrossTable(realtextl2_testcv_pred$class, realtextl2_test$label,
        prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
        dnn = c('predicted', 'actual'))

confusionMatrix(realtextl2_testcv_pred$class, realtextl2_test$label)


##################BAYES TF-IF#####################


set.seed(123)
trainlif <- sample(x = 1:nrow(realtextl2_dtmif), size = 0.8 * nrow(realtextl2_dtmif))
```

```
realtextl2_trainif <- realtextl2[trainlif, ]
realtextl2_testif  <- realtextl2[-trainlif, ]

realtextl2_corpus_trainif <- realtextl2_clean[trainlif ]
realtextl2_corpus_testif  <- realtextl2_clean[-trainlif ]

realtextl2_dtmif_train <- realtextl2_dtmif[trainlif, ]
realtextl2_dtmif_test  <- realtextl2_dtmif[-trainlif, ]

prop.table(table(realtextl2_trainif$label))
prop.table(table(realtextl2_testif$label))

realtextl2_dictif <- findFreqTerms(realtextl2_dtmif_train, 5)
#sms_dict <- Dictionary(findFreqTerms(sms_dtm_train, 5))
realtextl2_mdif_train <- DocumentTermMatrix(realtextl2_corpus_trainif, list(dictionary =
realtextl2_dictif))
realtextl2_mdif_test  <- DocumentTermMatrix(realtextl2_corpus_testif, list(dictionary =
realtextl2_dictif))

convert_counts <- function(x) {
  x <- ifelse(x > 0, 1, 0)
  x <- factor(x, levels = c(0, 1), labels = c("No", "Yes"))
}

realtextl2_mdif_train <- apply(realtextl2_mdif_train, MARGIN = 2, convert_counts)
realtextl2_mdif_test  <- apply(realtextl2_mdif_test, MARGIN = 2, convert_counts)

library(e1071)
library(gmodels)
set.seed(123)
realtextl2_classifierif <- naiveBayes(realtextl2_mdif_train, realtextl2_trainif$label)
realtextl2_test_predif <- predict(realtextl2_classifierif, realtextl2_mdif_test)


CrossTable(realtextl2_test_predif, realtextl2_testif$label,
           prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))

tablenbidf=table(predicho = realtextl2_test_predif, observado = realtextl2_testif$label)
tablenbidf

confusionMatrix(realtextl2_test_predif, realtextl2_test$label)

npd=precision(data=tablenbidf,reference = realtextl_test$label, relevant = "DETRACTORES")
npn=precision(data=tablenbidf,reference = realtextl_test$label, relevant = "NEUTROS")
```

```
npp=precision(data=tablenbidf,reference = realtextl_test$label, relevant = "PROMOTORES")

nrd=recall(data=tablenbidf,reference = realtextl_test$label, relevant = "DETRACTORES")
nrn=recall(data=tablenbidf,reference = realtextl_test$label, relevant = "NEUTROS")
nrp=recall(data=tablenbidf,reference = realtextl_test$label, relevant = "PROMOTORES")

f1_scorenbidfd=(2*npd*nrd)/(npd+nrd)
f1_scorenbidfn=(2*npn*nrn)/(npn+nrn)
f1_scorenbidfp=(2*npp*nrp)/(npp+nrp)


############SVM################
realtext=read.csv(file="carga_real_prev.csv",header=TRUE,sep=";")

realtext %>% group_by(label) %>% summarise(numero_label = n())

realtext$texto=as.character(realtext$texto)
realtext$label=as.character(realtext$label)

realtext <- realtext %>% mutate(texto_tokenizado = map(.x = texto,
                                                       .f = limpiar_tokenizar))
realtext %>% select(texto_tokenizado) %>% head()

realtext_tidy <- realtext %>% select(-texto) %>% unnest()
realtext_tidy <- realtext_tidy %>% rename(token = texto_tokenizado)
head(realtext_tidy)

realtext_tidy <- realtext_tidy %>% filter(!(token %in% sw))

write.table(realtext_tidy,"E:/TMT/realtext_tidy.txt")

###########carga lema############
library(rtweet)
library(tidyverse)
library(knitr)
library(quanteda)
library(tm)
library(twitteR)
realtextl=read.csv(file="carga_real_lema.csv",header=TRUE,sep=";")

realtextl %>% group_by(label) %>% summarise(numero_label = n())

realtextl$texto=as.character(realtextl$texto)
realtextl$label=as.character(realtextl$label)

#tweets_elon_ed <- tweets %>% filter(autor %in% c("elonmusk", "mayoredlee"))
##set.seed(123)
```

```
set.seed(369)
trainl <- sample(x = 1:nrow(realtextl), size = 0.8 * nrow(realtextl))
realtextl_train <- realtextl[trainl, ]
realtextl_test  <- realtextl[-trainl, ]

table(realtextl_train$label) / length(realtextl_train$label)

table(realtextl_test$label) / length(realtextl_test$label)

# Limpieza y tokenización de los documentos de entrenamiento
realtextl_train$texto <- realtextl_train$texto %>% map(.f = limpiar_tokenizar) %>%
  map(.f = paste, collapse = " ") %>% unlist()

# Creación de la matriz documento-término
sw <- readLines("E:/TMT/stopwordses.txt",encoding="UTF-7")
matriz_tl_tfidf_train <- dfm(x = realtextl_train$texto, remove = sw)

# Se reduce la dimensión de la matriz eliminando aquellos términos que
# aparecen en menos de 5 documentos. Con esto se consigue eliminar ruido.
matriz_tl_tfidf_train <- dfm_trim(x = matriz_tl_tfidf_train, min_docfreq = 5)

# Conversión de los valores de la matriz a tf-idf
matriz_tl_tfidf_train <- tfidf(matriz_tl_tfidf_train, scheme_tf = "prop",
                         scheme_df = "inverse")

matriz_tl_tfidf_train

# Limpieza y tokenización de los documentos de test
realtextl_test$texto <- realtextl_test$texto %>% map(.f = limpiar_tokenizar) %>%
  map(.f = paste, collapse = " ") %>% unlist()

# Identificación de las dimensiones de la matriz de entrenamiento
# Los objetos dm() son de clase S4, se accede a sus elementos mediante @
dimensiones_tl_matriz_train <- matriz_tl_tfidf_train@Dimnames$features

# Conversión de vector a diccionario pasando por lista
dimensiones_tl_matriz_train <- as.list(dimensiones_tl_matriz_train)
names(dimensiones_tl_matriz_train) <- unlist(dimensiones_tl_matriz_train)
dimensiones_tl_matriz_train <- dictionary(dimensiones_tl_matriz_train)

# Proyección de los documentos de test
matriz_tl_tfidf_test <- dfm(x = realtextl_test$texto,
                        dictionary = dimensiones_tl_matriz_train)
matriz_tl_tfidf_test <- tfidf(matriz_tl_tfidf_test, scheme_tf = "prop",
                         scheme_df = "inverse")
```

```
matriz_tl_tfidf_test

all(colnames(matriz_tl_tfidf_test) == colnames(matriz_tl_tfidf_train))

library(e1071)
modelo_tl_svm <- svm(x = matriz_tl_tfidf_train, y = as.factor(realtextl_train$label),
                     kernel = "linear", cost = 1, scale = TRUE,
                     type = "C-classification")
modelo_tl_svm

predicciones_tl <- predict(object = modelo_tl_svm, newdata = matriz_tl_tfidf_test)

tablesvm=table(predicho = predicciones_tl, observado = realtextl_test$label)
tablesvm

realtextl_test$label2=as.factor(realtextl_test$label)

confusionMatrix(predicciones_tl, realtextl_test$label2)

# Error de clasificación
clasificaciones_erroneas_tl <- sum(realtextl_test$label != predicciones_tl)
error_tl <- 100 * mean(realtextl_test$label != predicciones_tl)
paste("Número de clasificaciones incorrectas =", clasificaciones_erroneas_tl)

paste("Porcentaje de error =", round(error_tl,2), "%")

svm_cv_tl <- tune("svm", train.x =  matriz_tl_tfidf_train,
                  train.y = as.factor(realtextl_train$label),
                  kernel = "linear",
                  ranges = list(cost = c(0.1, 0.5, 1, 2.5, 5)))
summary(svm_cv_tl)


ggplot(data = svm_cv_tl$performances, aes(x = cost, y = error)) +
  geom_line() +
  geom_point() +
  geom_errorbar(aes(ymin = error - dispersion, ymax = error + dispersion)) +
  theme_bw()


predicciones_tl <- predict(object = modelo_cv_tl, newdata = matriz_tl_tfidf_test)

cpd=precision(data=tablesvm,reference = realtextl_test$label2, relevant = "DETRACTORES")
cpn=precision(data=tablesvm,reference = realtextl_test$label2, relevant = "NEUTROS")
cpp=precision(data=tablesvm,reference = realtextl_test$label2, relevant = "PROMOTORES")

crd=recall(data=tablesvm,reference = realtextl_test$label2, relevant = "DETRACTORES")
```

```
crn=recall(data=tablesvm,reference = realtextl_test$label2, relevant = "NEUTROS")
crp=recall(data=tablesvm,reference = realtextl_test$label2, relevant = "PROMOTORES")

f1_scored=(2*cpd*crd)/(cpd+crd)
f1_scoren=(2*cpn*crn)/(cpn+crn)
f1_scorep=(2*cpp*crp)/(cpp+crp)
```