PONTIFICIA UNIVERSIDAD JAVERIANA

MAESTRÍA ANALÍTICA PARA LA INTELIGENCIA DE NEGOCIOS

TRABAJO DE GRADO:

ANÁLISIS ETIQUETADO DE TEXTOS PARA PREDICCIÓN DE LA POLARIDAD, ENFOQUE SEMI SUPERVISADO Y ETIQUETADO AUTOMÁTICO

TEXTOLÍTICA CAOBA

MARIA ALEJANDRA LUQUE SÁNCHEZ

LUIS FELIPE CORTÉS DIAZ

Bogotá, 2020

TRABAJO DE GRADO: TEXTOLÍTICA CAOBA

El siguiente documento describe el contexto del negocio evaluado para desarrollar el proyecto de grado: Textolítica con el centro CAOBA. Este proyecto se enfoca en agregar valor a las soluciones ofrecidas por el centro a empresas colombianas para solucionar problemas de analítica relacionados con el etiquetado de la data para procesamiento de textos.

I. Entendimiento del negocio

Contexto:

Para entender el enfoque de este trabajo, primero es necesario entender qué es CAOBA. Alianza CAOBA es un centro de excelencia que apoya el uso de las tecnologías de *Big Data* y *Data Analytics* para fortalecer el análisis de la información en sectores reales, la alianza la conforman: el sector público a través de invitación directa del Ministerio de las tecnologías de la información y las comunicaciones (MINTIC) y Colciencias, la academia a través de las universidades Javeriana, Andes, EAFIT e ICESI, y el sector privado impulsado por patrocinadores como Bancolombia, Nutresa, SAS, IBM, entre otros. La potencia de esta alianza le permite interactuar en diferentes campos de análisis, tales como la investigación aplicada, la formación, el apoyo al emprendimiento y la consultoría.

Los servicios de consultoría de CAOBA se enfocan primordialmente en apoyar a empresas colombianas del sector real a implementar soluciones tecnológicas y de analítica de datos para afrontar los retos de la era digital a la que poco a poco ha ido migrando el país. La infraestructura de este centro, así como la experiencia de los profesionales que lo componen, permiten agregar un valor interesante en aras de que las empresas colombianas cada vez tengan mayor acceso a la tecnología y metodologías de punta para tomar mejores decisiones. Entendiendo a CAOBA como el beneficiario principal de este proyecto, es importante resaltar, que este centro presta servicio a diferentes clientes de todo el país de distintos sectores e industrias, y como dato a resaltar, el centro evidencia una demanda en particular por procesamiento de información en forma de textos, cerca del 50% de los proyectos que aborda el personal de CAOBA está relacionado con estos requerimientos.

Por este motivo, en el desarrollo de este trabajo de grado se buscó identificar áreas de las empresas que tuvieran relación con este tipo de data (textos) para ofrecer una solución a mayor escala. Dicho lo anterior, en discusión conjunta con los sponsors de este trabajo, una de las una de las áreas transversales a cualquier empresa es sin duda el área que da soporte y servicio a los clientes de dichas compañías y esta misma tiene relación estrecha con opiniones de los clientes que recibe en forma de audios, textos o mensajes a través de bots.

El área de servicio al cliente siempre ha sido una de las más relevantes dentro de las empresas, a pesar de ser un área operativa está ligada completamente a los objetivos estratégicos de las compañías. Si se parte del hecho que los clientes son el corazón de todo negocio, el llegar a entenderlos, canalizar sus dudas y quejas, para responder a tiempo con

calidad a sus problemas, es uno de los factores de fidelización más importantes hoy en día, los altos niveles de satisfacción de los clientes es una de las métricas más deseadas en cualquier industria. Sin embargo, a través de los años las áreas de soporte se han tenido que enfrentar a los retos de entender a esos clientes de maneras diferentes y con necesidades que evolucionan cada día, el contacto con el cliente cada vez se da más por medios digitales y la velocidad y asertividad de las respuestas deben apoyarse en nuevas tecnologías de la información, es ahí donde CAOBA puede potenciar la toma de decisiones de las empresas, facilitando el procesamiento y entendimientos de esta data a través de soluciones relacionadas con analítica.

Planteamiento del problema - Proceso Etiquetado:

En términos de analítica, el campo que explora los algoritmos asociados a procesamiento de textos es NLP[1] (Natural Lenguage Processing o Procesamiento de Lenguaje Natural) con diferentes variantes para modelar sobre esta tipología de data, tales como el pronóstico de la polaridad de un texto (sentiment analysis), o el análisis las emociones asociadas a algún contenido (emotions analysis) por nombrar los más comunes. Para los algoritmos de análisis de sentimientos, la evaluación busca determinar si el texto tiene una connotación positiva, negativa o neutra, lo que permitiría por ejemplo en el caso de las empresas a las que presta servicio CAOBA, tomar acciones diferenciadas para cada tipo o enfocar estrategias de respuesta diferenciadas. Sin embargo, la definición de lo que es positivo o negativo no deja de ser una connotación dada por el contexto en el cual se evalúa, y es por esto, que los análisis de NLP, cualquiera será el enfoque del que se hable, requieren de un trabajo de etiquetado por parte de expertos relacionados con dicha información.

Este trabajo de etiquetado (Data Labelling o Data Annotation) es clave para cualquier tipología de modelación en NLP y es una pieza fundamental que dota al modelo de calidad, sin embargo, es un trabajo que debe hacerse con rigurosidad si el objetivo es realmente contar con este nivel de calidad. El concepto de etiquetado manual se define como el proceso en que un experto se encarga de asignar para cada segmento del texto, una polaridad negativa, positiva o neutral según sea el caso, y llegar a estar en capacidad de realizar esto sobre la información requiere de capacidades y conocimientos muy específicos:

- *Conocimiento del contexto:* el etiquetador manual requiere tener el conocimiento del contexto al que pertenecen los datos que está analizando, es decir, entender primero la tipología de la información (estructura de los datos), entender por qué se recolectó dicha información, de qué entidad proviene, quienes participan, cuál es la finalidad de estos.
- *Experiencia en el dominio:* el etiquetador debe tener un entendimiento amplio en relación con el dominio de la data, conocer del entorno de la base que está analizando, esto le permite poder diferenciar la polaridad de palabras que pueden

[1] El Procesamiento de Lenguaje Natural o NLP por sus siglas en inglés, es un procedimiento de exploración que vincula la inteligencia artificial, las ciencias computacionales y el estudio de la lingüística del lenguaje humano.

tener connotaciones diferentes dependiendo el contexto en el que se evalúen, por ejemplo, la palabra vacío puede ser etiquetada como algo positivo o negativo dependiendo del entorno.

- *Conocimiento en lingüística:* esta es una de las capacidades más importantes que se deben tener en cuenta para el etiquetado manual, se requiere conocimiento en estructuras lingüísticas cuando se está analizando data de textos. Esto incluye entendimiento en gramática, sintaxis (composición del texto), lexicografía (manejo del vocabulario), semántica (significado de las palabras y lógica), pragmática (reconocimiento de expresiones dentro de un contexto con significado dentro del mismo), además claro de dominio del idioma en que se encuentre la información.
- *Dedicación*: la dedicación en el etiquetado se traduce en concentración para evitar equivocaciones, invertir tiempo y ser cuidadoso entendiendo con exactitud el texto analizado, es una tarea operativa que requiere aplicación y objetividad. Aproximadamente se requiere de 1 hora para etiquetar 100 sentencias de un texto por una sola persona.

Los conocimientos y capacidades que debe tener el experto que realizara el etiquetado manual se ven reflejadas en el proceso que esta tarea conlleva, algunas como:

- *Características del texto:* Entender la longitud de los textos a analizar, no es lo mismo la revisión de un tweet, un chat o un mensaje de correo.
- *Segmentación:* Identificar si el texto debe ser analizado por partes o su totalidad, por ejemplo, un correo electrónico podría contener en diferentes párrafos comentarios buenos como malos, lo que puede concluir en resultados no concluyentes.
- *Proceso Lingüístico:* Hacer la lectura de las unidades de texto definidas, donde aplicará los conocimientos lingüísticos presentados anteriormente, pero adicionalmente debe identificar palabras mal escritas, con errores de ortografía, modismos o palabras particulares del lenguaje o contexto.
- *Entidades:* Reconocer las entidades del texto, como, por ejemplo, personas, lugares, objetos, monedas, etc.
- *Ambigüedad:* Reconocer la ambigüedad del texto es una de las tareas más difíciles de trasladar a una máquina, palabras con connotación positiva y negativa pueden aparecer juntas en expresiones de ironía, y el etiquetador manual debe poder identificar el verdadero significado de la frase.
- *Correlación:* Reconocer cuando en un texto se hace referencia a entidades indirectamente (anáforas), el uso de expresiones como “aquel, el primero, el mismo” son un ejemplo de esta tarea que sólo viendo la totalidad del texto (frase, oración) puede tener sentido.
- *Objetividad:* Asignar la polaridad tras este análisis exhaustivo y mantener objetividad entre sentencias similares o determinar desempate de algunos términos comunes.

El proceso de etiquetado manual entonces es tan relevante porque recopila información sobre un entendimiento del texto propio de un experto, con habilidades que, aunque

pueden ser entrenadas en una máquina (Affective Computing) son naturales de los humanos, reconocer sentimientos en apartados de un texto y por lo tanto su valor es muy preciado. Además, el contraste de cualquier modelo a entrenar debe hacerse con data confiable y correctamente etiquetada, que se entiende hace parte de data etiquetada manualmente.

Entendiendo entonces la relevancia de un etiquetado manual en el análisis de textos, es posible entender con mayor claridad que esta tarea requiere de experiencia y dedicación que se traducen en costos y tiempos. Para aterrizar este entendimiento, a continuación, se presenta un benchmark estimado del costo para el etiquetado de 100 sentencias de 1 sola persona dedicada a dicha tarea entre diferentes compañías especializadas en etiquetado de información:

Entidad	Costo (USD)
Google	12.9
Amazon Recurso Interno	8
Amazon Manual	11.6
Amazon Integrado (sin costo entrenamiento)	9.2
iMerit	7.82
Cogito	6.44
SmartOne, Inc.	5.98
Vivetic	5.98
iVision- Foreign languages	10.58
Startek, Inc.	8.28

La estimación de estos costos supone que el tiempo promedio es de 36 segundos por sentencia, aproximadamente 1 hora para etiquetar 100 sentencias[2]. Considerando que en promedio una base de datos de texto como las que recibe CAOBA puede ascender a más de 20,000 sentencias, si el costo promedio por hora es de 8.68 usd, el costo estimado para etiquetar una base de este tamaño seria 1,736 usd. Es evidente que el costo en tiempo y por lo tanto en dinero escala rápidamente llegando a días de dedicación de una sola persona.

Es aquí entonces donde se plantea la problemática a resolver en este proyecto, visto desde la perspectiva de negocio, CAOBA tiene aquí una oportunidad de oferta importante, ante la demanda de sus clientes de procesar información en textos y conociendo que el proceso de etiquetado es el más costoso e importante para obtener resultados con valor, y conociendo además que este proceso es riguroso y puede estar sesgado por errores operativos vinculados a la experiencia de las personas que generan el etiquetado, CAOBA puede ofrecer dentro del portafolio de sus servicios, una solución de etiquetado manual de calidad y al mismo tiempo eficiente para sus clientes.

[2] CAOBA, referencia en tiempo estimado para etiquetar 100 sentencias cortas

Entendimiento de la solución:

Esta solución *Data Label Server*[3] que ofrecería CAOBA, es un servicio que ya ofrecen otras plataformas como Amazon con *Amazon Mechanical Turk*[4] y Cloudfactory con el módulo de *Data Procesing*[5] y cuyo enfoque principal es hacer más eficientes los procesos relacionados con el procesamiento de data, incluido el etiquetado de bases de datos.

Para CAOBA, este servicio podría presentarse dentro del flujo actual para modelado de NLP, junto con un módulo de Modelación y otro de Visualización. El módulo de Data Label Server permitiría cobrar por el etiquetado manual de una porción o de la totalidad de la base de datos que es entregada por el cliente. La decisión del cliente y de CAOBA de cuanto es necesario etiquetar, no solo radicará en su costo, sino también apoyada en un desempeño estimado del modelo con ciertos niveles de etiquetado.

El esquema del servicio que se propone para CAOBA tendría un diseño como el siguiente:

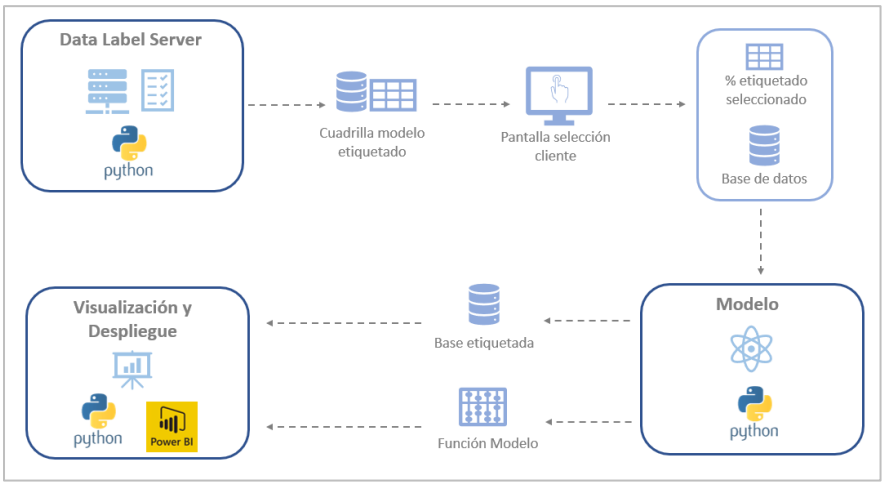
[3] Propuesta de nombre comercial a utilizar por CAOBA.

[4] "Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more." www.mturk.com

[5] "We staff CloudWorkers with the patience and consistency to perform repetitive tasks with precision. Focus on what matters while your team gives these important but time-consuming tasks the same attention to detail that you would if you had the time." www.cloudfactory.com

El alcance de este proyecto será todo lo relacionado con el módulo de etiquetado y el modelo asociado, sin embargo, el módulo de Modelo corresponde a todo el proceso de visibilidad interna para CAOBA que ejecuta los algoritmos para llegar a los resultados de nivel de etiquetado asociado a un desempeño.

El módulo de visualización se representará como una forma gráfica de mostrar algunos resultados del modelo al cliente, sin embargo, en el alcance de este proyecto se dejarán algunas bases para que pueda ser desarrollado con mayor detalle en estudios posteriores. A continuación, se presenta una propuesta de la arquitectura de información de esta solución:



Semi-supervised Learning y Transfer Learning:

Visto desde el punto de vista de analítica, la pregunta a responder es como entrenar un modelo para análisis de sentimientos[6] con un nivel de etiquetado manual para que con ese porcentaje manual el algoritmo aprenda y sea capaz de pronosticar el resto de las etiquetas faltantes. Para esto, se utilizarán dos conceptos de analítica: Semi-supervised Learning y Transfer Learning.

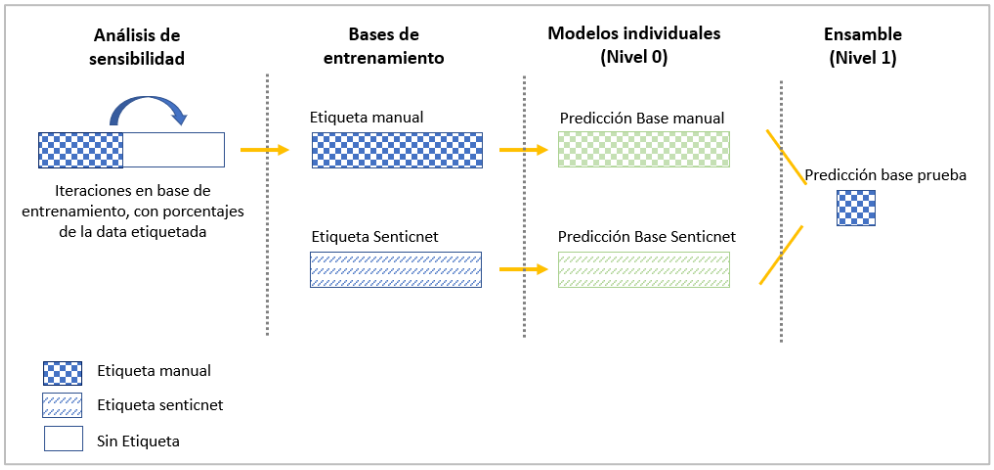
El aprendizaje semi supervisado, en un contexto de procesamiento de datos se entiende como la forma en que un algoritmo en su entrenamiento utiliza una cantidad de data etiquetada para pronosticar otra cantidad sin etiqueta. Considerando que un enfoque completamente supervisado implicaría tener etiquetado el 100% de la base, con los costos que esto implica en dinero y tiempo, y que un enfoque no supervisado implicaría trabajar con un base sin etiqueta alguna, con los costos al desempeño del modelo que esto traería, este enfoque semi supervisado trabaja con ambos escenarios, que para el caso de este proyecto ayuda a potenciar la relevancia de la data etiquetada, y aprender lo mejor que sea posible de esta. El planteamiento es entonces, mediante un análisis de sensibilidad iterar una base de entrenamiento con diferentes porcentajes de la data etiquetada manualmente

[6] El análisis de sentimientos o sentiment analysis es uno de los procesos más utilizados en NLP y puede ser el punto de partida para evaluar algoritmos más complejos sobre entendimiento de textos.

para pronosticar el porcentaje sin etiqueta y así obtener un pronóstico final sobre una base de prueba.

Con relación a Transfer Learning, si bien es cierto que el modelo de análisis de sentimientos podría cubrirse con el punto anterior (iteración de etiquetado manual), se quiso enriquecer este análisis de sensibilidad con otro modelo en paralelo que no utilizara etiquetas manuales, sino por el contrario dejara que esta función la cubriera una librería especializada en este tipo de tareas, *Senticnet*[7] desarrollada por Erik Cambria[8]. El objetivo de esto es aprovechar también el conocimiento que puede obtenerse de una librería que se ha construido mediante algoritmos de lenguaje también y entrenar un modelo con esta base de etiqueta sintética. El algoritmo para etiquetar la base con esta librería se referenció a un trabajo ya adelantado por CAOBA para modelos de NLP, específicamente el algoritmo desarrollado por Edwin Puertas.

Finalmente, con ambos resultados: modelo iterativo manual, modelo Senticnet, el objetivo es hacer un ensamble de tipo meta-modelo, Stacking, que permita aprender de ambos resultados con un pronóstico final. El ensamble de Stacking parte de los pronósticos realizados por modelos iniciales (modelos nivel 0) para crear un modelo final (modelo nivel 1) que entrena con dichos pronósticos, la cantidad de niveles a utilizar en este ensamble puede ser tanta como lo requiera el análisis, sin embargo, para este proyecto al tener sólo 2 modelos candidatos, se optará por llegar hasta el nivel 1 de un meta-modelo. La siguiente imagen ilustra el proceso.



Es importante resaltar que los resultados del modelo están sujetos a dos cosas: primero, al dominio asociado a las bases consideradas para el entrenamiento (Base de servicio al cliente de una entidad bancaría colombiana y partes de la base de Customer Compliant Dataset o CCD por sus siglas) si bien el dominio se amplía a textos de áreas de servicio al cliente, se

[7] SenticNet es una iniciativa concebida en el MIT Media Laboratory en 2009 con la colaboración del proyecto de Media Lab de la Universidad de Stirling y Sitekit Solutions basada en Singapur. El objetivo principal de SenticNet es desarrollar y aplicar análisis de sentimientos y emociones asociadas al lenguaje humano, para su interpretación a través de máquinas. (Sentic.net).

[8] Erik Cambria, creador de SenticNet, profesor asociado de la universidad NTU de Singapur y Provost Chair in Computer Science and Engineering. (Sentic.net).

debe tener precaución al aplicarlos sobre otro tipo de textos, así como se enfoca en bases con textos o transcripciones, sin considerar consumo de data no estructurada como videos, audios, imágenes; segundo, los resultados están sujetos también al tipo de ensamble seleccionado, la elección del método de Stacking se consideró como la más adecuada para el problema planteado, sin embargo, no se descarta que otro tipo de ensambles puedan tener resultados diferentes.

Objetivos y métricas:

Dicho lo anterior, los objetivos de este proyecto se alinean, el primero con una visión de beneficio hacia el cliente de CAOBA y el segundo con el interés hacia CAOBA:

- Desarrollar una solución analítica que pueda ofrecer CAOBA a sus clientes, mediante un servicio de etiquetado con algoritmos semi supervisados para el tratamiento de bases de datos relacionados con texto[9]. Esto con el objetivo de proponer ventajas en términos de costos, eficiencia y calidad a las empresas que asesora CAOBA.
- Construir una metodología de trabajo que pueda servir de base para CAOBA ante la implementación de la solución planteada, esta metodología incluirá las pautas principales a ofrecer, la matriz de resultados y la función de costos asociada.

El detalle de los objetivos planteados se alinea con los objetivos estratégicos de CAOBA: ofrecer servicios que den soluciones relevantes para las empresas colombianas que contribuyan a posicionar a Colombia como líder y referente en *Big Data* y *Data Analytics*.

[9] Inicialmente el segmento objetivo de la solución serían áreas de soporte, debido a que la base de entrenamiento del modelo de prueba corresponde con este dominio, sin embargo, la metodología puede expandirse a otras áreas y sectores realizando los respectivos ajustes.

Las métricas por considerar para alcanzar dichos objetivos se deben evaluar desde los diferentes enfoques a los que concierne, de la siguiente manera:

- Métricas analíticas: las métricas a considerar se asocian con el desempeño de las combinaciones de modelos a probar: Accuracy, especificidad, sensibilidad, F-Score, entre otras.
- Métricas CAOBA: las métricas de negocio serían dos, por un lado, medir el tiempo que conlleva realizar el etiquetado de X cantidad de datos, y una métrica asociada al dinero asociado al trabajo del etiquetado de esa X cantidad de datos.
- Métricas clientes: las métricas para los clientes de CAOBA podrían vincularse con las anteriores en relación con el costo en tiempo y en dinero que se ahorrarían las empresas al no contratar a CAOBA para el etiquetado de la su información, pero se plantea una métrica clave que sería el nivel de satisfacción de los clientes al percibir esta solución a su problemática actual y los beneficios obtenidos con la misma.

Entregables y Alcance:

Dentro de los entregables de este proyecto se presentan:

- Manual metodológico como soporte de la solución a ofrecer por CAOBA, enfatizando en el proceso completo de etiquetado y las ventajas que tendría tercerizar esta función con CAOBA.
- Códigos y documentación asociada al desarrollo de la modelación y pruebas.
- Matriz resultante con desempeño de las pruebas para diferentes niveles y categorías de etiquetado, ejecutado para la base de datos suministrada como prueba por CAOBA.
- Función de costos asociados a la solución y a los distintos niveles de etiquetado posibles.
- Vistas o ejemplos de un formato de entrada para los clientes y ejemplos de salidas en forma de tablero de mando con los resultados.

II. Entendimiento de la data

El conjunto de datos proporcionado por CAOBA pertenece al Servicio de Atención al Cliente en una entidad financiera. La información corresponde a las conversaciones entre los clientes y el personal de la entidad encargado de atender las solicitudes e inquietudes de clientes y usuarios. Por medio de este canal los clientes se ponen en contacto con el área de servicio para que les resuelvan algún problema específico, y las múltiples solicitudes y requerimientos como proporcionar información de productos y asesoría sobre los diferentes servicios y canales de atención que la entidad tiene disponible.

Por medio de la innovación en las técnicas de procesamiento del lenguaje natural se tiene acceso a una mejor comprensión de este tipo de datos junto con los modelos de conversación para el estudio de las prácticas modernas de atención al cliente y análisis de polaridad.

El conjunto de datos es un archivo csv (train.csv), donde cada fila corresponde a una línea de cada conversación entre el cliente y el asesor, incluye al menos una solicitud de cliente y una respuesta de la entidad. Se compone de la siguiente manera:

Característica	Descripción
Id	Contador para identificar cada una de las líneas
Comentario	Comentario del cliente o asesor, cada línea hace parte de una conversación.
Polaridad	Polaridad de cada comentario donde Neutral = 0 , Positivo = 1 , Negativo = -1

La primera exploración:

	id	tweets	polarity
0	1	buenas tardes, cordial saludo, dios te bendiga...	1
1	2	buenas tardes señora ACT_dd6762b3da08e42e3561...	0
2	3	bien gracias a dios	1
3	4	¡me alegra que se encuentre bien!	1
4	5	referente a mi tarjeta de crédito visa	1
5	6	permítame un momento por favor voy a validar l...	1
6	7	mi cédula es NUM_d83a2be7fc0ae0bf234164016a4ed...	1
7	8	señora ACT_6c565239f3bbb751f38573c023b3d627d7...	1

La base se compone de 23,685 registros con las tres características descritas.

La primera exploración nos muestra que es necesario aplicar un pre-procesamiento para remover todos los datos que hacen parte del texto, pero no ofrecen información para analizar.

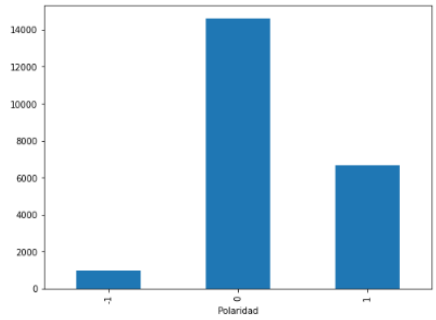
Una técnica común de pre-procesamiento es convertir el texto de entrada al mismo formato para que las palabras sean tratadas de la misma manera y reducir la duplicidad para obtener valores correctos de conteo. Eliminarla la puntuación es otra técnica común de pre-procesamiento, donde el objetivo es estandarizar los datos de texto. Adicionalmente se debe eliminar las palabras vacías, ya que son palabras que son utilizadas comúnmente en un idioma pero que deben ser eliminadas del texto la mayoría de las veces, ya que no proporcionan información valiosa para el análisis, de igual forma palabras frecuentes y raras.

Después de aplicar el proceso de limpieza descrito anteriormente obtenemos los siguientes resultados de una función simple de conteo:

Conteo de las 10 principales palabras:

Palabras	Frecuencia
Virtual	1087
Datos	1085
Cuenta	1070
Tarjeta	876
Clave	859
información	716
momento	700
Realizar	685
Crédito	660
Banco	653

En esta relación se puede identificar palabras y expresiones típicas de un contexto de servicio al cliente en una entidad financiera. La relación de la polaridad asociada a cada una de las líneas de las conversaciones se encuentra claramente desbalanceada con 66% para la polaridad Neutra 30% Positiva y tan solo 4% Negativa.



Negativa	Neutral	Positiva
968	14600	6652
4%	66%	30%

Para complementar la información anterior CAOBA nos proporcionó el conjunto de datos Consumer Complaint Database (CCD), el cual corresponde a las quejas que los consumidores de productos y servicios financieros han remitido a las instituciones con las que tienen algún tipo de relación. Esta es una base de acceso público[10] organizada por el gobierno de los Estados Unidos y que en trabajos previos de Caoba fue traducida al español.

Característica	Descripción
Id	Contador para identificar cada una de las líneas
Comentario	Reclamo del cliente de Productos y servicios Financieros

Se seleccionaron de manera aleatoria 33,904 registros para contar con un nivel similar de información a la base de Servicio al Cliente.

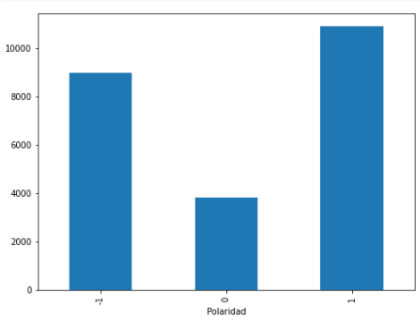
[10] https://catalog.data.gov/dataset/consumer-complaint-database

	id	tweets	polarity
14953	14954	Llamé a Transunion, XXXX y XXXX varias veces p...	1
3899	3900	Antes de partir de XXXX, mi esposa y yo congel...	1
27643	27644	He refinanciado mi casa el último día del XXXX...	-1

Después de aplicar el proceso de limpieza, pre-procesamiento y etiquetado manual[11], la primera exploración nos entrega el siguiente resultado de palabras con mayor frecuencia, donde se observa una mayor cantidad de palabras y mayor frecuencia de las más populares, lo que indica que nos puede aportar mayor cantidad de características para las tares de clasificación.

Palabras	Frecuencia
cuenta	21957
crédito	21028
pago	13574
préstamo	10543
informe	9481
información	9414
deuda	7849
después	7795
pagos	7271
tarjeta	7148

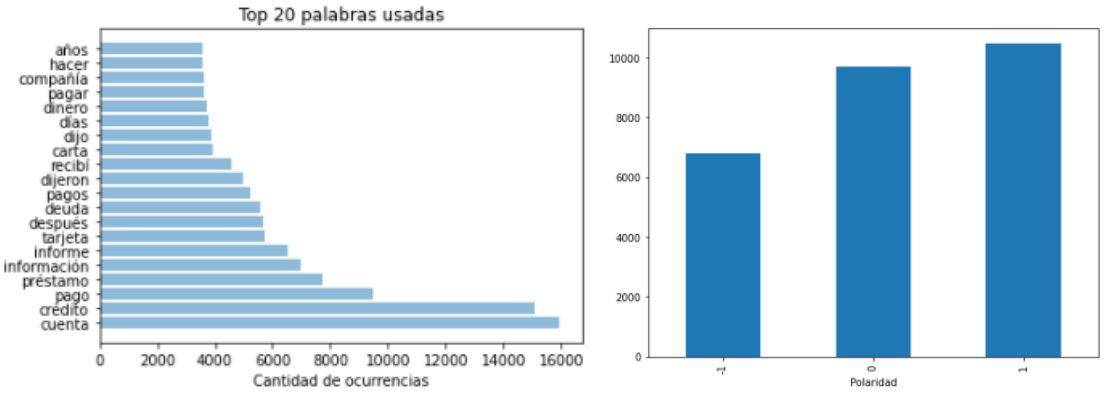
Con relación a la polaridad, la base CCD presenta un desbalanceo en la polaridad positiva con el 46%, la polaridad neutra con el 16% y la negativa con el 38%. Lo que tiene sentido, ya que es una base de reclamos donde se evidencia un nivel bajo de satisfacción de los clientes con sus entidades financieras.



Negativa	Neutral	Positiva
12755	5587	15562
38%	16%	46%

La distribución de frecuencias para la base total partir de los textos de las conversaciones de servicio al cliente de la entidad financiera y la base de reclamos CCD permite ver el top de las 20 palabras más usadas en toda la base.

[11] El proceso de etiquetado se describe en la sección Etiqueta Manual



Con la unificación de las dos bases se observa que ahora las categorías se encuentran mejor niveladas (Positiva 39%, Neutral 37%, Negativa 24%), lo que nos aporta más información sobre las características para cada polaridad, donde se espera que sean significativas en el momento de calibrar los clasificadores.

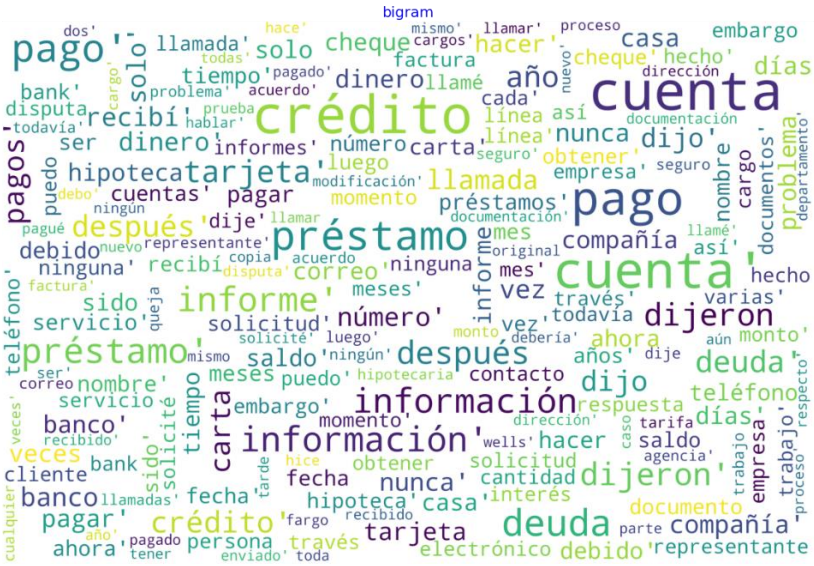
Por medio de la librería *sklearn.feature*_selection.chi2(X, y), se calcula las estadísticas de chi cuadrado entre las conversaciones y la polaridad. La prueba mide la dependencia entre las variables, por lo que al utilizar esta función "se eliminan" las características que tienen más probabilidades de ser independientes de la polaridad, por lo tanto, irrelevantes para la clasificación. Los unigramas más correlacionados muestran que si existe una relación coherente con la polaridad asignada, por ejemplo “negativo” o “cerró” en polaridad negativa y “perfecto” o “alegra” con la positiva. Un aspecto importante que llama la atención es la coincidencia de palabras entre la polaridad neutral y positiva.

. Most correlated unigrams:		
# '-1': Negativa	# '0': Neutral	# '1': Positiva
. negativo	. recibí	. bancolombia
. cerró	. pagar	. virtual
. amable	. disculpe	. legal
. aut	. años	. justos
. virtual	. casa	. legales
. tardes	. sucursal	. nueva
. crédito	. queja	. social
. pagos	. dijo	. clave
. claro	. bancolombia	. muchisimas
. comprendo	. debido	. interés

La información de unigramas es útil, sin embargo los bigramas y trigramas son estructuras que pueden ofrecer mayor contexto con relación a cada una de las polaridades, por ejemplo, en la negativa se presenta “robo identidad” o “información falsa”, en el caso de la polaridad positiva se puede ver conceptos como “pago total” o “ autenticación exitosa”

Los bigramas más correlacionados:

. Most correlated bigrams:		
# '-1': Negativa	# '0': Neutral	# '1': Positiva
. verificar cuentas	. cuota manejo	. encuentra línea
. queja formal	. pueda interesar	. virtual personas
. tarifa pago	. cuénteme puedo	. pagó totalidad
. robo identidad	. servicio cliente	. excelente preguntar
. canal virtual	. ejecución hipotecaria	. tipo contrato
. sucursal virtual	. pagos atrasados	. acciones legales
. visite canal	. continua linea	. ciudad encuentra
. información falsa	. permanece línea	. preguntar encuentre
. días retraso	. segunda clave	. seguridad social
. informe crédito	. disculpe continúa	. autenticación exitosa



Los trigramas más correlacionados

. Most correlated trgrams:		
# '-1': Negativa	# '0': Neutral	# '1': Positiva
. crédito encontré pago	. tardes visite canal	. emprender acciones legales
. discutí siendo pasado	. pantalla habilitará módulo	. excelente agradezco preguntar
. días pueden verificar	. inferior pantalla habilitará	. crédito best buy
. empresa archivo utilizó	. documento identidad clave	. tipo contrato maneja
. proporcione documentos empresa	. víctima robo identidad	. agente bienes raíces
. cuentas discutí siendo	. habilitará módulo ingrese	. solicité informe crédito
. eliminen rápidamente toda	. módulo ingrese documento	. canal utilizar servicio
. documentos empresa archivo	. datos ingresados seguros	. través canal utilizar
. marca días pueden	. tipo contrato maneja	. atenderle través canal
. pueda verificar disputado	. clic botón ingresar	. disculpe permanece línea

Los bigramas y trigramas más correlacionados muestran para la polaridad negativa expresiones como “quiero bloquear tarjetas” o “error aplicación” que están normalmente relacionadas con experiencias negativas de los clientes. De igual forma se puede ver en la nube, palabras relacionadas con la entidad financiera como productos y servicios bancarios, pero lo que más tiene relevancia en la polaridad Neutra y Positiva son las expresiones relacionadas con guiones de servicio al cliente, donde algunas características con los saludos y expresiones amables.

III. Transformación de la data

Tras la descripción de la data con la que se realizarán los experimentos de niveles de etiquetado, y considerando que se trata de bases de datos de textos, la transformación y procesamiento de esta está en línea con las técnicas tradicionales de limpieza. La estructura de la base de datos que está conformada por 3 columnas, explicadas en el capítulo anterior, se ve de la siguiente manera:

id	Comentario	polarity
1	buenas tardes, cordial saludo, dios te bendiga...	1
2	buenas tardes señora ACT_dd6762b3da08e42e3561...	0
3	bien gracias a dios	1
4	¡me alegra que se encuentre bien!	1
5	referente a mi tarjeta de crédito visa	1

Los comentarios que trae la base de datos contienen, como es de esperar, caracteres especiales como puntuación, o algún código especial, por ejemplo, y si bien es cierto que para algunos análisis de NLP la conjugación con estos caracteres les otorga un significado diferente a las oraciones, para este análisis, partiendo de una base de datos con comentarios de quejas y solicitudes, es posible que el uso gramatical no sea particularmente estricto y, por lo tanto, el aporte de estos signos pasa a un segundo plano. La primera tarea en el proceso de limpieza es eliminarlos del análisis, obteniendo una base que recoja un campo [Comentario_clean] que luzca de la siguiente manera:

Comentario	polarity	Comentario_clean
buenas tardes, cordial saludo, dios te bendiga...	1	buenas tardes cordial saludo dios te bendiga...
buenas tardes señora ACT_dd6762b3da08e42e3561...	0	buenas tardes señora CodACT bienvenida a ba...
bien gracias a dios	1	bien gracias a dios
¡me alegra que se encuentre bien!	1	me alegra que se encuentre bien

Con la limpieza genérica realizada, es importante determinar la unidad lingüística que será objeto del análisis, es decir, con qué parte del texto se trabajará: sentencias, oraciones o palabras como fuente de información primaria. A este proceso se le conoce como tokenización y además de ser el punto de partida para todo trabajo sobre procesamiento de textos, es la segmentación de la data para llegar a un nivel detallado de entendimiento, tal como mencionan [Ravi & Ravi] "*Los datos brutos adquiridos de diversas fuentes a menudo deben ser preprocesados antes de iniciar un análisis completo. Algunos de los pasos de preprocesamiento más populares son: la tokenización, la eliminación de las palabras vacias, el etiquetado de las partes de la oración (POS) y la extracción y representación de características. La "Tokenización" se utiliza para dividir una frase en palabras, frases,*

símbolos u otros tokens significativos mediante la eliminación de los signos de puntuación."[12]

En tanto, respecto a la elección de la unidad lingüística o token a utilizar dependerá en gran parte de los objetivos del análisis y el lenguaje a analizar, y esto lleva a que existan varias definiciones de lo que podría ser un token en NLP, sin embargo, se reduce en gran medida a un marco léxico, *"Hablando como lexicógrafo, J. McH. Sinclair propone definir un elemento léxico como "un elemento formal (de al menos un morfema de longitud) cuyo patrón de ocurrencia puede describirse en términos de una serie de otros elementos léxicos ordenados de forma única que se producen en su entorno"*[13].

En el idioma español, la unidad de morfema corresponde con la definición de palabra, no puede ser desagregada en más partes y por si sola puede otorgar un significado, si bien es cierto que algunos multi-morfemas puedan cumplir con esta función no refieren una mayoría importante en nuestro lenguaje, por lo tanto, para efectos de este análisis se tomará a las palabras como unidad de token a analizar, el resultado se evidencia en el campo [Comentario_tokenized]:

Comentario_clean	Comentario_tokenized
buenas tardes cordial saludo dios te bendiga...	[buenas, tardes, cordial, saludo, dios, te, be...
buenas tardes señora CodACT bienvenida a ba...	[buenas, tardes, señora, codact, bienvenida, a...
bien gracias a dios	[bien, gracias, a, dios]

Otra de las técnicas de limpieza que mencionan Ravi & Ravi, es la de eliminar las stopwords del texto. Estas stopwords hacen referencia a palabras que no agregan un valor semántico[14] al análisis del texto, tales como signos de puntuación, conectores, artículos, entre otros, con una frecuencia de uso muy elevada y que hacen parte de la sintaxis[15] de las oraciones con las que se comunica el ser humano, son parte importante del lenguaje, pero su interpretación es diferente para NLP, *"Las Stopwords o palabras vacías son palabras comunes que están presentes en el texto pero que generalmente no contribuyen al significado de una oración. Casi no tienen importancia a los efectos de la recuperación de*

[12] Kumar Ravi, Vadlamani Ravi. "A Survey on opinion mining and sentiment analysis: task, approaches and applications". 2015. Pag. 6.
[13] Jonathan J. Webster & Chunyu Kit. "Tokenization as the initial phase in NLP". 1992.
[14] Entiendo al valor semántico como aquel que agrega determinada parte del texto para dar significado y lógica a lo que se analiza del mismo.
[15] Noam Chomsky, padre de la gramática generativa disciplina que situó la sintaxis en el centro de la investigación lingüística. Con esta cambió la perspectiva, los programas y métodos de investigación en el estudio del lenguaje. Wikipedia.

información y el procesamiento del lenguaje natural. Pueden ser ignoradas con seguridad sin sacrificar el significado de la frase"[16].

Los mayores trabajos realizados en librerías para detectar stopwords se encuentran explorados ampliamente para el idioma inglés y existe una menor cantidad especializado en español, sin embargo, para este pre-procesamiento de la información, se trabajó con el corpus de la herramienta NLTK para Python.

Comentario_clean	Comentario_tokenized	Comentario_nostop
buenas tardes cordial saludo dios te bendiga...	[buenas, tardes, cordial, saludo, dios, te, be...	[buenas, tardes, cordial, saludo, dios, bendig...
buenas tardes señora CodACT bienvenida a ba...	[buenas, tardes, señora, codact, bienvenida, a...	[buenas, tardes, señora, codact, bienvenida, b...
bien gracias a dios	[bien, gracias, a, dios]	[bien, gracias, dios]
me alegra que se encuentre bien	[me, alegra, que, se, encuentre, bien]	[alegra, encuentre, bien]

El siguiente proceso de limpieza de la información, es llegar a normalizar un texto para poder procesarlo de manera correcta. Una de las formas de llegar a esto es estimizar (stemming) la data, este proceso es importante debido a que la mayoría de los textos, por razones gramaticales, contienen una cantidad de diferentes usos de una misma palabra, familias de palabras o simplemente la conjugación de una palabra para diferentes sujetos, pero con el mismo significado. Por este motivo, en aras de facilitar el análisis del texto, el proceso de stemming permite identificar la raíz de una palabra, *"La principal aspiración del stemming es cambiar una palabra derivada a su forma estándar y mantener la palabra raíz"*[17]. En el caso de la base de CAOBA, un ejemplo es el estema "*buen*" derivado de la palabra "*buenas*".

Comentario_clean	Comentario_tokenized	Comentario_nostop	Comentario_stemmed
buenas tardes cordial saludo dios te bendiga...	[buenas, tardes, cordial, saludo, dios, te, be...	[buenas, tardes, cordial, saludo, dios, bendig...	[buen, tard, cordial, salud, dios, bendig, des...
buenas tardes señora CodACT bienvenida a ba...	[buenas, tardes, señora, codact, bienvenida, a...	[buenas, tardes, señora, codact, bienvenida, b...	[buen, tard, señor, codact, bienven, bancolomb...
bien gracias a dios	[bien, gracias, a, dios]	[bien, gracias, dios]	[bien, graci, dios]
me alegra que se encuentre bien	[me, alegra, que, se, encuentre, bien]	[alegra, encuentre, bien]	[alegr, encuentr, bien]

[16] Sunakshi Mamgain. "Stopwords: Important for the language not so in NLP". 2019.
[17] Karthick , R John Victor, Manikandan, Bhargavi Goswami. " Professional Chat Application based on Natural Language Processing".

Sin embargo, el proceso de stemming, puede presentar algunas imperfecciones al momento de analizar los segmentos del texto, ya que se trata de una tarea de heurísticas que en ocasiones corta la parte final de la palabra con el fin de identificar la raíz inexistente, por ejemplo, es el caso de la palabra "*tardes*" cuyo estema es "*tard*", el token "*tard*" no tiene como tal un significado evidente para el lenguaje y menos para este texto. Es por esto que, para el tratamiento de esta base, se ejecutó un proceso de lematización. La lematización tiene como fin identificar el lema de una palabra, entendiendo al lema como la forma única canónica de una palabra, "*La lematización suele referirse a hacer las cosas correctamente con el uso de un vocabulario y un análisis morfológico de las palabras, normalmente con el objetivo de eliminar sólo las terminaciones inflexionales y devolver la base o la forma de diccionario de una palabra, lo que se conoce como el lema*"[18].

Volviendo al ejemplo de la palabra "*tardes*" de la base de datos, mediante la lematización, el lema que encuentra ya no es "*tard*", sino "*tarde*".

Comentario_clean	Comentario_tokenized	Comentario_nostop	Comentario_stemmed	Comentario_lemmatized
buenas tardes cordial saludo dios te bendiga...	[buenas, tardes, cordial, saludo, dios, te, be...	[buenas, tardes, cordial, saludo, dios, bendig...	[buen, tard, cordial, salud, dios, bendig, des...	bueno tarde cordial saludar dios bendecir dese...
buenas tardes señora CodACT bienvenida a ba...	[buenas, tardes, señora, codact, bienvenida, a...	[buenas, tardes, señora, codact, bienvenida, b...	[buen, tard, señor, codact, bienven, bancolomb...	bueno tarde señor codact bienvenido bancolombi...
bien gracias a dios	[bien, gracias, a, dios]	[bien, gracias, dios]	[bien, graci, dios]	bien gracia dios

Con los pasos descritos en esta sección, ya se cuenta con una base de datos pre-procesada para empezar el trabajo de modelación.

Extracción de características de texto

En el campo de análisis de texto es posible la aplicación de algoritmos de aprendizaje de máquinas. Sin embargo, es importante tener en cuenta que los algoritmos no pueden trabajar directamente con el texto en bruto, ya que la mayoría de ellos esperan vectores de características numéricas con un tamaño fijo en lugar de documentos de texto con longitud variable. En el procesamiento del lenguaje natural, una técnica común para extraer características del texto es colocar todas las palabras que aparecen en el documento en una representación de bolsa de palabras, este enfoque se denomina modelo BoW[19], porque no le importa en qué orden están las palabras.

La extracción de características consiste en transformar datos de un documento como palabras o frases, en características numéricas utilizables para el aprendizaje automático. La forma más común es por medio de la representación de una matriz con una fila por documento y una columna por token (por ejemplo, una palabra) que aparece en el texto. Al proceso general de recuento y normalización de la "bolsa de palabras o n-gramas" de le

[18] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. "Introduction to Information Retrieval". Cambridge University Press. 2008.
[19] (Scikit-learn. Text feature extraction, 2020)

denomina vectorización[20], ya que convierte una colección de documentos de texto en vectores de características numéricas.

La representación más común es contar las ocurrencias de palabras o tokens de una serie de textos. A cada término encontrado se le asigna un valor (0,1) si coincide o no con una columna de la matriz resultante. Para aumentar el vocabulario o el número de características y conservar un ordenamiento local se puede extraer n-gramas de palabras además de las palabras individuales.

La representación de matriz de un conteo normal para los siguientes textos de puede ver a continuación.

Texto1 = "podemos generar extracto de la Tarjeta de Crédito"
Texto2 = "enviar información y cupo del extracto de cuenta"
Texto3 = "negar cupo de Crédito rotativo"
Texto4 = "la información de aprobación de Crédito"

	aprobacion	credito	cuenta	cupo	enviar	extracto	generar	informacion	negar	podemos	rotativo	tarjeta
0	0	1	0	0	0	1	1	0	0	1	0	1
1	0	0	1	1	1	1	0	1	0	0	0	0
2	0	1	0	1	0	0	0	0	1	0	1	0
3	1	1	0	0	0	0	0	1	0	0	0	0

<u>Ponderación de términos Tf-idf:</u>

La frecuencia inversa de los documentos es el concepto de que las palabras que son más populares en todos los textos deben ser menos importantes, esto porque llevan muy poca información significativa sobre el contenido real del documento. Para ponderar las características de frecuencias de términos más raros pero más interesantes en un clasificador, es muy común utilizar la transformación tf-idf donde:

Tf significa término-frecuencia, es la suma de todas las ocurrencias o el número de veces que aparece un término en un documento.

Tf-idf significa término-frecuencia por frecuencia inversa del documento

$$tfidf(t,d) = tf(t,d).idf(t)$$

el término frecuencia, se multiplica con el componente idf, que se calcula como

$$idf(t) = \log\frac{1+n}{1+df(t)} + 1$$

[20] (Scikit-learn. Text feature extraction, 2020)

donde n es el número total de documentos del conjunto de documentos, y df(t) es el número de documentos del conjunto de documentos que contienen el término t. Los vectores tf-idf resultantes se normalizan entonces por la norma euclidiana:

$$v_{norm} = \frac{v}{||v||_2} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}}$$

La normalización de la tf-idf suele ser muy útil, sin embargo existe casos donde las representaciones binarias de ocurrencia puedan ofrecer mejores características.

	aprobacion	credito	cuenta	cupo	enviar	extracto	generar	informacion	negar	podemos	rotativo	tarjeta
0	0.000000	0.317993	0.000000	0.000000	0.000000	0.392784	0.498197	0.000000	0.00000	0.498197	0.00000	0.498197
1	0.000000	0.000000	0.508672	0.401043	0.508672	0.401043	0.000000	0.401043	0.00000	0.000000	0.00000	0.000000
2	0.000000	0.366747	0.000000	0.453005	0.000000	0.000000	0.000000	0.000000	0.57458	0.000000	0.57458	0.000000
3	0.702035	0.448100	0.000000	0.000000	0.000000	0.000000	0.000000	0.553492	0.00000	0.000000	0.00000	0.000000

Etiquetas Manuales:

El proceso de etiquetado manual de datos que seguimos para complementar la información sobre servicio al cliente en una entidad bancaria, se aplicó sobre la base CCD de reclamos de clientes. Adicional a la información sobre la narrativa de los reclamos se identifica información sobre productos y tipología de reclamos que podrían ser útiles para la clasificación.

- Selección de Muestra

  Se seleccionó una muestra de 33904 registros de la base CCD traducida por Caoba con el objetivo de complementar las observaciones y aumentar las características aplicables a un contexto de servicio al cliente.

- Identificación del contexto Reclamos y Solicitudes Entidades Financieras

  Adicional a la descripción de la muestra, se analiza la asociación de los reclamos con los productos financieros o con las tipologías identificadas de reclamos. Por lo general los clientes generan reclamaciones por situaciones que las entidades no ha dado tramite oportuno o la solución no resulta ser satisfactoria.

- Análisis de la Narrativa

  En general los reclamos a las entidades financieras deben cumplir con una serie de elementos como la descripción detallada del reclamo y como la entidad no ha

solucionado estas situaciones. Adicional es necesario aportar pruebas y soportes para demostrar cada uno de los hechos en reclamación.

- Asignación Inicial de Polaridad

  En esta etapa del proceso se procedió con la selección de una submuestra aproximada de 500 registros. Se analizó manualmente con el fin de identificar características, asociación de productos o tipo de reclamos que puedan evidenciar alguna de las polaridades objetivo. Como se propone en diferentes plataformas que ofrecen servicio de etiquetado, se crean especificaciones o instrucciones para que puedan ser aplicadas a todos los elementos del conjunto de datos[21].

- Asignación de Polaridad a la muestra total

  De acuerdo con las instrucciones o agrupaciones identificadas, se aplica sobre la base total la asignación de polaridad. Después del procesamiento de todos los registros, se verifica la consistencia de la asignación.

En las tareas de clasificación manual se sigue el procedimiento descrito anteriormente, lo que contribuye a mejorar el balance las categorías objetivo. Es importante mencionar que en la etapa de exploración, se aplicaron las técnicas de vectorización y extracción de características numéricas, con el objetivo de observar relación de alguna de las polaridades con grupos de vectores similares, sin embargo estos resultados no fueron de ayuda en el proceso de etiquetado manual.

Etiquetas Senticnet:

Como lo hemos mencionado en las secciones anteriores el etiquetado de datos es una etapa indispensable del pre-procesamiento en las tareas de analítica. El etiquetado de datos, también llamado anotación o clasificación, es el proceso de asignación de atributos, características o clasificaciones al conjunto de datos para el aprendizaje automático y reconocimiento de patrones. Obtener datos etiquetados de alta calidad es una tarea que se vuelve significativa en un proyecto en términos de funcionalidad y costo.

Adicional a los enfoques de etiquetado manual, se han identificado métodos que automatizan en parte el proceso y reducen la necesidad de participación humana, donde las diferentes herramientas hacen que el etiquetado sea más rápido y más barato. Estas herramientas agilizarán el flujo de trabajo de etiquetado para las tareas relacionadas con el procesamiento del lenguaje natural como el análisis de sentimientos, la vinculación de entidades, la categorización de textos, el análisis sintáctico y el etiquetado.

SenticNet es un conjunto de herramientas y técnicas para el análisis de sentimientos que combinan el razonamiento de sentido común, la psicología, la lingüística y el aprendizaje

[21] (Google cloud , Tareas de etiquetado de texto, 2020)

automático[22]. SenticNet realiza la detección de polaridad y el reconocimiento de emociones aprovechando la semántica y la lingüística, es decir, que el análisis de sentimiento se aplica a nivel de concepto y no depende únicamente de las frecuencias de co-ocurrencia de las palabras.

Parte fundamental para el desarrollo de este análisis fue la utilización de la base de conocimiento SenticNet[23], la cual proporciona un conjunto de semántica, sentimientos y polaridad asociados a 100.000 conceptos de lenguaje natural.

Dentro de las diferentes tareas de análisis de sentimientos, SenticNet realiza la detección de la polaridad por medio de patrones sensitivos[24]. Esos patrones se aplican al árbol sintáctico de dependencia de una frase, donde se identifica las palabras que tienen polaridad intrínseca y las palabras que modifican el significado de otras palabras. Después de eliminar las palabras no utilizadas para el cálculo de la polaridad (en blanco), el resultado es un número flotante entre -1 y +1 (donde -1 es negatividad extrema y +1 es positividad extrema).

El proceso de asignación automático de etiquetas para determinar la polaridad en la base de servicio al cliente y base de reclamos CCD, se realizó por medio de la librería senticnet 1.3 desarrollada en Python. [25] Adicionalmente se utilizó el código desarrollado por Edwin Puerta[26] para el análisis de los textos, procesamiento y asignación de polaridad por medio de SenticNet. A continuación, se puede ver un ejemplo de asignación de polaridad.

texto = 'buenas noches señor, le doy la bienvenida a nuestro can al virtual ¿cómo se encuentra?'

Language: es
Pipe: ['emoji', 'sentencizer', 'tagger', 'parser', 'stemmer', 'ner']
['noches', 0.0]
['buenas noches', 0.121]
['señor', 0.036]
['la bienvenida', 0.0]
['nuestro canal', 0.0]
['a nuestro canal', 0.0]
['virtual nuestro canal', 0.0]
Polarity: 0.022
Polaridad: Positiva

[22] https://sentic.net/

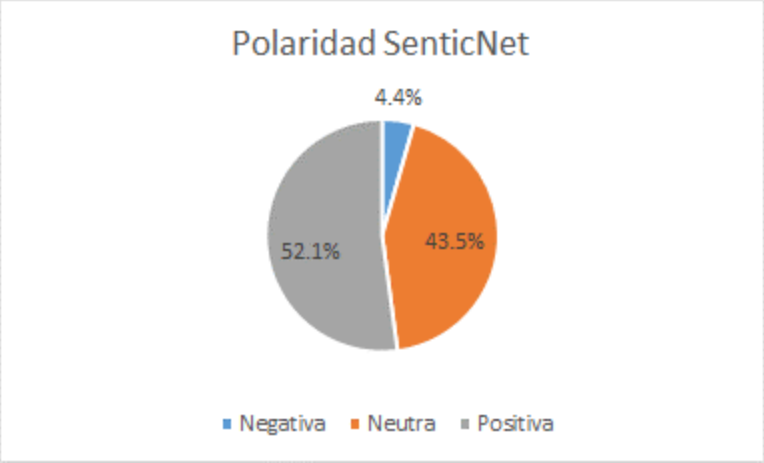
[23] SenticNet 5 alcanza los 100.000 conceptos empleando redes neuronales recurrentes. Está disponible en 40 idiomas diferentes.

[24] (Poria, Cambria, Winterstein, & Huang, 2014)

[25] (Cambria, Poria, Hazarika, & Kwok, 2018)

[26] https://grid.utb.edu.co/user/epuerta/notebooks/TransferLearning_CustomerServices/logic/models/linguistic.ipynb

El resultado del proceso de etiquetado automático por medio de SenticNet se resume en la siguiente grafica



IV. Modelo

El detalle de los pasos anteriores explica la descripción de la información que recibida en este proyecto y las transformaciones que se realizaron para hacer que la data se encontrara en la mejor forma para entrar al modelo. En este apartado se explicará el proceso que se llevó a cabo para seleccionar un modelo que cumpliera con los objetivos de este trabajo, los resultados aquí planteados dependen exclusivamente de los algoritmos y funciones a utilizar como herramientas para cumplir ese objetivo de una manera adecuada, sin embargo, eso no excluye que cualquier otra elección sobre parámetros no pueda reflejar resultados diferentes.

Antes de comenzar con la descripción del proceso, es importante retomar algunos conceptos planteados en apartados anteriores: a) el dataset de entrada cuenta con información de servicio al cliente de una entidad financiera colombiana, sin embargo, los resultados de este tipo de ejercicios de pronósticos de polaridad, están muy relacionados con el dominio de la base de datos que se evalúa, motivo por el cual se determinó ampliar el dataset de entrada incluyendo información de una base de servicio al cliente CCD (customer compliant dataset); b) la transformación de la data permitió el proceso de tomar un comentario crudo y procesarlo para tener diferentes versiones de él, comentario con limpieza de puntuación, sin stopwords, lematizado o estimizado, este comentario procesado será el insumo para el modelo a modo de variable explicativa con su respectiva vectorización; c) el data set de entrada contiene un campo de etiquetado manual [polarity] y un campo con los resultados arrojados por senticnet [senticnet] para 3 polaridades, positiva, negativa y neutra; d) el primer nivel de modelo de este proyecto contempla la iteración de un enfoque semi supervisado que sensibilizando un porcentaje de las etiquetas de entrada genera las etiquetas faltantes; e) los pronósticos de este modelo se realizarán a través del uso de dos metodologías de etiquetado, una manual y otra por medio de

senticnet, para finalmente por medio de un ensamble consideran las mejores predicciones de ambos enfoques.

Selección muestras:

Con la base de datos completa, el primer paso a considerar es la selección de la muestra de entrenamiento y prueba, la base de prueba seleccionada será contra la que se validarán cada uno de los modelos a construir (senticnet y manual) y la que será la referencia para el resultado final del ensamble. La selección de muestras en un proceso de modelado puede recurrir a diferentes técnicas, entre las que están y se recomienda el uso de tuneo del porcentaje a seleccionar, sin embargo, en este caso se determinó utilizar un punto fijo para la partición, fijado en 80% para entrenamiento y 20% para prueba. Esta decisión corresponde a que la base de prueba debe estar etiquetada en su totalidad, es decir que de cara al servicio que implementará CAOBA esta base debe etiquetarse por completo y no está dentro de la selección de porcentaje a iterar (seleccionado por CAOBA o el cliente) que solo aplicará a la base de entrenamiento ya que la base de prueba es el soporte real contra el que se corroborarán las predicciones y desempeño del modelo; entonces con el fin de optimizar los costos de etiquetado manual (que es una de los métricas de este trabajo) se plantea un porcentaje fijo para seleccionar la base de test y se fija en 20% ya que es un balance utilizado con frecuencia, y que al utilizarlo en este caso concluyó en buenos resultados del modelo. Otro de los motivos importantes para esta decisión corresponde a que el otro 80% restante de la base, que también se fija en ese porcentaje, se prefiere mantener constante ya que, al ser la parte de la base susceptible a la iteración de sensibilidad, no se desea que los resultados de las particiones estén afectados por la selección inicial de la muestra de entrenamiento.

Variables del modelo:

Este modelo de análisis de sentimientos recibe dos parámetros, por un lado, el comentario y por otro la polaridad, siendo el primero (comentario) la variable explicativa de la segunda, sin embargo, la forma como este comentario entra al modelo es en forma de una vector de palabras. Como se menciona en el apartado de transformación de la data, las unidades que conforman este vector son los tokens definidos en el procesamiento, que para este caso son palabras; cuando estos tokens se evalúan de forma unitaria se entienden como unigramas, cuando se evalúan de pares se entienden como bigramas, y así sucesivamente dependiendo del conjunto de tokens que se seleccione. Para este caso en particular, se decidió considerar como input del vector tanto a los unigramas como a los bigramas del texto, principalmente para recoger también significados de la unión de dos palabras consecutivas, por ejemplo, evaluar la palabra *servicio,* por un lado, y evaluar la aparición de *mal-servicio* o *buen-servicio* por el otro.

Una vez definido el conjunto de vectores de entrada, el proceso de modelado recurre a la función de TF-IDF, también explicada en detalle en el apartado anterior, que en términos

generales calcula el TF (Term Frequency) la frecuencia de aparición de un término (unigrama o bigrama) en un comentario, y a su vez calcula el IDF (Inverse Document Frequency) la frecuencia con que el término aparece en todos los comentarios del texto, penalizando (Inverse) apariciones muy recurrentes en todos los comentarios, ya que palabras que aparecen con mucha frecuencia en todos los documentos se asocian a palabras con poca relevancia por muy comunes. La función de TF-IDF genera entonces las variables finales que entran al modelo para análisis de sentimientos.

La otra parte relevante en términos de variables es la variable a explicar, en este caso es la polaridad asociada a cada comentario, sin embargo, esta base de datos tiene dos campos relacionados con la polaridad, 1) el campo [polarity] que es una etiqueta manual asignada por un experto tras evaluar detalladamente cada comentario del dataset, y 2) el campo [senticnet] que es la polaridad que generó automáticamente el algoritmo de la librería de senticnet. El tener dos variables con la etiqueta de polaridad, se debe a que este proceso de modelación se conforma por tres partes: 1) iteración de pronóstico semi supervisado con la etiqueta manual, es decir con el campo [polarity] ya que se requiere variar el porcentaje de iteración de lo que para un nuevo cliente decide CAOBA será lo que realmente etiquetará de forma manual. 2) construcción de un modelo que recibe como input la base de etiqueta manual tras la iteración anterior, y que pronóstica con el campo [polarity] finalmente sobre la base de prueba y 3) un modelo que entrena con la polaridad de senticnet, este modelo usa entonces el campo [senticnet] y pronóstica sobre la base de prueba. La conjunción de estos tres puntos termina en un meta-modelo (ensamble) que toma los resultados del punto 2 y 3 y genera un pronóstico final. Es importante mencionar, que todos los ajustes de pronósticos contra la base de prueba se realizarán con el campo [polarity] de etiqueta manual independiente de si se trata del modelo 2 o 3, ya que se asume que ese es valor real que toma la base de prueba, el valor asignado manualmente.

Parámetros del modelo:

En el apartado de transformación de la data, se enfatiza en el procesamiento que se le realiza a la base de datos, con el fin de convertir el comentario original en una expresión gramatical que permita al modelo hacer mejores predicciones. Entre esas transformaciones se encuentran 3 importantes, la primera removiendo signos de puntuación y stopwords, la segunda estimizando el comentario y la tercera llegando hasta el proceso de lematizado, cada uno de estos procesos resulta en un conjunto de palabras que, aunque representan al comentario original son diferentes.

Las predicciones del modelo estarán entonces relacionadas con el contenido que reciba de la variable comentario (en cualquiera de sus transformaciones) por lo tanto para este ejercicio se decidió que la transformación del comentario será uno de sus parámetros, sin embargo, solo se iterará con dos de las tres transformaciones mencionadas: comentario sin stopwords, y comentario lematizado. El motivo de esta decisión se basa en que la transformación a estemas del texto es un proceso que trata de homologar la aparición de

palabras buscando la raíz de estas, sin embargo, en este intento puede llegar a cortar parte de la palabra haciendo que el resultado no tenga ninguna relación con la palabra original. Por otro lado, el considerar la transformación de stopwords mantiene el texto escrito de manera original, eliminando palabras irrelevantes y puntuación, y el proceso de lematizado, reduce la cantidad de palabras buscando su forma única canónica para poder asociar un mismo concepto a palabras que pertenecen a la misma familia. En resumen, los parámetros a utilizar serán:

- Comentario sin stopwords: corresponde a la limpieza básica e inicial del texto que mantiene el estado original de las palabras, eliminando las palabras más comunes el idioma.
- Comentario lematizado: corresponde a la limpieza del comentario completa, pasando por la eliminación de stopwords, pero llevándolo al nivel de encontrar el lema del texto.

Igualmente otro de los parámetros del modelo tiene que ver con la variable a explicar, la polaridad, que en el estado original de la base de datos a probar toma tres valores, positivo, negativo y neutro; los resultados del análisis de sentimiento es sensible entonces a la cantidad de categorías que se evalúan, por lo tanto se realizarán pruebas tomando la variable polaridad con tres valores así como sólo para dos: positivo y negativo, asumiendo que la categoría neutra es clasificada como alguna de las primeras. En este caso se asumió que la asignación de la categoría neutra se transfiriera a la categoría positiva, principalmente para dar visibilidad de las verdaderas palabras asociadas al carácter negativo de la base de datos, cosa que se entiende puede ser de mayor relevancia para los clientes.

Otro parámetro para considerar en este modelo es el balance de la variable polaridad, tanto para el ejercicio con tres o dos categorías. Para la primera parte del proceso completo, es decir, la sensibilización del nivel de etiquetado de la base de datos se decidió no balancear la variable de polaridad, ya que este ejercicio busca aprender de la base de la parte etiquetada en su estado original, precisamente para probar la potencia de ese aprendizaje para etiquetar el resto de la base. Para el modelo generado por senticnet tampoco se balanceó la base de datos para poder probar la verdadera efectividad de senticnet al momento de evaluarlo en conjunto con el modelo manual, sin embargo para la generación del ensamble final, si se balancearon las categorías para toda la base con los predichos tanto manuales como de senticnet, esto mostró alguna mejora frente a no balancear la data, sin embargo no evidencia mejoras muy distintas, lo que soporta la decisión de no realizar este proceso en todo las etapas del modelado.

Finalmente, otro parámetro a utilizar en todo este proceso es el clasificador para pronosticar, en el caso de los 3 primeros modelos se decidió tunear este hiper parámetro considerando 4 de los clasificadores de aprendizaje supervisados más utilizados para desarrollo de modelos en NLP:

- Random Forest: ensamble de distintos árboles de decisión para clasificar en alguna de las categorías.
- Support Vector Machine: en este caso representado por Support Vector Classifier (SVC) que trabaja con más de dos clases para pronosticar, para abarcar el escenario de usar tres categorías.
- Naive Bayes: trabajando el clasificador multinomial de este algoritmo.
- Regresión Logística.

Sensibilización – Enfoque etiquetado semi supervisado:

Esta etapa del proceso se refiere a cómo tomando como punto de partida unas etiquetas de la base de datos se construye un modelo para generar el resto de las etiquetas faltantes. Esto se ejecutará sobre la base de entrenamiento, y busca sustentar la hipótesis de que no es necesario considerar una base etiquetada por completo para generar un modelo de análisis de sentimientos, sino que con un porcentaje de la base etiquetada, un modelo puede aprender lo suficiente para hacer buenas predicciones sobre una porción no etiquetada, para probar esto se generaran particiones de la base de datos con y sin etiquetas para generar una base completa etiquetada que será el insumo para el modelo 2) de pronóstico manual.

Los rangos de iteración se definieron en deciles, recorriendo desde el 10% al 100%[27]

de la base con etiqueta:

10%	20%	30%	40%	50%	60%	70%	80%	90%	99%

Inicialmente se consideraron estos niveles como punto de partida para la sensibilización, sin embargo, los resultados que arrojó el modelo demuestran que una partición más pequeña no genera una ganancia en desempeño, adicional a que, de cara a la oferta de CAOBA, es más intuitivo ofrecer un etiquetado por deciles que entrar al detalle de percentiles, ofrecer un 20% de etiquetado es más sencillo de entender por un cliente que ofrecer un 18%, por ejemplo.

Con los rangos de iteración definidos, se generó una función que selecciona de la base completa de entrenamiento, una muestra aleatoria del porcentaje seleccionado, es decir, para la primera iteración se selecciona una muestra aleatoria que represente el 10% de la base completa y se considera la etiqueta manual que trae esa porción de la base, para el 90% restante se elimina la etiqueta y se asume que no trae etiqueta alguna. Con esta primera partición (10%), se construye un clasificador para predecir la etiqueta del resto de la base (90%), este proceso se lleva a cabo iterando con el porcentaje de selección inicial.

Este modelo para el proceso iterativo se construye entonces con las variables que entran al modelo, polaridad y comentario. Para no sesgar los resultados con la selección de un solo

[27] Como punto de referencia del decil se seleccionó 99% para que el modelo aún tuviera un valor a pronosticar.

clasificador, se construyó un algoritmo que tunea este parámetro y encuentra el mejor modelo en términos de desempeño entre un Random Forest, un SVM, Naive Bayes y una Regresión logística en cada una de las particiones. El desempeño del modelo se evalúa contra la etiqueta real de la porción seleccionada para no tener etiqueta.

Adicionalmente, para el caso de los clasificadores también se realizó un tuneo por ejemplo para el algoritmo de Random Forest[28], se iteró con diferentes valores de árboles para que el modelo seleccionara la mejor opción de este clasificador antes de competir con los otros tres. Dicho esto, entonces para cada partición el clasificador seleccionado será el que mejor se ajuste a los datos entre los cuatro mencionados.

Adicionalmente, agregando otras dos dimensiones a este ejercicio, este proceso se ejecutó variando los parámetros de tipo de comentario de entrada (stopword y lematizado) y variando la cantidad de categorías para la polaridad (3 o 2).

Polaridad	3 categorías			
Comentario	Lematizado		No stopword	
Pct Etiquetado	Mejor modelo	Accuracy	Mejor modelo	Accuracy
10%	Logistic Regresion	0,6292	Logistic Regresion	0,6276
20%	Logistic Regresion	0,6572	Logistic Regresion	0,6502
30%	Logistic Regresion	0,6654	Logistic Regresion	0,6626
40%	Logistic Regresion	0,6802	Logistic Regresion	0,6734
50%	Logistic Regresion	0,6880	Logistic Regresion	0,6784
60%	Logistic Regresion	0,6888	Logistic Regresion	0,6816
70%	Logistic Regresion	0,6990	Logistic Regresion	0,6852
80%	Logistic Regresion	0,6931	Logistic Regresion	0,6884
90%	Logistic Regresion	0,6952	Logistic Regresion	0,6950
99%	Logistic Regresion	0,7109	Logistic Regresion	0,6783
Polaridad	**2 categorías**			
Comentario	**Lematizado**		**No stopword**	
Pct Etiquetado	**Mejor modelo**	**Accuracy**	**Mejor modelo**	**Accuracy**
10%	Logistic Regresion	0,7892	Logistic Regresion	0,7826
20%	Logistic Regresion	0,8087	Logistic Regresion	0,7965
30%	Logistic Regresion	0,8183	Logistic Regresion	0,8115
40%	Logistic Regresion	0,8253	Logistic Regresion	0,8133
50%	Logistic Regresion	0,8310	Logistic Regresion	0,8175
60%	Logistic Regresion	0,8338	Logistic Regresion	0,8309
70%	Logistic Regresion	0,8334	Logistic Regresion	0,8321
80%	Logistic Regresion	0,8411	Logistic Regresion	0,8389
90%	Logistic Regresion	0,8329	Logistic Regresion	0,8483
99%	Naive Bayes	0,8326	SVM	0,8348

[28] El tuneo de cantidad de árboles a considerar para el Random Forest se evalúo para [10,50,100,200].

Como se puede ver en la tabla anterior, las predicciones sobre una polaridad con sólo dos categorías son mejores con relación al ajuste (fit), sin embargo, es importante recordar que esta métrica se está evaluando sobre la base de entrenamiento y si bien es cierto que estos resultados empiezan a mostrar unas conclusiones del modelo, el resultado final se evaluara bajo la luz de otras métricas de desempeño adicionales, y sobre la base de prueba final después de haber realizado el ensamble completo. Estos resultados son entonces indicadores del aprendizaje de cada una de las iteraciones.

El resultado de todas estas iteraciones es una base de datos de entrenamiento etiquetada al 100% para cada una de las particiones, es decir, de la partición del 10% etiqueta y 90% sin etiqueta, saldrá una base de datos con el 100% de las etiquetas, solo que 90% de estas habrán sido estimadas en el proceso iterativo a partir del 10%. Cada una de estas bases nuevas completadas serán el insumo para probar un modelo de predicción de sentimientos que se evaluará sobre la misma base de prueba que se separó originalmente.

Modelo etiqueta manual:

El pronóstico del modelo de etiqueta manual recibe por cada iteración una base de datos con la totalidad de las etiquetas, el objetivo entonces es predecir sobre una base de prueba separada. Para este proceso también se construyó un modelo que evalúa 4 clasificadores: Random Forest, SVM, Naive Bayes y Regresión logística (con sus tuneos propios) a modo de tunear este parámetro para el pronóstico que como se mencionó puede tomar resultados distintos dependiendo del clasificador.

El modelo encuentra entonces el mejor clasificador que se ajusta con cada nueva base que va recibiendo, este proceso itera también para cada una de las particiones que tuvo la base inicial ya que se corre sobre 9 bases de entrenamiento diferentes.

Al igual que en el proceso anterior, este modelo también recibe como parámetro la cantidad de categorías del campo polaridad y el tipo de comentario que entra. A continuación, se muestran los resultados del fit del modelo contra la base de prueba, recordando que los resultados finales se mostrarán en la descripción del ensamble:

Polaridad	3 categorías			
Comentario	Lematizado		No stopword	
Pct Etiquetado	Mejor modelo	Accuracy	Mejor modelo	Accuracy
10%	Logistic Regresion	0,6321	SVM	0,6362
20%	Logistic Regresion	0,6577	Logistic Regresion	0,6564
30%	Logistic Regresion	0,6663	SVM	0,6670
40%	SVM	0,6760	Logistic Regresion	0,6776
50%	Logistic Regresion	0,6835	SVM	0,6820
60%	Logistic Regresion	0,6854	SVM	0,6801
70%	Logistic Regresion	0,6894	Logistic Regresion	0,6924
80%	Logistic Regresion	0,6951	SVM	0,6894
90%	Logistic Regresion	0,6968	Logistic Regresion	0,6934

99%	Logistic Regresion	0,7051	Logistic Regresion	0,7002
Polaridad	**2 categorías**			
Comentario	**Lematizado**		**No stopword**	
Pct Etiquetado	**Mejor modelo**	**Accuracy**	**Mejor modelo**	**Accuracy**
10%	SVM	0,7894	Logistic Regresion	0,7826
20%	Logistic Regresion	0,8053	Logistic Regresion	0,7953
30%	Logistic Regresion	0,8147	Logistic Regresion	0,8093
40%	Logistic Regresion	0,8240	Logistic Regresion	0,8157
50%	Logistic Regresion	0,8282	Logistic Regresion	0,8197
60%	Logistic Regresion	0,8302	Logistic Regresion	0,8294
70%	Logistic Regresion	0,8337	Logistic Regresion	0,8305
80%	Logistic Regresion	0,8349	Logistic Regresion	0,8365
90%	Logistic Regresion	0,8390	Logistic Regresion	0,8361
99%	Logistic Regresion	0,8404	Logistic Regresion	0,8398

Entre los resultados presentados anteriormente se destaca que el clasificador de regresión logística es seleccionado como mejor modelo en gran parte de los casos, adicionalmente se puede ver que para todas las particiones los resultados son muy similares, con desviaciones entre el 1% y 3% dependiendo la dimensión evaluada. Así mismo se analizaron los resultados de la predicción (con la iteración del 99%) de este método manual y se equivoca una mayor cantidad de veces pronosticando la polaridad negativa.

Modelo etiqueta SenticNet:

Para la fase del modelo de senticnet, se realizó el mismo proceso que para la base etiquetada e iterada con la etiqueta manual, la diferencia de este modelo es que su variable a predecir ya no es [polarity] que es manual, sino [senticnet] que es el resultado de haber sometido a la base inicial a la clasificación bajo la librería senticnet.

Para este caso en particular no se tiene en cuenta ningún tipo de iteración ni sensibilidad de porcentajes de etiquetado, ya que el objetivo de usar este pronóstico en el ensamble es enriquecer el pronóstico manual (Transfer learning) y modificar su entrada de etiquetas no tiene genera un beneficio para CAOBA ya que este proceso de etiquetado es automático y funciona igual para el 10% o 100% de la base, salvaguardando claro el costo en tiempo de procesamiento que si podría variar, sin embargo, es un costo marginal que se puede considerar como parte del módulo de modelado y no discriminado para el módulo de Data Label Server.

3 categorías			
Lematizado		**No stopword**	
Mejor modelo	**Accuracy**	**Mejor modelo**	**Accuracy**
Random Forest	0,5465	Random Forest	0,5438
2 categorías			
Lematizado		**No stopword**	

Mejor modelo	Accuracy	Mejor modelo	Accuracy
Random Forest	0,7597	Random Forest	0,7599

Entre los resultados presentados para el modelo manual, era evidente el protagonismo del clasificador Logistic Regresion, sin embargo, al ejecutar el proceso para los diferentes parámetros sobre el modelo de senticnet, en todos los casos el mejor clasificador es un Random Forest[29].

Analizando los resultados predichos por SenticNet, este método se está equivocando más en predecir las palabras dentro de la categoría de palabras neutrales, posiblemente debido al entrenamiento que tenga esta librería con palabras más definidas y menos neutras.

Ensamble Stacking:

Un ensamble en machine learning es una forma de combinar diferentes predicciones de uno o varios modelos a modo de mejorar el rendimiento final, la forma de llegar a este objetivo cambia según la necesidad de cada modelo y abre un abanico de posibilidades a métodos para ensamblar, principalmente los ensambles responden a combinaciones de distintos modelos o a combinaciones de un solo modelo probando con un set de muestras distintas. A partir de esto, un modelo ensamblado puede recorrer opciones desde votaciones sencillas de las partes como promedios, max voting, ponderaciones, entre otras, hasta técnicas más avanzadas como: Stacking, Blending, Bagging y Boosting.

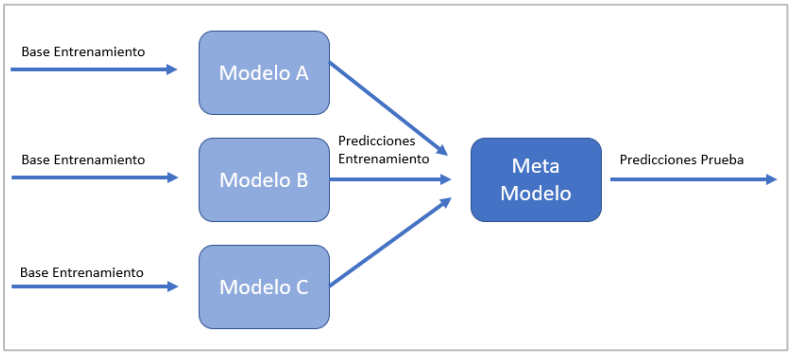
A grandes rasgos, las técnicas de Bagging y Boosting trabajan con las predicciones de un clasificador (que puede ser un ensamble en sí mismo como un Random Forest) y generan resultados bajo diferentes muestras de la base de entrenamiento, el primero realiza una votación de cada uno de los participantes y generaliza los errores en el total, mientras que Boosting va aprendiendo en cada partición de los errores anteriores y los mejora. Las técnicas de Stacking y Blending trabajan con las predicciones de dos o más clasificadores, se diferencian en que el primero toma todas las predicciones de la base de entrenamiento mientras que el segundo toma solo las predicciones de una parte de validación de esa base de entrenamiento.

Recapitulando el desarrollo de este proyecto, se han explicado dos procesos de predicciones sobre la misma base de datos, uno que toma como variable dependiente la etiqueta manual y otro que lo hace con la etiqueta automática de senticnet, partiendo entonces que el objetivo de utilizar una librería automatizada es poder transferir el conocimiento que esta aporta a la predicción final, y rescatando las ventajas y estructura que conlleva el etiquetado manual, se tomó la decisión de combinar ambos modelos para rescatar lo mejor de ambas partes. Es aquí entonces donde entra a jugar el mencionado

[29] Los hiper parámetros del Random Forest también fueron tuneados por lo que los resultados pueden tener diferentes ensambles con distintas cantidades de árboles.

ensamblaje, cómo el objetivo es utilizar las predicciones de dos clasificadores la decisión fue utilizar un proceso de Stacking[30].

En líneas generales el Stacking (también conocido como modelos apilados o meta-model ensemble) toma las predicciones de dos o más clasificadores, a los que denomina clasificadores de nivel 0, apila estás predicciones y genera a partir de estas un meta-modelo al que denomina clasificador de nivel 1, que por lo general es una regresión logística. Con estas predicciones nuevas y combinadas, testea contra una base de prueba.



Esta técnica puede evaluar más niveles, dependerá de la cantidad de modelos en el nivel 0 y las combinaciones que quieran hacerse, para este caso en particular, debido a que se cuenta con 2 clasificadores, el meta-modelo se construirá con sólo con dos niveles (0 y 1).

Volviendo a este caso en particular, anteriormente se explicó el proceso de construcción de las predicciones de cada modelo (manual y senticnet) y se plasmaron los resultados para cada dimensión adicional (tipo de comentario y cantidad de categorías), en el ensamblado se unen ambas predicciones para generar un clasificador final, sin embargo, previo a esto, se realizó un trabajo de balanceo de la base de datos combinada, El método de balanceo utilizado fue sobre muestreo que implica llevar a todos las categorías a la misma cantidad de aquella que más registros tiene, esta opción permite conservar la cantidad original de registros (agregando más) pero sin perder información posiblemente valiosa para el modelo como es el caso de un sub muestreo. Adicionalmente se utilizó el muestreo con reemplazo.

Con las predicciones apiladas y balanceadas, la construcción del clasificador final se realizó considerando un último tuneo de hiper parámetros entre los dos métodos más utilizados en Stacking para ensamblar, una regresión logística y un clasificador de Naive Bayes. El ensamble selecciona entonces el mejor modelo en cada caso, y esto último es muy importante ya que el ensamble también está sujeto a iteraciones debido a que recibe en cada caso las predicciones manuales de una base de datos manual que proviene de cierto nivel de etiquetado inicial, recordemos el proceso de sensibilización. En otras palabras, se tendrán tantos ensambles como particiones iniciales de etiqueta manual, cada ensamble

[30] La selección de Stacking sobre Blending busca dotar al modelo con un número mayor de predicciones al utilizar la base de entrenamiento completa y no sólo de una muestra de validación.

correrá con las diferentes iteraciones de la base manual, ensamblara con el único modelo de senticnet (no tiene particiones) y arrojará los resultados generalizados. Esto, adicionalmente para las dimensiones de parámetros (tipo comentario, categorías).

Cómo este es el resultado final de todos los procesos anteriores, los resultados que se presentan a continuación evalúan al modelo contra la base de prueba desde las siguientes métricas de clasificación:

- Accuracy: La exactitud del modelo en clasificar bien todas las categorías.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Recall: Exhaustividad, esta métrica mide la capacidad clasificar correctamente una categoría (1) con relación a todo lo que en realidad pertenecía a dicha categoría.

$$recall = \frac{TP}{TP + FN}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

- Precision: Esta métrica mide la habilidad del modelo de clasificar correctamente una categoría (1) con relación a todo lo que el modelo pronóstico era esa categoría.

$$precision = \frac{TP}{TP + FP}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

- F-Score: Conocida como la métrica de F1-Score, vincula tanto el resultado de precisión y exhaustividad.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Como en este caso se cuenta con un variable dependiente es multi -categoría, él cálculo de Recall y precisión se lleva a cabo con la función de *sklearn* que estima los valores para cada una de las 3 categorías y las pondera en un promedio, esto puede derivar en valores de F-score que no se encuentran entre la precisión y el Recall, sin embargo, es el método recomendado para trata con etiquetas de múltiples clases.

Para el comentario lematizado, los resultados en cada categoría son los siguientes:

Comentario Lematizado		**3 categorías**			
Etiquetado Inicial	Meta clasificador	Accuracy	Recall	Precision	F-Score
10%	Logistic Regresion	0,5944	0,5733	0,5924	0,5833
20%	Logistic Regresion	0,6197	0,5992	0,6215	0,6084
30%	Logistic Regresion	0,6228	0,6023	0,6268	0,6121

40%	Logistic Regresion	0,6255	0,6088	0,6286	0,6173
50%	Logistic Regresion	0,6331	0,6160	0,6382	0,6234
60%	Logistic Regresion	0,6280	0,6102	0,6328	0,6190
70%	Logistic Regresion	0,6455	0,6299	0,6522	0,6368
80%	Logistic Regresion	0,6373	0,6232	0,6413	0,6292
90%	Logistic Regresion	0,6404	0,6259	0,6478	0,6315
99%	Logistic Regresion	0,6445	0,6285	0,6528	0,6360
		2 categorías			
Etiquetado Inicial	Meta clasificador	Accuracy	Recall	Precision	F-Score
10%	Logistic Regresion	0,7883	0,6405	0,7669	0,7677
20%	Logistic Regresion	0,8027	0,6631	0,7855	0,7845
30%	Logistic Regresion	0,8047	0,6691	0,7883	0,7878
40%	Logistic Regresion	0,8167	0,6964	0,8038	0,8043
50%	Logistic Regresion	0,8179	0,6930	0,8048	0,8040
60%	Logistic Regresion	0,8218	0,7026	0,8097	0,8095
70%	Logistic Regresion	0,8207	0,7097	0,8093	0,8108
80%	Logistic Regresion	0,8269	0,7134	0,8160	0,8160
90%	Logistic Regresion	0,8234	0,7090	0,8120	0,8124
99%	Logistic Regresion	0,8274	0,7188	0,8169	0,8178

Para el comentario que sólo elimina stopwords, los resultados en cada categoría son los siguientes:

Comentario no stopwords		3 categorías			
Etiquetado Inicial	Meta clasificador	Accuracy	Recall	Precision	F-Score
10%	Logistic Regresion	0,6032	0,5757	0,5993	0,5909
20%	Logistic Regresion	0,6153	0,5915	0,6161	0,6059
30%	Logistic Regresion	0,6281	0,6047	0,6282	0,6187
40%	Logistic Regresion	0,6278	0,6045	0,6313	0,6174
50%	Logistic Regresion	0,6382	0,6196	0,6397	0,6311
60%	Logistic Regresion	0,6327	0,6138	0,6365	0,6240
70%	Logistic Regresion	0,6388	0,6182	0,6443	0,6297
80%	Logistic Regresion	0,6399	0,6239	0,6416	0,6352
90%	Logistic Regresion	0,6401	0,6210	0,6451	0,6328
99%	Logistic Regresion	0,6383	0,6216	0,6439	0,6309
		2 categorías			
Etiquetado Inicial	Meta clasificador	Accuracy	Recall	Precision	F-Score
10%	Logistic Regresion	0,7849	0,6234	0,7605	0,7586
20%	Logistic Regresion	0,7949	0,6507	0,7755	0,7754
30%	Logistic Regresion	0,8087	0,6638	0,7932	0,7884
40%	Logistic Regresion	0,8108	0,6792	0,7960	0,7951
50%	Logistic Regresion	0,8185	0,6844	0,8056	0,8017
60%	Logistic Regresion	0,8222	0,6998	0,8101	0,8090
70%	Logistic Regresion	0,8267	0,7076	0,8155	0,8143
80%	Logistic Regresion	0,8290	0,7084	0,8182	0,8161
90%	Logistic Regresion	0,8290	0,7103	0,8182	0,8166

99%	Logistic Regresion	0,8315	0,7169	0,8212	0,8202

V. Evaluación de Resultados

Las conclusiones que se describirán en este apartado son el resultado de la construcción de este modelo en particular, de las decisiones que se tomaron en cada uno de los procesos detallados antes y corresponden entonces a la combinación de todos estos factores, pueden ser de referencia para trabajos similares (como hará CAOBA con la llegada de nuevos datasets) pero no son directamente extrapolables ni asumibles como la mejor opción para todos los casos.

En la descripción de los resultados presentados hasta el momento se evidencia una gran cantidad de salidas vinculadas a cada una parte del modelo y al modelo de ensamble final, sin embargo, en dicha descripción el modelo final se evalúo en relación a las métricas de desempeño propias de un modelo de clasificación, y se expusieron también los resultados de estas métricas para dos tipos de parámetros mencionados recurrentemente, el tipo de comentario evaluado y la cantidad de categorías que tiene el campo de polaridad.

Evaluando primero los resultados de los modelos por separado, es interesando ver como el Accuracy del modelo de etiquetado manual es mayor en todos los casos (para todas las particiones) con relación a las predicciones de senticnet, como se muestra en la siguiente tabla evaluando el Accuracy promedio (para las particiones manuales):

Accuracy promedio	3 categorías		2 categorías	
	Lematizado	Stopwords	Lematizado	Stopwords
Manual	0,6787	0,6775	0,8240	0,8195
Senticnet	0,5465	0,5438	0,7597	0,7599

Para cada parámetro de la tabla anterior, senticnet está por debajo del pronóstico manual en cerca del 20% del Accuracy, a simple vista pareciera que es posible obtener predicciones buenas (superiores al 65%) con solo probar el modelo manual, sin embargo, recordemos que el objetivo de este proyecto es tomar el aprendizaje de ambas metodologías y reflejarlas en el ensamble, además la situación de senticnet también puede verse afectada por el dominio de la data a evaluar (parte de la data pertenece a una entidad financiera) y puede perder exactitud con algunos términos muy particulares.

Con esto sobre la mesa, pasemos a evaluar entonces los resultados del ensamble, primero de forma agrupada y luego en detalle. De forma agrupada, se calculó la desviación estándar para cada partición de etiquetado obteniendo que se mantienen en rango entre el 1% y 3%.

Desviación Estándar partición	3 categorías		2 categorías	
	Lematizado	Stopwords	Lematizado	Stopwords
Accuracy	1,51%	1,23%	1,26%	1,57%

Recall	1,73%	1,56%	2,56%	3,05%
Precision	1,80%	1,48%	1,61%	2,02%
F-Score	1,60%	1,41%	1,63%	2,03%

En forma agrupada para resolver la pregunta sobre que tipo de comentario y cuantas categorías en la polaridad serían lo más convenientes para este modelo.

Métricas promedio	3 categorías		2 categorías	
	Lematizado	Stopwords	Lematizado	Stopwords
Accuracy	0,6291	0,6302	0,8150	0,8156
Recall	0,6117	0,6094	0,6916	0,6844
Precision	0,6334	0,6326	0,8013	0,8014
F-Score	0,6197	0,6217	0,8015	0,7995

Evaluando primero el parámetro de tipología de comentario, tanto para la polaridad vista desde tres como dos categorías, los resultados entre un comentario lematizado (que incluye también el proceso de quitar stopwords) en comparación con el comentario que sólo realiza la limpieza de stopwords, son muy similares, sin superar para ninguna de las métricas el 2% entre ambos, tanto para el promedio, tanto como si se miran los resultados para cada partición del apartado anterior. Los comentarios de stopwords parecen comportarse mejor con relación al Accuracy, pero un poco peor con relación al Recall y al Precision, esto podría ser un justificante para preferir usar el comentario lematizado (Recall y Precision) terminan viendo que tan bueno es el modelo cometiendo errores tipo I y tipo II, sin embargo, esta decisión puede estar más enfocada a un tema de negocio.

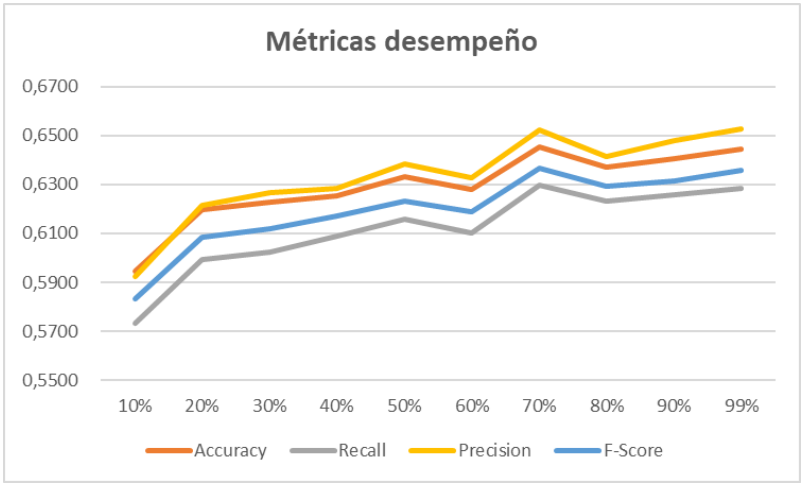
El comentario lematizado, como se describió en apartados anteriores, conlleva un proceso lingüístico mucho más complejo que una limpieza sencilla, y este proceso implica que se buscan asociar palabras que comparten la misma familia, esto ayuda al modelo a ser más preciso en la asignación del sentimiento ya que la reduce el tamaño de la matriz de entrada, pero para un cliente, por ejemplo, también es más sencillo entender la polaridad asociado a un lema que a todas las palabras relacionadas a su familia. Por lo tanto, sería la recomendación de este trabajo, trabajar con el comentario con todo el pre-procesamiento hasta llegar al lema de la palabra a evaluar.

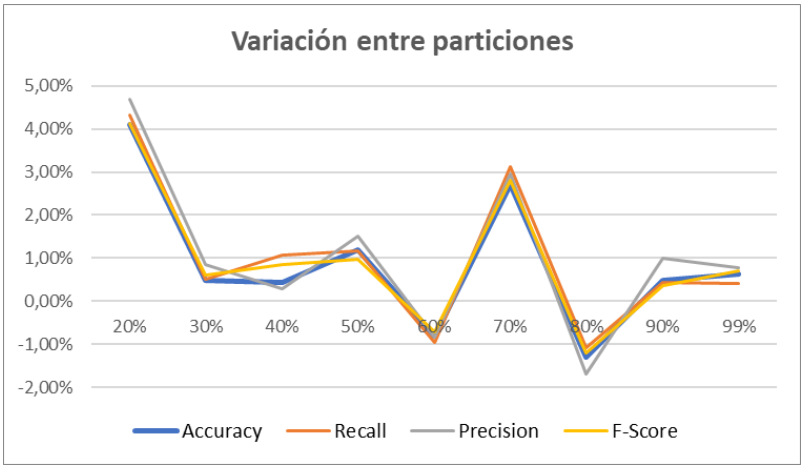
Continuando con las conclusiones sobre los resultados, evaluemos entonces los resultados para las categorías de la polaridad, es evidente, que cuando el modelo entrena con sólo dos categorías: positiva y negativa, se obtienen mejores resultados en comparación a cuando se involucra una tercera categoría, la neutra. Entonces a simple vista estos resultados podrían llevar rápidamente a la conclusión de proponer realizar este ejercicio con solo 2 categorías, sin embargo, esto puede tener ciertos impactos asociados. Lo primero, el ejercicio con 2 categorías supera al de 3 en métricas como el Accuracy y la Precision, pero cae drásticamente cuando se trata del Recall, lo que se puede concluir de estos resultados es que el modelo con 2 categorías “asume” un error de clasificación (al perder la categoría neutra) como un acierto de clasificación al otorgarla a una de las dos opciones, lo cual

claramente no es correcto ya que realmente hay sentencias que no indican una polaridad tan marcada como buena o mala.

De todas maneras, la respuesta no es que deba hacerse con una opción o la otra, en realidad la conclusión es que se debe tener cuidado con la interpretación de los resultados entre ambas. De igual manera, aunque metodológicamente parece tener más robustez tener 3 categorías, es posible que el cliente que llega a CAOBA en busca de ahorrar tiempos y por facilidad de interpretación quiera visualizar sus resultados en los dos extremos del espectro y esto es válido, sin embargo, debe estar siempre sobre la mesa que la categoría a la que se le asignen las sentencias neutras no está revelando la realidad en términos, por ejemplo de las palabras que más se correlacionan con ella.

Tomando como referencia los resultados del modelo con comentario lematizado y con 3 categorías de polaridad, se evaluaron las diferencias de cada iteración para sensibilización del etiquetado manual de entrada, validando el cambio relativo en relación con la iteración anterior, como se muestra en el siguiente gráfico:





Entre mayor cantidad de etiquetas de entrada las métricas aumentan por lo general, sin embargo, existen puntos de porcentaje de etiquetado que generan ganancias más significativas (por arriba del 1%) que otros, por ejemplo, las particiones en 20, 50 y 70 para este caso. En todo caso, esta conclusión y estos puntos importantes son el resultado de este modelo en particular y todas las decisiones tomadas a lo largo de todo el proceso, y no son necesariamente extrapolables para otro dataset o para otras metodologías.

VI. Despliegue

El despliegue de este modelo se puede explicar a través de la aplicación que podrá darle CAOBA a los resultados y a la metodología aquí planteada. Antes de entrar al detalle de este apartado, es importante recalcar que los resultados de este proyecto corresponden a las decisiones de analítica tomadas para dar cumplimiento a los objetivos planteados y en aras de encontrar resultados que CAOBA pudiese aplicar desde la perspectiva de negocio. El mantenimiento de estos modelos expuestos se deberá ajustar por parte de la Entidad conforme sea necesario.

Retomando entonces los objetivos de este proyecto, pueden expresar a grandes rasgos como, por un lado, encontrar una metodología que pueda ser utilizada por CAOBA para ofrecer un servicio de etiquetado para modelos de análisis de sentimientos que sea eficiente, de calidad y con beneficios en reducción de costos y tiempos para el centro, y por otro el objetivo de encontrar una forma de traducir esa eficiencia a los clientes externos. Este modelo visto desde su explicación se apoya en metodologías de machine learning para reducir el tiempo de etiquetado manual, maximizar la ganancia por hacer esta tarea y generar un modelo que combine además una parte automática del etiquetado para mejorar eficiencia, sin embargo, su implementación en el plano tangible permitirá que esta construcción no sólo quede en el ámbito teórico.

Con CAOBA ofreciendo el módulo completo de Data Label Server, Modelo y Visualización debe venir previo una explicación para los clientes externos, de modo que entiendan el universo completo que implica realizar un modelo de NLP, pero sobre todo para que entiendan la importancia del etiquetado de una base de datos por parte de expertos que justifique realizar esta tarea con CAOBA. Para dar cumplimiento con esto, uno de los entregables de este proyecto es el “Manual Metodológico de Data Label Server”[31] que recoge entre otras cosas una introducción de CAOBA y al servicio y un detalle de conceptos previos básicos de NLP, detalle del proceso de etiquetado y explicación de los principios clave que puede ofrecer CAOBA con este servicio.

[31] Para ver el detalle del manual, referirse a la versión completa en “2-Manual Metodológico Data Label Server CAOBA”

Alianza CAOBA
Centro de Excelencia y Apropiación en Big Data y Data Analytics
http://www.alianzacaoba.co

MANUAL METODOLÓGICO DATA LABEL SERVER

La información, la herramienta más valiosa que han tenido las organizaciones a lo largo de los años para tomar decisiones respecto a qué, cómo, a quién y cuándo ofrecer sus productos y servicios, está encontrando ahora un nuevo aliado, la inteligencia artificial.

Gracias a la tecnología y a las nuevas formas de interacción con los usuarios, las empresas en la actualidad cuentan con un poderoso activo que empieza a cobrar relevancia para entender a los clientes, la data individual de cada uno de ellos, para detectar comportamientos, patrones, tendencias, y hasta sentimientos de estos.

Este manual metodológico, les permitirá a las empresas enfocadas en recaudar información de servicio al cliente, el valioso proceso de entender y pre procesar la data con la que cuentan para explotarla con la ayuda de Machine Learning y de CAOBA.

TABLA DE CONTENIDO

1. ¿QUÉ ES CAOBA?
2. CONCEPTOS INICIALES
3. DATA LABELLIN - ETIQUETADO
4. PRINCIPIOS CLAVES
   a. CALIDAD
   b. EFICIENCIA
   c. ESCALABILIDAD
   d. SEGURIDAD
   e. TECNOLOGÍA
5. DATA LABEL SERVER - CAOBA

PRINCIPIOS CLAVES

Entendiendo ya que el etiquetado manual de la data a procesar cumple un rol muy importante dentro de cualquier algoritmo a probar dentro de NLP y en general en cualquier rama de Machine Learning, y conociendo que este trabajo puede ser ejecutado de diferentes maneras, planteemos el escenario en que, con una ayuda inicial de contexto de la data por parte de la empresa, CAOBA se encarga del proceso de etiquetado combinando lo mejor de personas, procesos y tecnología para lograrlo.

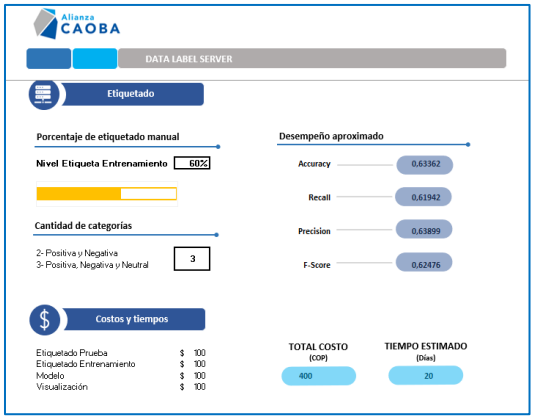
Por qué esto es una excelente opción lo detallamos a continuación, bajo lo que llamamos, los principios claves del etiquetado de la data para modelar con Machine Learning.

El proceso de etiquetado manual que requiere la data de las empresas es un trabajo muy específico que debe recorrer los siguientes pasos, por nombrar los principales:

- *Características del texto:* Entender la longitud de los textos a analizar, no es lo mismo la revisión de un tweet, un chat o un mensaje de correo.
- *Segmentación:* Identificar si el texto debe ser analizado por partes o su totalidad, por ejemplo, un correo electrónico podría contener en diferentes párrafos comentarios buenos como malos, lo que puede concluir en resultados no concluyentes.
- *Proceso Lingüístico:* Hacer la lectura de las unidades de texto definidas, donde aplicará los conocimientos lingüísticos presentados anteriormente, pero adicionalmente debe identificar palabras mal escritas, con errores de ortografía, modismos o palabras particulares del lenguaje o contexto.
- *Entidades:* Reconocer las entidades del texto, como, por ejemplo, personas, lugares, objetos, monedas, etc.
- *Ambigüedad:* Reconocer la ambigüedad del texto es una de las tareas más difíciles de trasladar a una máquina, palabras con connotación positiva y negativa pueden aparecer juntas en expresiones de ironía, y el etiquetador manual debe poder identificar el verdadero significado de la frase.
- *Correlación:* Reconocer cuando en un texto se hace referencia a entidades indirectamente (anáforas), el uso de expresiones como "aquel, el primero, el mismo" son un ejemplo de esta tarea que sólo viendo la totalidad del texto (frase, oración) puede tener sentido.
- *Objetividad:* Asignar la polaridad tras este análisis exhaustivo y mantener objetividad entre sentencias similares o determinar desempate de algunos términos comunes.

Este proceso además debe ir acompañado de características por parte del etiquetador para garantizar la calidad de la información a entrar en el modelo:

El siguiente entregable para desplegar este modelo, es una herramienta ofimática para que CAOBA pueda dar visibilidad de este servicio. Esta herramienta contiene una plantilla inicial para que el cliente proporcione un breve contexto de la data y seleccione el nivel de etiquetado que desea realizar de forma manual, junto con la cantidad de categorías de polaridad que desea sean incluidas en dicho etiquetado. Adicionalmente en esta plantilla se le mostrarán los resultados de las métricas del modelo construido para dar una orientación de resultados estimados bajo las condiciones seleccionadas. Finalmente, esta plantilla mostrará al cliente los costos y tiempos asociados a su decisión. Con estas opciones el cliente o CAOBA mismo podrá decidir la combinación que mejor se ajuste a su base en términos de desempeño, tiempo y costo.



Esta herramienta estará soportada por la matriz resultantes del modelo construido, en donde para cada combinación de porcentaje de etiquetado y cantidad de etiquetas mostrará las métricas de Accuracy, Recall, precisión y F-Score, es importante hacer claridad con el cliente que los resultados aquí mostrados son una estimación basada en el entrenamiento de una base de servicio al cliente que podría ajustarse al dominio del nuevo cliente, pero cuyos valores finales dependerán exclusivamente de la base en particular. Esta herramienta también contará internamente con la matriz de costos construida para este proyecto en el que se detalla a partir de la cantidad de registros que tenga la base a evaluar, el valor por el etiquetado requerido necesariamente para la base de prueba (20%) y el valor asociado a la partición que se desea hacer para la base de entrenamiento; adicionalmente se dejará la opción (a evaluar por CAOBA) sobre los costos asociados a la ejecución del modelo, y el costo de la visualización de los resultados.

Finalmente, esta herramienta contiene también un tablero de resultados, una vez CAOBA aplique la metodología para desarrollar el modelo de análisis de sentimiento, podrá

mostrarle al cliente indicadores clave como nivel de polaridad de toda la base, palabras con mayor frecuencia para cada una de las polaridades, y bigramas más correlacionados con cada valor. Esto sería el equivalente al módulo de visualización de la solución completa.

En relación con los costos que se podrán visualizar en esta herramienta, y tras evaluar las tarifas de la industria para servicios de etiquetado similares al que ofrecería CAOBA, la propuesta se movería en los siguientes rangos:

Complejidad	Tiempo asociado	Costo (USD)	Cota
100 sentencias cortas	1 hora	2.78	mínima
100 sentencias largas o complejas	1.5 hora	7.64	máxima

De acuerdo con benchmark presentado en la primera sección de este documento, el rango de precio de las compañías listadas varía entre 5 y 12 dólares por la tarea de etiquetado de 100 sentencias, sin embargo, todas esas empresas son extranjeras y esta propuesta debe ajustarse un poco. Empezando por el cálculo del tiempo para las 100 sentencias, es en general lo que otras empresas emplean, pero sobre todo es la referencia que se tiene de CAOBA llevando a cabo esta tarea y del trabajo de este proyecto etiquetando la base de CCD, el tiempo se referenció en 1 hora aproximadamente para sentencias cortas y cerca de 1.5 horas para sentencias más complejas, textos más largos, folios y documentos más extensos en general.

El precio asociado al tiempo de etiquetado se calculó consideran dos cosas, primero el perfil requerido para el trabajo, de acuerdo con la complejidad del texto y como se ha detallado en este documento y en el manual de usuario de la herramienta de Data Label Server, el etiquetado manual requiere de ciertas capacidades y experiencia, entonces se plantean dos perfiles, solo para fijar una cota mínima y máxima en la que se puede encontrar el cobro final. Para etiquetado manual de sentencias cortas y sencillas, se calculó el costo de hora de salario para un profesional[32] y para sentencias más complejas se estimó un máximo tomando las referencias de mercado y el ingreso promedio según oferta laboral para un perfil Científico de Datos.[33]

Por último, para completar el despliegue de esta solución, se entregará a CAOBA un código que explica el paso a paso para realizar el proceso de pre procesamiento, y construcción de los modelos de iteración de porcentajes, predicciones de etiqueta manual y senticnet y ensamble final que se construyeron en este proyecto, para que pueda aplicarlo ante la llegada de un cliente y base nuevos, además tiene una sección para generar las tablas que permiten generar los indicadores del tablero de control de la herramienta. Este código contará además con el detalle de cómo estructurar la data para que pueda ejecutarse sin

[32] https://www.larepublica.co/economia/los-profesionales-ganan-71-mas-que-personas-con-personas-con-basica-primaria-2960985
[33] https://www.elempleo.com/co/ofertas-empleo/trabajo-cientifico-de-datos

problema. Toda la codificación y salidas del modelo aquí construido también serán entregados a CAOBA.

Las recomendaciones sobre el despliegue de este trabajo son, primero contextualizar a los clientes externos de CAOBA sobre las principios de este modelo en términos de dominio de información (áreas de servicio al cliente), aclarando que los resultados aquí plasmados son el reflejo de una metodología específica, conforme esta metodología y sus resultados sean puesto en práctica por CAOBA, este último puede ver necesario calibrar nuevamente el modelo con información reciente o que incluya otros sectores o industrias, así como mantener actualizado al modelo sobre novedades de la librería de senticnet, entre otras consideraciones, velando siempre por cumplir con los objetivos de generar valorar para los clientes a través de metodologías de big data y machine learning.