

**HERRAMIENTA DE INTELIGENCIA DE NEGOCIOS PARA EL ANÁLISIS DE INFORMACIÓN ESTRUCTURADA Y NO ESTRUCTURADA EN EL ÁREA DE SERVICIO AL CLIENTE**

**ESTEBAN ANTONIO LLANOS MILLÁN**

**UNIVERSIDAD DEL VALLE**

**FACULTAD DE INGENIERÍA**

**ESCUELA DE INGENIERÍA DE SISTEMAS**

**Y CIENCIAS DE LA COMPUTACIÓN**

**CALI, VALLE DEL CAUCA**

**2018**



**HERRAMIENTA DE INTELIGENCIA DE NEGOCIOS PARA EL ANÁLISIS DE INFORMACIÓN ESTRUCTURADA Y NO ESTRUCTURADA EN EL ÁREA DE SERVICIO AL CLIENTE**

**ESTEBAN ANTONIO LLANOS MILLÁN**

**DIRECTOR**

**OSWALDO SOLARTE PABÓN Ms.C.,**

**TRABAJO DE GRADO PARA OPTAR POR EL TÍTULO DE:**

**INGENIERO DE SISTEMAS**

**UNIVERSIDAD DEL VALLE**

**FACULTAD DE INGENIERÍA**

**ESCUELA DE INGENIERÍA DE SISTEMAS**

**Y CIENCIAS DE LA COMPUTACIÓN**

**CALI, VALLE DEL CAUCA**

**2018**

**NOTA DE ACEPTACIÓN**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**OSWALDO SOLARTE PABÓN – DIRECTOR**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**JURADO**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**JURADO**

**Santiago de Cali, 2018**

***DEDICATORIAS***

Pendiente…

Esteban Antonio Llanos

Tabla de contenido

**Resumen**1

**Introducción**1

**Planteamiento del Problema**1

**Objetivos**2

*Objetivo General*3

*Objetivos Específicos*3

**Marco de Referencia**3

***Marco Teórico***3

*Selecc*ión *de Herramienta de Análisis*3

*Características de las Herramientas Encontradas*5

***Marco Conceptual***7

***Estado del Arte***9

*Análisis de Sentimientos*9

*Inteligencia Sobre el Cliente*10

*Integración de Información Estructurada y No Estructurada*10

*Bodegas de Datos Orientadas al Servicio al Cliente*11

**Alcances y Limitaciones**12

**Metodología**13

*Actividades por Objetivo*14

**Cronograma**15

**Presupuesto**16

*Presupuesto General*16

*Presupuesto Detallado*17

**Bibliografía**18

Lista de Figuras

Propuesta de Solución1

Lista de Tablas

Resumen1

RESUMEN

A lo largo del presente trabajo se buscó desarrollar una herramienta que permitiera encontrar el valor oculto en los grandes volúmenes de información almacenados en el área de servicio al cliente, desde información captada personalmente de los clientes a través de la atención brindada por trabajadores de la empresa hasta información obtenida de los clientes en redes sociales y por medio de llamadas telefónicas recibidas por el Call-Center.

Para esto, en primera instancia se consideraron varias herramientas existentes en el mercado y que simplifican la tarea de analizar información no estructurada a través de técnicas de Procesamiento de Lenguaje Natural. La selección de entre todas las herramientas encontradas en la red pasó por diversas consideraciones entre las que se dio mayor prioridad a tres factores principales: costo, robustez y pertinencia de acuerdo al problema abordado, el cual será descrito en mayor detalle más adelante. La robustez de la herramienta fue medida en términos de las opciones de desarrollo que brindara y la documentación con la que se contara. Por otra parte, saber si la herramienta era útil para el manejo y análisis de información propia del área de servicio al cliente dependería de que la estructura proveída por el software encontrado soportara los procesos necesarios para dichas tareas.

Una vez hecho esto fue necesario identificar la estructura de la Bodega de Datos que se usaría para almacenar todo lo extraído de las fuentes de información no estructurada (pero ya procesada) y donde sería relacionada con información estructurada para llevar a cabo tareas propias de Minería de Datos que evidenciaran puntualmente los resultados obtenidos del proceso.

Posterior a esta labor se inició el desarrollo de la herramienta de análisis con el Framework seleccionado siguiendo lineamientos que permitieran ejecutar una tarea de minería de opinión y análisis de sentimientos precisa y que diera resultados óptimos en términos de detección de intención y correcta interpretación del mensaje leído. Los resultados entregados por la herramienta debieron ser almacenados en una base de datos temporal con el fin de clasificar la información, hasta ese momento, no estructurada.

Una vez migrada y procesada toda la información concerniente al área de servicio al cliente de la organización sería necesario poder visualizar los resultados del análisis final sobre lo almacenado en la bodega de datos. Para esto se exploró la red en busca de una API adaptable al Framework, al lenguaje y a la herramienta de análisis con el fin de permitir al usuario el despliegue de reportes gráficos que simplificaran la interpretación de los resultados finales.

INTRODUCCIÓN

En el constante crecimiento de las redes sociales y por la imperante necesidad de sacar provecho a la información que se genera rápidamente en los tantos y tan variados medios que existen en la actualidad, se hace necesario contar con herramientas que puedan explotar apropiadamente estas inmensas fuentes de datos para desarrollar estrategias de mercadeo, planes de negocio, etc. acordes a la realidad que se manifiesta por medio de esa información que es producto de análisis.

Contar con una amplia gama de posibilidades en el análisis y el procesamiento de esa información que se extrae a diario es fundamental, ya que el rápido desarrollo de las tecnologías de la información demanda que las corporaciones de todo tipo se adapten a estas nuevas tendencias en pro de mantenerse a la vanguardia del mundo de los negocios que se transforma en torno a estas nuevas formas de ver el mundo.

El BIG DATA relacionado a los procesos de atención al cliente ha sido desarrollado con el objetivo de mejorar los servicios prestados al mismo por medio de la predicción de comportamientos y la definición de patrones presentes en la información extraída de estos medios de comunicación en los cuales interactúa quien vende un producto u ofrece un servicio, con quienes han adquirido su producto o servicio o inclusive con potenciales clientes a futuro. La implementación de elementos propios del **Manejo de Relaciones con el Cliente** (en adelante, **CRM**) para mejorar el servicio brindado a los clientes se ha convertido en un tema prioritario para las empresas de toda índole, ya que servicios como el Social CRM representan una ventaja frente a los competidores

El uso de algoritmos para el Procesamiento del Lenguaje Natural (en adelante PLN) y la Minería de Texto sobre estas fuentes de Información No Estructurada (descritas más adelante en este documento) permite que, junto a métodos de análisis de la información, se simplifique y se potencialice la búsqueda de mejores elementos en la proyección empresarial deseada.

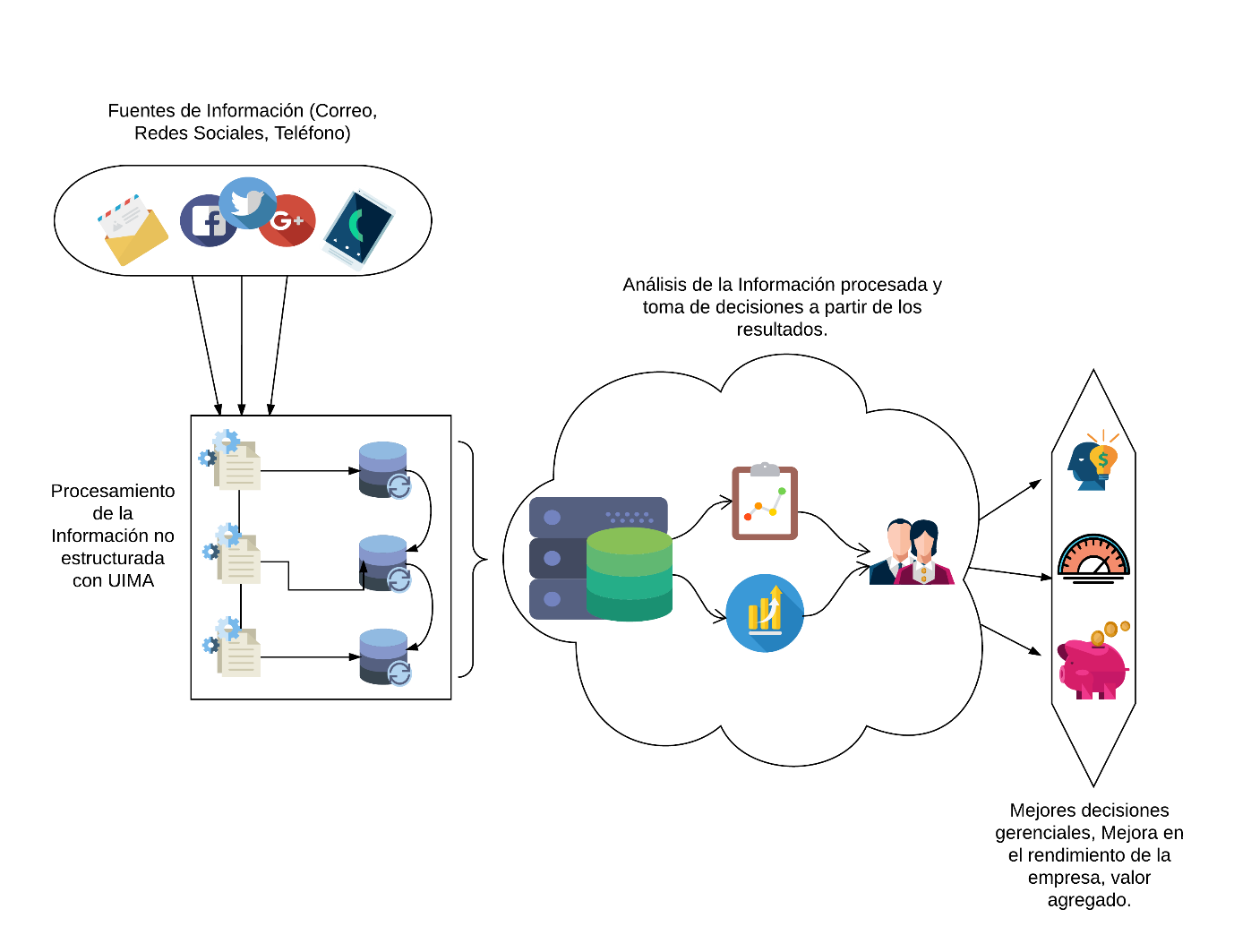
A través de este proyecto se pretende profundizar en la Minería de Datos en torno a la información no estructurada manejada por las empresas en torno al servicio al cliente y que suele ser ignorada por su propia naturaleza carente de una apropiada categorización y de características identificables que permitan su procesamiento y posterior análisis. Poder explorar este amplio universo de conocimiento que se esconde detrás de la información que se genera en la red es una necesidad que no da espera, ya que el crecimiento de la información que se almacena a diario es exponencial. Un estudio realizado por **IBM**[[1]](#footnote-1) apunta que diariamente se generan cerca de **2,5 quintillones de bytes** de información y considerando que el mismo estudio plantea que, actualmente el **90%** de la toda la información almacenada en internet se creó en los últimos **2 años** podríamos concluir con facilidad que la cantidad de data creada cada día irá en aumento, así como van en aumento las tecnologías para la captación de datos de los usuarios de productos y servicios.

**PLANTEAMIENTO DEL PROBLEMA**

Conocer la opinión de los clientes en torno a una organización y a los servicios que esta ofrece, es de gran utilidad a la hora de buscar y desarrollar estrategias más innovadoras y efectivas que le permitan a la empresa mejorar sus procesos y así mismo los servicios y productos que ofrece. Este análisis de información nos permite también encontrar los quiebres y puntos débiles de la organización en términos de mercadeo, publicidad, servicio al cliente, etc.

Tener en cuenta la percepción que tiene el usuario final nos permite considerar una serie de elementos que se escapan de la perspectiva interna de la marca y llegar a contemplar más ampliamente todas las aristas del problema, sin embargo, ¿cómo se puede conocer de manera más asertiva la opinión de los clientes respecto a nuestra marca?, ¿cómo poder analizar dichas opiniones a partir de los comentarios dejados por los usuarios finales en redes sociales, a través de llamadas telefónicas entre otras fuentes de información no estructurada para poder generar indicadores que permitan explorar nuevas posibilidades de negocio? Esto nos lleva a pensar en la elaboración de una herramienta que, a través de Algoritmos de PLN y de estándares para el análisis de información elaborados por distintas entidades, permitan la lectura y procesamiento de todas estas fuentes de información que puedan posteriormente ser analizados haciendo uso de herramientas de Minería de Texto y de Big Data.

Esta herramienta a desarrollar nos va a permitir llevar a cabo todo el proceso de ETL (Extracción, Transformación y Carga) sobre las fuentes de texto obtenidas a partir del levantamiento de información de la empresa para su posterior análisis y visualización en la Bodega de Datos.



***Figura 1.*** *Representación Gráfica de Propuesta de Solución*

**OBJETIVOS**

**Objetivo General:**

* Desarrollar una aplicación para analizar información estructurada y no estructurada en el área de servicio al cliente usando técnicas de inteligencia de negocios y bodegas de datos.

**Objetivos Específicos:**

● Revisar algunas herramientas y propuestas orientadas a la integración y análisis de información estructurada y no estructurada en una bodega de datos.

● Diseñar un esquema de bodega de datos orientada al servicio al cliente para la integración de la información estructurada y no estructurada.

● Implementar un módulo que permita extraer, procesar y cargar información estructurada y no estructurada en la bodega de datos

● Implementar un módulo que facilite la visualización de la información procesada a partir de alguna herramienta de generación de reportes ya existente.

● Diseñar un plan de pruebas que permita validar la herramienta implementada y verificar el cumplimiento de los requerimientos.

**MARCO DE REFERENCIA**

**Marco Teórico**

Para el actual proyecto es necesario considerar las repercusiones en el ámbito comercial y los resultados que produce la implementación de este sistema dentro de los procesos de mercadeo y proyección ejecutados por la organización que haga uso de un prototipo que permita el análisis de datos no estructurados como herramienta para el desarrollo de estrategias empresariales y para sustentar la toma de decisiones a nivel gerencial. Para la comprensión del problema de investigación y la adecuada lectura de los elementos a desarrollar a lo largo de este documento es necesario establecer ciertos conceptos que limitarán la ejecución de la herramienta con la cual se pretende explotar el potencial estratégico de la información contenida en medios que no suelen ser explorados a fondo como redes sociales, correos electrónicos, llamadas telefónicas, etc.

Los datos no estructurados componen un cúmulo de información en movimiento que no suelen ser almacenados y que fluyen en inmensas cantidades a través de diversos medios. Estos datos se generan a muy altas velocidades y provienen de múltiples fuentes, lo cual sumado a la gran cantidad de ruido que se genera en dichas fuentes, dificulta su procesamiento para su posterior análisis.

***Selección de Herramienta de Análisis***

Si bien existen diversas plataformas para el análisis de texto y para la ejecución de tareas de Procesamiento de Lenguaje Natural, el problema específico a abordar requiere del uso de una herramienta más robusta que complemente las tareas de Text Mining y PLN con opciones de interoperabilidad y escalamiento de la información, y que adicionalmente permita el manejo de los datos resultantes del análisis de los elementos a explorar (Texto, Audio, Vídeo, Imágenes, etc.). En este sentido la **Organización para el Avance en los Estándares sobre Información Estructurada** (**OASIS** por sus siglas en inglés)[[2]](#footnote-2) desarrolló el estándar de **Arquitectura para el Manejo de Información No Estructurada** (**UIMA**)[[3]](#footnote-3), con el fin de establecer un marco de referencia que ofreciera las condiciones base para el desarrollo de Aplicaciones **UIM** (Aplicaciones para el Manejo de Información no Estructurada)**.**

Con el fin de establecer qué plataforma sería la más apropiada para la construcción de la herramienta de análisis en cuestión, se realizó una comparativa entre tres proyectos desarrollados por Apache**: Apache OpenNLP, Apache Gate, Apache UIMA.**

* **Apache OpenNLP (Open Natural Language Processing)** (Apache Software Foundation - OPENNLP, 2017) es un proyecto que contiene un conjunto de herramientas de aprendizaje automático que soporta tareas comunes como tokenización, segmentación de oraciones, POS-tagging, NER (Identificador de Entidades), chunking (análisis sintáctico superficial por bloques), análisis sintáctico, etc.

El conjunto de herramientas que **OpenNLP** provee trabaja de manera independiente al idioma, se le puede entrenar en el idioma que se requiera y para algunos idiomas específicos ofrece modelos pre entrenados. Cuenta con una **API** para la integración con otras herramientas de cada una de sus funcionalidades, así como también con una interfaz de línea de comandos. La biblioteca está hecha en **Java** y es accesible desde el repositorio público de Maven. Existen también facilidades para utilizarlo desde **.NET**.

* **Apache GATE (General Architecture for Text Engineering)** (Universidad de Sheffield, 2017) es una suite de herramientas de código abierto en Java para tareas de PLN. Se compara usualmente con NLTK (Natural Language Toolkit). Es una infraestructura para desarrollar y desplegar componentes que procesan lenguaje natural. Contiene un entorno de desarrollo integrado (**IDE**) propio **(GATE Developer)** para la construcción de componentes de software para la plataforma, ayuda en la construcción de máquinas de estados, support vector machines y otras estructuras de datos complejas y permite la visualización de los datos procesados. La aplicación web GATE Teamware provee un ambiente colaborativo para realizar anotación semántica. Entre las prestaciones, la más importante es la arquitectura, la cual permite una organización de alto nivel de composición de componentes de PLN. Entre los lenguajes soportados se encuentra el español.
* **Apache UIMA (Unstructured Information Management Architecture)** (Apache Software Foundation - UIMA, 2017) es un Framework que permite la integración del análisis de texto u otros tipos de información no estructurada. Tiene implementaciones en Java, C++ entre otros lenguajes. Funciona haciendo uso de Anotaciones Stand-Off lo cual permite no modificar el texto que está siendo analizado a diferencia de las Anotaciones In-Line que pueden desordenarse rápidamente.

Algunas de las ventajas más remarcables en **Apache UIMA** son: la implementación de una **Estructura Común de Análisis** que permite la indexación de las anotaciones encontradas en los artefactos analizados, así como también proporciona una mayor eficiencia en el uso de memoria. Del mismo modo el manejo de **Sistemas de Tipos** da al desarrollador una estructura de control de tipos más estable y ordenada.

***Características de las Herramientas Encontradas***

* **Apache GATE**
  + Procesamiento en Paralelo (Motores de Análisis)
  + Sistema Modular Basado en Reglas (separador de oraciones, tokenizer, rotulador y un revisor que encuentra las entidades reconocidas en el análisis)
  + Arquitectura abierta y ampliable
  + Comprende un conjunto de recursos básicos de procesamiento (PRs):
    - Tokenizers (delimitadores de cadenas)
    - Gazetteers (revisores de encuentro de entidades)
    - Rotuladores POS
    - Separadores de Cadenas
    - Rotuladores Semánticos (JAPE traducer)
    - Orthomatcher (Co-referencias Ortográficas)
  + Permite el uso de UIMA como un 'plugin'
  + Buena Documentación + Ejemplos + Interfaz Gráfica para el manejo de Anotaciones Lingüísticas
  + Manejador del cuerpo del documento, editor de texto y un depurador de resultados
  + No mide los tiempos de procesamiento
  + Recursos de procesamiento y lingüística disponibles en distribución
  + ‘Plugins’ para:
    - OpenNLP
    - UIMA
    - OpenCalais
    - WordNet
    - StanfordParser
  + Familia de herramientas GATE
    - GATE Developer – un entorno de desarrollo integrado para componentes de procesamiento de lenguaje empaquetado con el sistema de Extracción de Información más ampliamente usado y un conjunto de ‘plugins’ comprensible
    - GATE Embedded – una librería de objetos optimizada para la inclusión de diversas aplicaciones
    - GATE Teamware – aplicación Web, un entorno para generar anotaciones colaborativo
    - GATE Cloud – Procesamiento distribuido en paralelo
  + GATE es software libre bajo la licencia GNU
* **Apache UIMA**  
  + Soporte nativo para elementos de computación distribuida.
  + Se enfoca en el ‘performance’ y la escalabilidad, con computación distribuida.
  + Incluye un conjunto de ‘plugins’ de desarrollo para Eclipse (esto en contraste con GATE que tiene su propio entorno de desarrollo).
  + Estructura:
    - Motores de Análisis (AE) - componentes básicos usados para analizar documentos
    - Estructura Común de Análisis (CAS) - usada para representar y compartir los resultados del análisis entre anotadores
  + La descripción de los archivos está contenida en archivos XML (Los flujos de información pueden ser configurados por medio de la definición de Descriptores XML)
  + Las estructuras de características son fuertemente tipadas (deben declarar el tipo de las anotaciones esperadas desde cada módulo)
  + Buena documentación
  + ‘Plugins’ limitados, algunos recursos externos deben ser implementados para algunas funcionalidades
  + Puede usar OpenNLP como un ‘plugin’
  + Mide de tiempo de procesamiento para cada módulo.
* **Apache OpenNLP**
  + Tiene la estructura de proyectos ‘Open Source’ relacionados al PLN
  + Las herramientas de OpenNLP usan modelos estadísticos basado en máxima entropía
  + Puede usarse en UIMA como un ‘plugin’
    - Tokenization
    - Segmentación de oraciones
    - Rotuladores POS (part of the speech)
    - Extracción de entidades nombradas
    - Chunking
    - Análisis sintáctico completo
    - Resolución co-referencial

*5. Tabla Comparativa de Herramientas Analizadas*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| A continuación, se comparan las herramientas descritas previamente con el fin de esclarecer las fortalezas y debilidades de cada una de ellas y de este modo concluir cuál de ellas es la opción más conveniente para el desarrollo del aplicativo. (Bank and Schierle, 2018) | | | | | | |
| **…** | **Menor Curva de Aprendizaje** | **Mayor Documentación** | **Estructura de desarrollo más Amplia** | **Mayores Recursos de Procesamiento** | **Mayor**  **Interoperabilidad** | **Mayor**  **Escalabilidad** |
| **Apache GATE** |  |  |  |  |  |  |
| **Apache UIMA** |  |  |  |  |  |  |
| **Apache OpenNLP** |  |  |  |  |  |  |

Dado que la herramienta que se planea desarrollar requiere de elementos más completos para el manejo de la información independiente del procesador que permita Procesamiento de Lenguaje Natural y con el objeto de explorar la amplia gama de opciones que ofrece este Framework, se opta por hacer uso de **Apache UIMA** que brinda todo un ambiente para el desarrollo de una aplicación más robusta y de mayor complejidad que las que se podrían crear con alguna de las otras herramientas ya mencionadas.

Adicionalmente resulta conveniente a la hora de desarrollar una aplicación **UIM** contar con una estructura base como la que proporciona el Framework para seguir un flujo en la implementación de los módulos necesarios para la extracción, el procesamiento y el posterior análisis de la información a tratar.

**Marco Conceptual**

De acuerdo a los criterios de realización establecidos, es necesario aclarar los siguientes conceptos para facilitar la lectura del presente documento y el entendimiento del proyecto a desarrollar:

* **Bodega de Datos (Data Warehouse):** Las bodegas de datos son macro estructuras que soportan el proceso de toma de decisiones a nivel gerencial y administrativo basándose en el manejo de grandes volúmenes de datos provenientes de diversas fuentes o diversos tipos que proveen un panorama más amplio de cada parte del negocio.
* **Almacén de Datos (Data Mart):** Son subconjuntos dela Bodega de Datosque se especializan y se enfocan en un sector particular del negocio con el fin de explotar de mejor manera la información propia de esta área para poder así dar lugar a la toma de decisiones más óptimas.
* **Minería de datos (Data Mining):** La minería de datos, es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. El análisis de esta información y el descubrimiento de estos patrones permiten la consecución de nueva información a partir de los datos recogidos.
* **Big Data:** Son sistemas a gran escala usados para el manejo de grandes volúmenes de información. Adicionalmente este proceso sirve para señalar los procedimientos que pretenden la búsqueda de patrones dentro de los datos que almacena.
* **Procesamiento de Lenguaje Natural:** El procesamiento del lenguaje natural (PLN) es el campo que combina las tecnologías de la ciencia computacional (como la inteligencia artificial, el aprendizaje automático o la inferencia estadística) con la lingüística aplicada, con el objetivo de hacer posible la comprensión y el procesamiento asistidos por ordenador de información expresada en lenguaje humano para determinadas tareas, como la traducción automática, los sistemas de diálogo interactivos, el análisis de opiniones, etc.
* **Recuperación de Información (Information Retrieval):** es la ciencia de la búsqueda de información en documentos electrónicos y cualquier tipo de colección documental digital, encargada de la búsqueda dentro de éstos mismos, búsqueda de [metadatos](https://es.wikipedia.org/wiki/Metadato) que describan documentos, o también la búsqueda en [bases de datos relacionales](https://es.wikipedia.org/wiki/Base_de_datos), ya sea a través de [internet](https://es.wikipedia.org/wiki/Internet), una [intranet](https://es.wikipedia.org/wiki/Intranet), y como objetivo realiza la recuperación en textos, imágenes, sonido o datos de otras características, de manera pertinente y relevante.
* **Aprendizaje Automático (Machine Learning):** El aprendizaje automático [es un tipo de inteligencia artificial](http://searchdatacenter.techtarget.com/es/cronica/Cuatro-tendencias-de-big-data-y-AI-a-las-que-echarles-un-ojo) (AI) que proporciona a las computadoras la capacidad de aprender, sin ser programadas explícitamente. El aprendizaje automático se centra en el desarrollo de programas informáticos que pueden cambiar cuando se exponen a nuevos datos.
* **Minería de Texto (Text Mining):**  Conjunto de técnicas que nos permiten extraer información relevante y desconocida de manera automática dentro de grandes volúmenes de información textual, normalmente en lenguaje natural y por lo general no estructurada.
* **Extracción, Transformación y Carga (ETL):** Proceso que permite mover datos desde múltiples fuentes, reformatearlos, limpiarlos y cargarlos en otra base, bodega o almacén de datos.
* **Razonamiento Automatizado (Automated Reasoning):** El razonamiento es un proceso mediante el cual se realizan inferencias; el razonamiento automatizado tiene que ver con la construcción de sistemas que automatizan este proceso. Aunque el objetivo general es mecanizar diferentes formas de razonamiento, el término se emplea para referirse a la automatización del razonamiento deductivo válido como se practica en matemáticas y lógica formal.
* **Análisis de Sentimientos (Sentiment Analisys):** Es el proceso de determinar el tono emocional que hay detrás de una serie de palabras, y se utiliza para intentar entender las actitudes, opiniones y emociones expresadas en una mención online.
* **Inteligencia de Negocios (Business Intelligence):** Es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.
* **Metadato:** Son aquellos elementos que contienen información que le es útil a usuarios, desarrolladores y administradores de la bodega de datos. Componen un cúmulo de información del que puede ser extraída dicha información tras su procesamiento.
* **KANBAN:** Técnica para la gestión de las tareas de un proyecto basada en el desarrollo incremental y la división del trabajo. Una de sus principales características es el uso de mecanismos visuales para observar el estado de las tareas. Si bien este marco de trabajo se asemeja a SCRUM es bastante más flexible y adaptable.

**ESTADO DEL ARTE**

**Aplicaciones para el Análisis de Información No Estructurada**

Si bien existen en el mercado diversas tecnologías que permiten por medio de la Minería de Texto, el Procesamiento de Lenguaje Natural, Modelos Estadísticos, etc., el análisis de información no estructurada, algunas se destacan por sus mejoras en rendimiento y opciones para la detección de estructuras.

Una de estas herramientas es el **Information Discovery** desarrollado por **Averbis *text analytics.*** Si bien esta plataforma es compatible con Apache UIMA, Averbis se ha encargado de enriquecer el Framework en una amplia gama de aspectos que van desde la infraestructura hasta los componentes que conforman su núcleo. Esta herramienta de análisis de texto y machine learning almacena y analiza todo tipo de documentos con el objetivo de buscar información relevante al usuario ya sea en documentos de patentes, literatura investigativa, sitios web, bases de datos, entre otras fuentes. El Information Discovery no solo le permite al usuario la estructuración de la información no estructurada en las fuentes procesadas, sino que también permite el despliegue de resultados en distintos formatos de acuerdo a los filtros que el usuario defina para descubrir conocimiento a partir de lo encontrado (Averbis GmbH, 2017).

Desde su creación el Framework UIMAse ha convertido en una herramienta frecuentemente usada para el desarrollo de aplicaciones para el procesamiento de información no estructurada en el ámbito comercial pero también en el ámbito académico. Investigadores del Centro Tecnológico de Cataluña desarrollaron un **Sistema basado en UIMA para el Análisis de Sentimientos** en las Opiniones de los clientes de un hotel. Este sistema se encargaba de detectar las opiniones puntuales de los clientes sobe el hotel y asignaba una polaridad que era lo que determinaba si el cliente en su comentario daba o no una calificación positiva sobre el establecimiento. Para la realización de este análisis de información se tomaron como anotaciones objetivos (sujeto al cual e hacía referencia en el comentario) e indicios (calificativos que el cliente asignaba al que sería el objetivo de su opinión), una vez hecho esto los desarrolladores optaron por usar distintos métodos que les permitieran hacer la asignación de la polaridad a la Unidad de Opinión, métodos tales como la asignación de una polaridad a los indicios detectados (polaridad que luego sería extendida a la Unidad de Opinión) y también métodos más complejos como el uso de Máquinas Vectoriales de Soporte que pudiesen ser entrenadas para la clasificación de dichas Unidades de Opinión basadas en un cúmulo de características compuestas por palabras, sus polaridades, negaciones y algunos cuantificadores. Los investigadores adicionalmente hicieron uso de **Apache Solr** lo cual les permitió la visualización de los resultados obtenidos. Los resultados permitieron concluir que de **700** opiniones entregadas al sistema un **88,5%** fueron correctamente identificadas y de este porcentaje de unidades de opinión leídas a aproximadamente al **70%** se les asignó correctamente una polaridad, dejando un margen de acierto aceptable (Rodríguez-Penagos et al., 2017).

Otra herramienta que fue elaborada haciendo uso de este Framework es **BLUIMA,** plataformaque en el 2013 fue desarrollada por una iniciativa de la Escuela Politécnica Federal de Lausana denominada **Blue Brain Project[[4]](#footnote-4).** Esta herramientafue desarrollada específicamente para la extracción de información Neurocientífica proveniente de artículos publicados en revistas especializadas. Para esto se desarrollaron modelos de análisis a partir de Procesamiento de Lenguaje Natural biomédico (BioNLP)[[5]](#footnote-5). A través de este desarrollo se propusieron dos componentes para el Análisis de Información No Estructurada: el primer componente permitiría la configuración e instanciación de flujos de desarrollo en UIMA por medio de un simple lenguaje de comandos permitiendo diseñar y ejecutar flujos de análisis a personas que no fuesen expertas en el manejo del Framework y el segundo componente fue una Estructura Común de Análisis de Almacenamiento basada en MongoDB la cual permitiría al desarrollador llevar a cabo anotaciones incrementales sobre documentos extensos extraídos de las fuentes seleccionadas. Los investigadores encontraron que estos componentes desarrollados agilizaron el proceso de análisis y manejo de información, así como también el despliegue de todo el flujo de información a través de los componentes de PLN (Bluebrain.epfl.ch, 2017).

**Análisis de Sentimientos**

Dada la importancia en el reconocimiento de la opinión del usuario al momento de analizar su perspectiva sobre el producto o servicio ofrecido, se hace de vital importancia explorar elementos referentes al **Análisis de Sentimientos** en relación al servicio al cliente. Si bien el Análisis de Sentimientos suele ser usado con el objetivo mejorar la experiencia de usuario también puede ser de utilidad en la construcción de estrategias que le permitan a la organización mejorar el servicio ofrecido.

En este sentido casi todas las organizaciones toman medidas de diferentes maneras reconociendo la importancia de esta labor. Ya que el Análisis de Sentimientos es funcional en una amplia gama de negocios y tiene múltiples aplicaciones, los casos de uso de esta estrategia no solo están presentes en el ámbito de negocios y en el entorno comercial.

Un ejemplo de cómo el Análisis de Sentimientos está presente y es de utilidad en muchos aspectos del mundo que nos rodea son las pasadas elecciones presidenciales de los Estados Unidos del año 2016donde el Análisis de Sentimientos fue utilizado para conocer la percepción que tenían los potenciales votantes en torno a los candidatos presentes en dichas elecciones para de esta manera obtener una primera impresión de las intenciones de voto de la población muestreada. En este caso en específico los investigadores procuraron clasificar la información de los usuarios de la plataforma de twitter dentro de cinco sentimientos que pudieran englobar las opiniones recopiladas en el estudio, dichas emociones de clasificación fueron: **felicidad**, **tristeza**, **miedo**, **risas** y **enojo**. Como objeto de estudio se tomaron cerca de **trescientos mil comentarios** que estuvieran compuestos por ciertos términos clave como “**política**”, “**candidatos políticos**” o el nombre completo de alguno de los candidatos a las pasadas elecciones. Así mismo se hizo uso de un elemento propio de las comunicaciones digitales de la actualidad, así como de las redes sociales tales como lo son los **Emojis** que estuvieron presentes en cerca de **tres mil** de los comentarios descargados para la ejecución del análisis. En particular el análisis de estos datos y el modelamiento se hizo a través de tres algoritmos con los cuales se puso a prueba cuál era el de mayor efectividad considerando el tipo de problema a tratar, estos algoritmos fueron: **Support Vector Machines** (**Máquinas de Vectores de Soporte**), **Nearest Neighbors** (**K-vecinos más cercanos**) y el **Algoritmo de Clasificación de Naive Bayes**. A partir de este estudio se pudo conocer cuál era la opinión de la población estadounidense en torno a cada uno de los candidatos presidenciales en los estados donde la relación tweets - habitantes era significativa, dichos resultados de este análisis fueron desplegados gráficamente al mapear por estado cuál era la emoción que mejor se adaptada de acuerdo a los comentarios recogidos (Chin, D., Zappone, A., & Zhao, J. 2017).

Atendiendo más puntualmente a los aspectos referentes al servicio al cliente, hay plataformas en el mercado especializadas en el análisis de sentimientos y que buscan captar las opiniones de los clientes en redes sociales y por otros medios de interacción de los usuarios con las compañías. Un ejemplo de este tipo de proyectos que buscan soportar la relación cliente-compañía haciendo uso de análisis de sentimientos es **Brand24**.

**Brand24 -** Este aplicativo cuenta con una amplia estructura de análisis de la información que se puede de las menciones de una marca en redes sociales. A través de la plataforma se puede conocer la cantidad de menciones que tuvo la marca de manera rápida para facilitar la interacción con los usuarios, se puede conocer el volumen de discusión de los clientes en redes en relación al producto o la marca analizada, cuenta con herramientas para el análisis comercial de la información que permitan la toma de decisiones, provee alertas a la compañía sobre cambios significativos en el patrón de comportamiento de los clientes en términos de la cantidad o la calidad de los comentarios dejados en redes sociales y claro permite la clasificación de los comentarios en tres categorías (positivo, neutral o negativo) basados en las emociones expresadas por los usuarios. Una de las grandes ventajas que ofrece este proyecto es la facilidad y rapidez con la que los directivos de la compañía pueden acceder a la información, ya que cuenta con una versión para dispositivos móviles que les permite a sus usuarios ver en tiempo real los movimientos de sus clientes a través de la red. Entre las marcas más reconocidas que hacen uso de esta plataforma de monitoreo de redes están: **Intel**, **Uber**, **AirFrance** y **Discovery Network**. (www.brand24.com, 2017)

**Inteligencia Sobre el Cliente (Customer Intelligence)**

En cuanto a la Inteligencia de negocios aplicada al servicio al cliente surge el **Customer Intelligence (CI)** como una evolución de los sistemas de **Customer Relationship Management (CRM)** tradicionales. El **CI** busca mejorar la experiencia del usuario ampliando la perspectiva del negocio con el fin de entender de fondo las necesidades más profundas del cliente y entender esos comportamientos que están generando un impacto en mi empresa.

Un claro ejemplo de uso que demuestra los grandes cambios que se pueden generar a partir de la apropiada aplicación de este concepto estuvo a manos de la compañía **Neuromobile Mall** que en un informe de 2015, al ser consultada para analizar el comportamiento de compra de cierto público en una determinada franja horaria en un centro comercial, descubrió que entre las 9:00 am y las 10:00 am disminuía en un 60% el tiempo de permanencia en el centro comercial de mujeres con familia entre los 25 y los 45 años. Para dar solución a este problema, se decidió enviar un mensaje al teléfono celular de las visitantes con un cupón para un café gratis al interior del centro comercial. Tras el paso de dos semanas los resultados fueron visibles, En una semana más del 70% de los cupones enviados fueron canjeados y la duración de las clientes con el perfil determinado aumentaron en aproximadamente el 70%, logrando con éxito menguar el efecto negativo (www.neuromobilemarketing.com, 2017).

**Integración de Información Estructurada y No Estructurada**

En la constante búsqueda de nuevas formas que permitan el uso de los grandes volúmenes de información no estructurada que son desaprovechados debido a su difícil manejo, pero que aun así tienen un crecimiento exponencial en la actualidad, se han intentado implementaciones de diversas herramientas y métodos que logren hacer una integración efectiva en una bodega de datos que permita el análisis posterior de dicha información.

Uno de los casos expuestos es el estudio publicado en el 2015 por la Revista de la **Asociación Americana de Informática Médica** en torno a la integración de información estructurada y no estructurada con el fin de asignar los códigos de diagnósticos y procedimientos médicos a los pacientes internos en los hospitales. Esta investigación también tuvo como objetivo demostrar que el manejo conjunto de datos estructurados y no estructurados mejora considerablemente la calidad de los resultados que se obtienen con cada tipo de información de manera independiente.

La información, en su mayoría, fue tomada de los registros médicos electrónicos de la bodega de datos del **Hospital Universitario de Antwerp (Bélgica)** donde se condensaba a manera de texto la información de los pacientes y a partir de la cual se podían realizar las predicciones en cuanto a diagnósticos y procedimientos. Adicional a esta fuente, también fueron tomados: reportes de cirugía, cartas entre paciente y especialista o entre especialistas de la institución médica, notas diarias de la evolución del estado del paciente, protocolos diagnósticos como la interpretación de exámenes realizados a los pacientes por parte del personal médico, etc. Se evaluaron dos enfoques de integración por separado: el primer tipo de integración buscaba tomar y combinar los datos y características de distintas fuentes para así después insertarlas en un mismo modelo y el segundo método de integración desarrolla un modelo a partir de cada fuente para después mezclar los modelos en una misma estructura. Cada una de estas formas de integración era evaluada en fuentes de datos y códigos clínicos de un amplio conjunto de especialidades médicas.

Cuando se compararon los resultados de este estudio con métodos de predicción que separan información estructurada de no estructurada, los modelos de análisis híbridos superan ampliamente el poder predictivo de las herramientas conocidas, por ejemplo, el **Valor-F** (medida de precisión estadística evaluada en el test) aumentó del **30,6%** al **38,3%** para la Clasificación Internacional de Enfermedades, demostrando así que los modelos que utilizan múltiples fuentes de datos electrónicos de registro de salud sistemáticamente, superan a los modelos que hacen uso de fuentes de datos de forma aislada en la tareas predictivas en una amplia gama de especialidades médicas (Scheurwegs, E., Luyten, L., Daelemans, W., & Van den Bulcke, T., 2015).

**Bodegas de Datos Orientadas al Servicio al cliente**

Desde que surge la necesidad de plantear un sistema de inteligencia de negocios enfocado en mejorar la manera en que se usa la información arrojada por los clientes y que pueda optimizar los procesos ejecutados de las empresas, se desarrollan múltiples iniciativas que den solución a esta problemática dando lugar al análisis de las particularidades que representan el tipo de negocio que ejecuta la organización.

Uno de estos estudios fue desarrollado en la Universidad de Friburgo, Suiza en torno a la importancia del manejo de una Bodega de Datos especializada en los clientes. El objetivo de esta investigación se centró en cómo fundamentar la creación de dicha bodega, reconociendo la importancia que tiene lograr dilucidar la mayor cantidad de información posible de los usuarios en relación a cinco interrogantes principales:

* ¿Qué caracteriza a un cliente o a un grupo de clientes de interés?
* ¿Cuál es el valor de un cliente en el pasado, el presente y el futuro?
* ¿Qué tan leal es un cliente o un grupo de clientes en específico?
* ¿Qué cambios en las peticiones de los clientes o en la calidad del servicio pueden ser rastreados?
* ¿Cuáles son los canales de comunicación preferidos por el cliente?

Así mismo a través de este estudio también se buscaba exponer la necesidad de automatizar el manejo de la información de los clientes enfatizando en que de esta manera su análisis y posterior estudio se facilitaría si este proceso no era ejecutado manualmente.

Otro estudio en esta área fue realizado en la Universidad de Estudios Económicos de Bucarest, Rumania, en donde la investigadora Mónica Lía planteó un modelo de análisis de datos de los clientes de compañías de telecomunicaciones haciendo uso de herramientas de inteligencia de negocios. Este estudio tenía como objetivo dar al mercado rumano una herramienta para mejorar las previsiones en torno al comportamiento de los usuarios y que esto pudiese desencadenar una mejor toma de decisiones por parte del medio de las telecomunicaciones en general.

Monica Lia resalta la importancia de contar con un correcto despliegue de los resultados del análisis de la información de los clientes a través de tablas dinámicas, gráficos interactivos y diagramas que revelen, a quienes deben tomar las decisiones y para quienes usualmente es más sencillo interpretar de esta forma la información, los aspectos de la compleja relación cliente-compañía.

Tras diversos análisis, principalmente enfocados en aspectos como: análisis de los clientes de acuerdo al tráfico en sus telecomunicaciones y por tipo de usuario, posicionamiento geográfico de los clientes, caracterización de usuarios por tipo de productos consumidos, etc., mostró la importancia de la previsión en ventas que resultaba de estudios de este tipo. Así mismo se demostró que las campañas publicitarias, ofertas y demás métodos de interacción con el usuario se ven alterados de acuerdo al resultado de los reportes dinámicos que dan luces del comportamiento y de los intereses del usuario. En síntesis, el propósito de este estudio fue mostrar la gran ayuda que aportan al mercado este tipo de análisis y que son estudios que pueden relativamente ser de fácil aplicación si se hace uso de herramientas de Inteligencia de Negocios (Monica Lia, 2015).

*5. Tabla Comparativa de Desarrollos Previos – Estado del Arte*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A continuación, se exponen algunos de los aspectos más relevantes de los estudios y desarrollos previos, producto de la etapa investigativa y se destacan los que aportan mayor cantidad de resultados de interés en relación al trabajo propuesto a manera de conclusión. | | | | |
| **…** | **Especialidad abarcada** | **\*Pertinencia respecto a estudio actual** | **\*\*Consistencia del estudio** | **Elementos de utilidad** |
| **Information Discovery (Averbis Text Analytics)** | Aplicaciones para el Análisis de Información No Estructurada | *MEDIA* | El software desarrollado por Averbis presenta un sistema completo que parece dar solución al problema del manejo de la información no estructurada en una compañía. Sin embargo, al ser un sistema de pago no hay forma de verificar la efectividad del sistema. | 1. El sistema cuenta con un módulo de carga de archivos a procesar y permite la configuración de un flujo de análisis y reconocimiento de entidades. Todo esto soportado por el Framework Apache UIMA con que ha sido desarrollado. 2. Cuenta con una interfaz intuitiva que aleja de usuario de la complejidad de los componentes internos que permiten el procesamiento documental. |
| **Sistema basado en Apache UIMA para el Análisis de Sentimientos (Centro Tecnológico de Cataluña, Rodríguez-Penagos)** | Aplicaciones para el Análisis de Información No Estructurada  Análisis de Sentimientos | *ALTA* | El sistema desarrollado cuenta con un sustento teórico consistente en torno a la estructura definida para la construcción del flujo de información y su posterior análisis  Adicionalmente lo largo de todo el estudio realizado por los investigadores del Centro Tecnológico de Cataluña se publican los resultados del procesamiento como evidencia de lo planteado. | 1. Provee una estructura para la separación de Unidades de Opinión que facilita la extracción de elementos al interior de Información No Estructurada. 2. Expone como pueden ser acoplados a Apache UIMA otros sistemas como OpenNLP, Lemmatizer, etc., que permiten el robustecimiento del sistema con el objetivo de optimizar las funciones de procesamiento y extracción de la información. |
| **Análisis de Sentimientos en Twitter para Elecciones Presidenciales en Estados Unidos del 2016 (D. Chin, A. Zappone, J. Zhao)** | Análisis de Sentimientos | *MEDIA* | A lo largo del documento donde se presenta el estudio se justifican los parámetros de clasificación que sustentan el análisis y al mismo establecen una comparativa de métodos y algoritmos de clasificación que permitan concluir cuál es la mejor opción, basados en resultados. | 1. Establecen conjuntos de entrenamiento, de prueba, un proceso de evaluación de los datos sustraídos del texto plano y algunos parámetros en la selección de la información a usar. 2. Presentan una clara comparativa en métodos de clustering y clasificación con los cuales se ejecuta la Minería de Texto sobre los comentarios de Twitter. 3. Dentro de las conclusiones se incluye un análisis detallado del error generado tras el procesamiento de la información. |
| **Brand24 (Aplicativo)** | Análisis de Sentimientos | *BAJA* | Debido a que este también es un sistema de pago, no se cuenta con suficiente información sobre el desarrollo del aplicativo y como este realiza el análisis de Sentimientos. | 1. Provee una vistosa interfaz de usuario que facilita el acceso a la información encontrada y que hace evidente la clasificación dada por el sistema. Así mismo establece un sistema de categorías que permite encapsular los resultados en comentarios **Positivos**, **Neutros** y **Negativos.** 2. Si bien en el sistema que se planea desarrollar en el actual trabajo no está contemplado el análisis de información en tiempo real, **Brand24** permite extraer de las fuentes (redes sociales, foros, blogs, etc.) la información generada hasta el momento real de la consulta. |
| **Neuromobile Mall (Aplicativo)** | Inteligencia sobre el Cliente (Customer Intelligence) | *BAJA* | Al igual que los sistemas comerciales analizados y revisados previamente el acceso a la información de fondo sobre el funcionamiento del aplicativo es limitado. Sin embargo Neuromobile provee ciertos elementos resultantes de la experiencia sobre la implementación del sistema que evidencian los resultados obtenidos y las ventajas que provee. | 1. Adaptabilidad del sistema para la rápida interacción del usuario. Reconocer que es de interés para el usuario y darle acceso rápido a este contenido aún antes de su solicitud. 2. Amplia gama de reportes generados con distintos filtros que dan todo un abanico de opciones de consulta y de visualización de resultados de la información procesada. 3. Impactar distintos puntos del proceso, brindando resultados de interés a distintos niveles. ¿Cómo hacer útil la información procesada no solo a nivel gerencial sino también en las áreas de la empresa en particular? |
| **Data Integration os structured and unstructured sources for assigning clinical codes for patient stays (Scheurwegs, E., Luyten, L., Daelemans, W., & Van den Bulcke, T., 2015)** | Integración de Información Estructurada y No Estructurada | *MEDIA* | El documento revisado cuenta con un sustento real basado en el proceso de experimentación y como conclusión son presentados todo un conjunto de resultados visibles al público. |  |
|  |  |  |  |  |
| **\***La pertinencia del estudio se estima en torno al grado de relación que exista entre el elemento analizado y el estudio en desarrollo. La clasificación se dará en 3 niveles de acuerdo a esta relación (**ALTA**: Múltiples elementos relacionados, **MEDIA**: Algunos elementos de utilidad para el estudio, **BAJA**: Pocos elementos que relacionen un estudio con otro o elementos inconexos entre sí)  **\*\***Laconsistencia teórica de cada elemento analizadopondrá en consideración el sustento real de la teoría desarrollada y su efectividad práctica. | | | | |

**ALCANCES Y LIMITACIONES**

Este desarrollo tiene como propósito el desarrollo de una herramienta que permita conocer con mayor precisión las impresiones que los clientes de una corporación tienen en cuanto al servicio que esta ofrece. Los detalles de la opinión que los usuarios hayan plasmado a través de redes sociales, por registros telefónicos que hayan resultado del cruce entre el cliente y el servicio de atención al cliente o por otras fuentes usadas para la recopilación de datos serán analizados y posteriormente desplegados gráficamente para una fácil lectura e interpretación por parte del usuario final de la herramienta.

Los datos analizar dependerán única y exclusivamente de las fuentes proporcionadas por la empresa ya que se pretende hacer un estudio con información real adquirida de los clientes que han plasmado a través de algún medio su opinión en torno a los servicios ofrecidos por la organización. Ya que el movimiento en redes sociales de la empresa es relativamente nuevo, no se cuenta con un gran monto de datos que permitan, posteriormente, garantizar una alta precisión en los patrones encontrados por la herramienta sobre el global de los clientes con los que cuentan.

Es importante resaltar que el análisis ejecutado por la herramienta no sucede en tiempo real. Hay una etapa de recopilación de información que, dependiendo de la cantidad de datos que se desean recoger, tomará una cantidad específica de tiempo, con lo cual sí se podrá realizar un procesamiento y análisis sobre estos datos recogidos. Es posible que entre la recolección de datos y el despliegue de resultados haya sucedido alguna eventualidad que conlleve a un cambio en la percepción del usuario respecto a la marca.

**METODOLOGÍA**

Para este proyecto será necesario el diseño de una bodega de datos orientada al área de servicio al cliente de la empresa de la cual será tomada la información. Esta bodega almacenará todos los datos condensados tras el procesamiento realizado y deberá considerar el funcionamiento interno de dicha organización y los canales de interacción con sus clientes para poder ejecutar el análisis de resultados de manera óptima. Para conocer a fondo el funcionamiento de la empresa se hará un proceso de levantamiento de requerimientos y de indagación con el encargado del área de TI, lo cual dará luces para la construcción de la bodega de datos en el cual se basará la herramienta para el análisis de la información adquirida en las distintas fuentes de Servicio al Cliente con las que cuenta la organización.

Ya que el Framework seleccionado para el análisis de la información no estructurada cuenta con una implementación de sus funcionalidades en JAVA y que se pueden integrar como plugins al IDE ECLIPSE, este será el lenguaje y la plataforma en las cuales se desarrolle la herramienta en su totalidad, incluyendo interfaz gráfica y el módulo de despliegue de reportes que se implementará con ayuda de la API Open Source JasperReports.

Para hacer uso de la plataforma será necesario obtener la opinión, quejas y comentarios de los clientes de la empresa de sus redes sociales, correo electrónico y del servicio de call center, estos comentarios y opiniones de los usuarios van a permitir captar las sensaciones en torno al servicio ofrecido. Para poder llegar a resultados más concluyentes, las opiniones y comentarios de los usuarios después de ser procesados serán clasificados en una de tres categorías posibles al momento de ser almacenados en la bodega de datos. Esta clasificación sintetizará correctamente la intencionalidad de dicha publicación o comentario. Las tres categorías serán: **Positivo, Neutral, Negativo.** Para garantizar una exitosa clasificación de los contenidos extraídos se tomarán un grupo de términos por categoría que permitan a la herramienta conocer las emociones que han impulsado al cliente a dar su punto de vista en cuanto al servicio recibido de parte de la empresa.

Las tareas a desarrollar estarán guiadas por la metodología ágil KANBAN que servirá para procesar de manera organizada y por etapas la ejecución del proyecto. Esta metodología va a permitir controlar de manera visual los alcances, el progreso, los pendientes y las limitaciones en la implementación de la herramienta.

**Actividades por Objetivo**

Las actividades a desarrollar por objetivo son las siguientes:

**Objetivo 1:** Revisar algunas herramientas y propuestas orientadas a la integración y análisis de información estructurada y no estructurada en una bodega de datos.

**Actividad 1.1:** Consultar artículos, libros y tesis previas en el estudio del manejo de información No Estructurada.

**Actividad 1.2:** Probar herramientas y tecnologías desarrolladas en este contexto con el fin de establecer un punto de partida para el proyecto.

**Actividad 1.3:** De entre las tecnologías exploradas seleccionar la que mejor se adapte a las condiciones requeridas para dar solución al problema puntual que se está tratando.

**Objetivo 2:** Diseñar un esquema de bodega de datos orientada al servicio al cliente para la integración de la información estructurada y no estructurada.

**Actividad 2.1:** Hacer el respectivo levantamiento de requerimientos con la entidad que facilita la información para el desarrollo de esta herramienta para saber cuáles son sus necesidades específicas y para conocer a fondo su estructura.

**Actividad 2.2:** Consultar las fuentes de datos disponibles en la entidad en relación al Servicio al Cliente y establecer las fuentes a escoger para la etapa de procesamiento de datos.

**Actividad 2.3:** Elaborar el esquema de Bodega de Datos para cruzar datos estructurados y datos no estructurados con los que se cuente, de acuerdo a los requerimientos de la empresa que otorga los datos.

**Objetivo 3:** Implementar un módulo que permita extraer, procesar y cargar información estructurada y no estructurada en la bodega de datos.

**Actividad 3.1:** Con base en los requerimientos levantados y la información obtenida a partir de las fuentes de información ya mencionadas, hacer un pre-procesamiento que permita **extraer** y condensar en documentos de texto bajo el mismo estándar todos los datos no estructurados a manejar con el aplicativo.

**Actividad 3.2:** Construir un módulo que permita analizar y **transformar** desde los documentos leídos la información relevante y que será objeto de estudio al finalizar el proceso.

Este análisis contemplará aspectos de utilidad tales como: Tipo se servicio (Salud, Educación, Recreación, etc.), especialista, la ubicación del especialista y la intensión del mensaje leído.

**Actividad 3.3:** Cargar la información estructurada y no estructurada en la bodega de datos para su posterior análisis.

**Actividad 3.4:** Construir la interfaz gráficaque soporteel funcionamiento del aplicativo.

**Objetivo 4:** Implementar un módulo que facilite la visualización de la información procesada a partir de alguna herramienta de generación de reportes ya existente.

**Actividad 4.1:** Buscar, probar y seleccionar librerías de visualización de reportes que permitan mostrar apropiadamente los resultados.

**Actividad 4.2:** Integrar dicha librería con la herramienta implementada y establecer una interfaz de consultas que permita al usuario observar en detalle la información recolectada de acuerdo a los parámetros que defina en sus filtros.

**Objetivo 5:** Diseñar un plan de pruebas que permita validar la herramienta implementada y verificar el cumplimiento de los requerimientos.

**Actividad 5.1:** Realizar la búsqueda de una Herramienta para la realización de pruebas que se adapten a la metodología ágil seleccionada para la ejecución del proyecto. Una opción para la realización de pruebas es el uso de pruebas bajo el enfoque del **Behaviour Driven Development**.

**Actividad 5.2**: Determinar un conjunto de pruebas de aceptación y de pruebas unitarias para ser usadas con la herramienta seleccionada en el punto **5.1**.

**Actividad 5**.**3:** Ejecutar con la herramienta los casos de prueba y analizar los resultados obtenidos en aras de verificar el cumplimiento de los requerimientos.

**CRONOGRAMA**

De acuerdo a las tareas establecidas se establecen unos tiempos que contemplan la etapa de diseño de la herramienta, así como la implementación y su posterior testeo. En estas tareas propuestas se consideran los plazos de duración de cada tarea, así como las relaciones entre ellas. Este cronograma se construye considerando un plazo de diez meses en los que deberán concluir las tareas presupuestadas.

*Tabla 1. Cronograma de Actividades Etapa de Investigación*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBJ 1.** Revisar algunas herramientas y propuestas orientadas a la integración y análisis de información estructurada y no estructurada en una bodega de datos. | | | | | | | | | | |
|  | **Mes 1** | **Mes 2** | **Mes 3** | **Mes 4** | **Mes 5** | **Mes 6** | **Mes 7** | **Mes 8** | **Mes 9** | **Mes 10** |
| **ACT 1.1** |  |  |  |  |  |  |  |  |  |  |
| **ACT 1.2** |  |  |  |  |  |  |  |  |  |  |
| **ACT 1.3** |  |  |  |  |  |  |  |  |  |  |

*Tabla 2. Cronograma de Actividades Etapa de Diseño*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBJ 2:** Diseñar un esquema de integración de información estructurada y no  estructurada en una bodega de datos con información de servicio al cliente. | | | | | | | | | | |
|  | **Mes 1** | **Mes 2** | **Mes 3** | **Mes 4** | **Mes 5** | **Mes 6** | **Mes 7** | **Mes 8** | **Mes 9** | **Mes 10** |
| **ACT 2.1** |  |  |  |  |  |  |  |  |  |  |
| **ACT 2.2** |  |  |  |  |  |  |  |  |  |  |
| **ACT 2.3** |  |  |  |  |  |  |  |  |  |  |

*Tabla 3. Cronograma de Actividades Etapa de Desarrollo – Módulo ETL*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBJ 3:** Implementar un módulo que permita extraer, procesar y cargar información estructurada y no estructurada en la bodega de datos. | | | | | | | | | | |
|  | **Mes 1** | **Mes 2** | **Mes 3** | **Mes 4** | **Mes 5** | **Mes 6** | **Mes 7** | **Mes 8** | **Mes 9** | **Mes 10** |
| **ACT 3.1** |  |  |  |  |  |  |  |  |  |  |
| **ACT 3.2** |  |  |  |  |  |  |  |  |  |  |
| **ACT 3.3** |  |  |  |  |  |  |  |  |  |  |
| **ACT 3.4** |  |  |  |  |  |  |  |  |  |  |

*Tabla 4. Cronograma de Actividades Etapa de Desarrollo*

*Módulo de Visualización y Reportes*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBJ 4:** Implementar un módulo que facilite la visualización de la información procesada. | | | | | | | | | | |
|  | **Mes 1** | **Mes 2** | **Mes 3** | **Mes 4** | **Mes 5** | **Mes 6** | **Mes 7** | **Mes 8** | **Mes 9** | **Mes 10** |
| **ACT 4.1** |  |  |  |  |  |  |  |  |  |  |
| **ACT 4.2** |  |  |  |  |  |  |  |  |  |  |

*Tabla 5. Cronograma de Actividades Etapa de Diseño*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OBJ 5:** Diseñar un esquema de pruebas para validar la herramienta implementada. | | | | | | | | | | |
|  | **Mes 1** | **Mes 2** | **Mes 3** | **Mes 4** | **Mes 5** | **Mes 6** | **Mes 7** | **Mes 8** | **Mes 9** | **Mes 10** |
| **ACT 5.1** |  |  |  |  |  |  |  |  |  |  |
| **ACT 5.2** |  |  |  |  |  |  |  |  |  |  |
| **ACT 5.3** |  |  |  |  |  |  |  |  |  |  |

**PRESUPUESTO**

**Presupuesto General**

*Tabla 6. Presupuesto General*

|  |  |
| --- | --- |
| **RUBROS** | **VALOR (Pesos)** |
| Personal | $4’425.000,00 |
| Hardware | $1’400.000,00 |
| Software | $0,00 |
| Otros | $±20.000,00 |
| **TOTAL PRESUPUESTO** | $±5’845.000,00 |

**Presupuesto Detallado**

A continuación, se describe en detalle el presupuesto mostrado en la sección anterior:

*Tabla 7. Costo Personal*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Nombre de Participante** | **Función** | **Meses de Vinculación** | **Dedicación (Horas x Semana)** | **Costo por Hora** | **Costo Total (Pesos)** | **Fuente de Financiación** |
| Oswaldo Solarte Pabón M. Sc. | Director | 10 | 5 | $10.125,00 | $2’025.000,00 | Propia |
| Esteban Antonio Llanos Millán | Investigador, Analista y Desarrollador | 10 | 20 | $3.000,00 | $2’400.000,00 | Propia |
| TOTAL COSTO PERSONAL | | | | | $4’425.000,00 | |

*Tabla 8. Costo Hardware*

|  |  |  |  |
| --- | --- | --- | --- |
| **Equipo** | **Justificación** | **Costo (Pesos)** | **Fuente de Financiación** |
| Computador Portátil | Necesario para el proceso de investigación y desarrollo de la herramienta. | $1’400.000,00 | Propia |
| TOTAL COSTO HARDWARE | | $1’400.000,00 | |

*Tabla 9. Costo Personal*

|  |  |  |  |
| --- | --- | --- | --- |
| **Equipo** | **Justificación** | **Costo (Pesos)** | **Fuente de Financiación** |
| Impresiones y Fotocopias | En el proceso de investigación se hace necesario contar con cierto tipo de material en físico (libros, artículos, etc.) | $±20.000,00 | Propia |
| TOTAL COSTO IMPRESIONES | | $±20.000,00 | |

**BIBLIOGRAFÍA**

Bluebrain.epfl.ch. (2017). Bluebrain | EPFL. [online] Disponible en: http://bluebrain.epfl.ch/ [Consultado el 19 oct. 2017].

Rodríguez-Penagos, C., García Narbona, D., Massó Sanabre, G., Grivolla, J. and Codina Filbá, J. (2017). Sentiment Analysis and Visualization using UIMA and Solr. 1st ed. [ebook] Barcelona: Barcelona Media Innovation Centre. Disponible en: http://ceur-ws.org/Vol-1038/paper\_5.pdf [Consultado el 19 Oct. 2017].

Developing Language Processing Components with GATE. (2017). 8va ed. [ebook] Sheffield, Reino Unido. Disponible en: https://gate.ac.uk/sale/tao/tao.pdf [Consultado el 19 oct. 2017]

Apache OpenNLP Developer Documentation. (2017). 2da ed. [ebook] Apache Software Foundation. Disponible en: https://opennlp.apache.org/docs/1.8.2/manual/opennlp.html [Consultado el 19 oct. 2017].

Foster Kevin, Nathan Senthil, Rajan Deepak, Ballard Chuck, IBM InfoSphere Streams: Assembling Continuous Insight in the Information Revolution, IBM RedBooks, 2011.

Dossier Neuromobile Mall. (2017). [ebook] Disponible en: http://neuromobilemarketing.com/wp-content/uploads/2015/10/Dossier-Neuromobile-Mall-ESv1.pdf [Consultado el 19 oct. 2017].

Moncecchi, G., & Prada, J. (2015). Integración de Herramientas para Procesamiento de Lenguaje Natural (1ra ed., pp. 20-23). Montevideo, Uruguay: universidad de la República. Disponible en: https://www.fing.edu.uy/inco/grupos/pln/prygrado/Informe\_libPLN.pdf

Chin, D., Zappone, A., & Zhao, J. (2017). *Analyzing Twitter Sentiment of the 2016 Presidential Candidates* (1ra ed.). Palo Alto, California. Tomado de https://web.stanford.edu/~jesszhao/files/twitterSentiment.pdf

*Case Study UBER*. (2017). *Brand24*. Tomado el 18 abril de 2017, de https://brand24.com/case-study/uber/

Rouse, M. (2010). *Customer Intelligence*. *Search Business Analytics*. Tomado el 20 abril de 2017, De http://searchbusinessanalytics.techtarget.com/definition/customer-intelligence-CI

Soto, M. (2017). *Qué es el customer intelligence en Retail*. *Wivoanalytics*. Tomado el 20 abril de 2017, de http://blog.wivoanalytics.com/qu%C3%A9-es-el-customer-intelligence-en-retail

Douglas, T. (2017). *The Enterprises Guide to Customer Intelligence* (1ra ed.). Tomado de https://www.visioncritical.com/wp-content/uploads/2015/06/enterprise-guide-to-customer-intelligence.pdf

Scheurwegs, E., Luyten, L., Daelemans, W., & Van den Bulcke, T. (2015). Data integration of structured and unstructured sources for assigning clinical codes to patient stays. *JAMIA*, *23*(1). http://dx.doi.org/https://doi.org/10.1093/jamia/ocv115

Lia, M. (2015). *Customer Data Analysis Model using Business Intelligence Tools in Telecommunication Companies*. 1ra ed. [Artículo] Bucarest, Rumania. Disponible en: http://www.dbjournal.ro/archive/20/20\_5.pdf [Leído el 3 de mayo de 2017].

Meier, A. and Werro, N. (2005). *Customer Data Warehouse*. 1ra ed. [Artículo] Friburgo, Suiza. Disponible en: http://diuf.unifr.ch/is/studentprojects/pdf/reports/CRM\_SS05\_Customer\_Data\_Warehouse\_(AlainWahl).pdf [Leído el 1 de mayo de 2017].

Averbis GmbH. (2017). Averbis Information Discovery - Products for Your Success. [online] Disponible en: https://averbis.com/en/information-discovery/ [Consultado el 19 oct. 2017].

Bank, M. and Schierle, M. (2018). A Survey of Text Mining Architectures and the UIMA Standard. [Ebook] Estambul, Turquía. Disponible en: http://lrec.elra.info/proceedings/lrec2012/pdf/183\_Paper.pdf [Consultado el 9 enero de 2018].

1. 1 SINTEF. (2013, May 22). Big Data, for better or worse: 90% of world's data generated over last two years. *ScienceDaily*. Publicado el 18 de octubre, 2017 tomado de www.sciencedaily.com/releases/2013/05/130522085217.htm [↑](#footnote-ref-1)
2. ## Oasis-open.org. (2017). OASIS. [online] Disponible en: https://www.oasis-open.org/org [Consultado el 19 oct. 2017].

   [↑](#footnote-ref-2)
3. 3 Unstructured Information Management Architecture (UIMA). (2017). 1ra ed. [ebook] Adam Lally (IBM), Karin Verspoor (University of Colorado Denver), Eric Nyberg (Carnegie Mellon University), p.99. Disponible en: http://docs.oasis-open.org/uima/v1.0/os/uima-spec-os.pdf [Consultado el 19 Oct. 2017]. [↑](#footnote-ref-3)
4. 4 **Bluebrain.epfl.ch.** (2017). *Bluebrain | EPFL*. [online] Disponible en: http://bluebrain.epfl.ch/ [Consultado el 19 oct. 2017]. [↑](#footnote-ref-4)
5. 5 **2016.bionlp-st.org**. (2017). *BioNLP-ST 2016*. [online] Disponible en: http://2016.bionlp-st.org/home [Consultado el 19 oct. 2017]. [↑](#footnote-ref-5)