ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Maestría en Sistemas de Información Gerencial

“IMPLEMENTACIÓN DE UN MODELO DE MINERÍA DE OPINIÓN DE PRODUCTOS Y CONTENIDOS WEB PARA EL SOPORTE DE DECISIONES DEL DEPARTAMENTO DE MARKETING**”**

EXAMEN DE GRADO (COMPLEXIVO)

Previa a la obtención del título de:

MAGISTER EN SISTEMAS DE INFORMACIÓN GERENCIAL

JUAN CARLOS GARCIA PLUA

GUAYAQUIL - ECUADOR

2015

AGRADECIMIENTO

A Dios por sus bendiciones y ayuda que continuamente me provee, a mi esposa e hijos que les sacrifiqué tiempo; y a mis padres que siempre estuvieron apoyándome.

DEDICATORIA

A Dios, a Gisella, a Virginia, a Juan Sebastián, a Ethan Nicolás, y a Emilio Rafael.

TRIBUNAL DE SUSTENTACIÓN

.....

Ing. Lenin Freire Cobo

DIRECTOR DEL MSIG

.....

Ing. Carlos Alberto Salazar

PROFESOR DELEGADO

POR LA UNIDAD ACADÉMICA

.....

Ing. Jorge Rodríguez Echeverría

PROFESOR DELEGADO

POR LA UNIDAD ACADÉMICA

RESUMEN

Por primera vez en la historia de la humanidad, tenemos un gran volumen de contenido digital generado por los usuarios a través de las redes sociales, foros de discusión, blogs, y en opiniones de productos. La sociedad y los negocios tienen la imperiosa necesidad de extraer conocimiento de estos datos y para el área de mercadeo es importante cuantificar en tiempo real el sentimiento que tienen los clientes acerca de sus marcas o productos.

El análisis manual de estos datos no es factible ya que por su naturaleza son de gran volumen y no están estructurados. Esto implica que los resultados necesitan mayor cantidad recursos con un alto porcentaje de errores y no están disponibles a tiempo.

La solución consiste en diseñar e implementar un sistema de soporte decisiones que procesa los datos usando componentes de lingüística computacional, ejecuta un modelo de minería de texto, y expone los datos a través de un sitio web.

El proceso de limpieza y análisis semántico de los datos fue fundamental para que el modelo que clasifica una opinión en positiva o negativa obtenga un nivel de precisión del 78% al procesar 300.000 opiniones en un tiempo menor a 10 minutos.

El análisis automático de sentimientos es de bajo coste y permitió entender lo que el mercado piensa sobre los productos mejorando el servicio al cliente y la reputación de la marca a través de estrategias que generan una percepción positivo basado en información útil acerca de las preferencias de compra, gustos y disgustos revelados por el análisis de sentimiento.

ÍNDICE GENERAL

ABREVIATURAS Y SIMBOLOGÍA

API	Application Program Interface
DSS	Decision Support System
NLP	Natural Language Processing
PLN	Procesamiento Natural del Lenguaje

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El rápido crecimiento de la Web 2.0 nos ha dejado como resultado una gran cantidad de información que por su inmediatez y espontaneidad tienen datos reveladores sobre la opinión del mercado acerca de nuestros productos. Estos datos por su volumen, variedad y forma hacen poco eficiente su análisis por medios humanos por lo que se requiere automatizar este proceso conocido como análisis de sentimiento.

La automatización del análisis de sentimiento tiene la capacidad procesar grandes volúmenes de datos con un mínimo retardo y una precisión aceptable a un bajo coste.

El departamento de mercadeo de una empresa que vende productos en una tienda de comercio electrónico necesita de un sistema de soporte decisiones que automatice el proceso de análisis de sentimiento. Los datos están disponibles en la base de datos, en las redes sociales como Twitter, Facebook, y en blogs especializados de acuerdo a la línea de productos. La extracción de estos datos es diaria y se concentra en una base de datos.

La plataforma de desarrollo escogida es R por su facilidad de extraer información de múltiples fuentes de datos y la alta variedad de librerías para minería de datos. Los paquetes usados para la extracción de la información son “twitterR”, “facebookR”, y “jsonlite”. Para la limpieza y análisis semántico se usó el paquete de procesamiento de natural del lenguaje “tm”. El modelo de análisis de sentimiento fue desarrollado con el clasificador “naive bayes” a través de los paquetes “e1071” y “RTextTools”. Los resultados del análisis son mostrados al usuario a través de un sitio desarrollado en la plataforma Shiny.

Este documento describe en su capítulo 1 el planteamiento del problema y la solución propuesta, junto con los objetivos. En el capítulo 2 se detalla la metodología aplicada en las siguientes fases: extracción, limpieza, modelo de análisis de sentimiento, pruebas y la presentación de los resultados. En el capítulo 3 se analizan los resultados para determinar las conclusiones y recomendaciones.

CAPÍTULO 1

GENERALIDADES

1.1 Descripción del Problema

El rápido crecimiento de la Internet y las redes sociales junto con la habilidad de los usuarios de crear y publicar contenido ha desarrollado comunidades electrónicas que proveen una extensa información sobre las características, precios, promociones, fortalezas y debilidades de un producto e inclusive un servicio. Sin embargo, el alto volumen de opinión que son publicados sobre un producto están dispersos por toda la Internet lo que hacen difícil la tarea de entender la verdad que existe sobre la calidad y reputación de una marca o producto.

La disponibilidad de esta información ha desplazado el poder de negociación a los clientes al tener la facultad de decidir la compra de un producto basado en la experiencia de otros usuarios. Cuando esta experiencia es negativa se crea una mala reputación que afecta a las ventas y sin una reacción oportuna el producto está destinado a desaparecer del mercado.

La automatización de un análisis de opinión es una aplicación de la minería de texto que está compuesto de áreas multidisciplinarias como extracción de información, aprendizaje automático, estadísticas y lingüística computacional. Sin una herramienta de software es complejo realizar un análisis de opinión para diferentes productos a través de varias perspectivas o dimensiones de análisis.

Una opinión expresada en datos no estructurados tiene algunas complicaciones que requieren de la ayuda de componentes de software de procesamiento natural del lenguaje (Natural Processing Languaje, NLP) para determinar ambigüedades en el léxico de las oraciones de una opinión que en su mayoría son subjetivas.

1.2 Solución Propuesta

Usando estratégicamente las opiniones puede convertiste en una poderosa herramienta para persuadir a los consumidores a comprar y además crear una relación de confianza, pero se requiere de un

proceso automatizado a través de una aplicación de software que extraiga la información de múltiples fuentes de datos, la cargue a un almacén de datos y la transforme para su posterior análisis. La aplicación permite hacer un análisis exploratorio usando técnicas de estadística descriptiva y gráficos de relaciones de los términos asociados con la percepción del producto. El análisis de opinión consiste en aplicar un modelo de minería de texto para clasificar los documentos de opinión una escala de 3 (negativo, neutro, positivo) o 5 niveles (muy negativo, negativo, neutro, positivo, muy positivo) usando algoritmos basados en léxico. Los resultados son mostrados en forma gráfica y como informes.

La aplicación además considera aspectos generales cualquier tipo de software como seguridad a nivel de autenticación y autorización, ya que existen usuarios que pueden analizar ciertos datos por ejemplo una región especifica.

La versión inicial de la aplicación necesita ser simple y rápida de tal forma que permita validar rápidamente su funcionamiento. Por esta razón va a ser desarrollada en R para la carga de información desde archivos planos, redes sociales en formato JSON, y la base de datos que puede ser Oracle o PostgreSQL. El resultado del análisis es expuesto en una interfaz WEB basada en Shiny.

Los beneficios de la solución propuesta son:

a. Analizar un alto volumen de información disponible en texto como opiniones que permita entender lo que piensan los clientes sobre un producto.

b. Permitir tomar decisiones oportunas cuando la reputación de un producto es negativa y explotar las oportunidades disponibles cuando la valoración es positiva.

c. Extraer y limpiar la información desde múltiples fuentes de datos.

d. Abstraer la complejidad de modelos estadísticos a usuarios finales incrementando su productividad.

1.3 Objetivo General

Implementar un modelo de minería de opinión (análisis de sentimientos) de productos y contenidos WEB para el soporte de decisiones del departamento de marketing.

CAPÍTULO 2

METODOLOGÍA DE DESARROLLO

2.1 Definición de la Situación Actual

La venta de productos a través de una tienda electrónica genera una gran cantidad de información no estructurada disponible en las opiniones de los productos y en la red social Twitter. También existe información interna generada por encuestas a empleados y grupos de opinión.

El departamento de Marketing necesita obtener información sobre la percepción de un producto analizando una gran cantidad de opiniones de los clientes en el menor tiempo posible y a un bajo coste. Actualmente está información que se encuentra en diferentes fuentes

de datos no es explotada correctamente ya que el análisis es superficial sobre unas pocas opiniones. Hacer un análisis sobre toda la información requiere de mayores recursos y llevaría semanas con un alto riesgo de que la información no sea exacta.

Analizar las opiniones es de vital importancia porque permitiría corregir errores cuando la percepción de un producto no es la esperada evitando una mala reputación y por ende el retiro del producto del mercado. Cuando una opinión es positiva se abre las puertas a nuevas oportunidades a ser explotadas con el afán de incrementar la confianza con el cliente.

El sitio WEB mantiene las opiniones por producto en una base de datos relacional e incluye la fecha y hora de creación, el cliente que lo creo, y la descripción de la opinión. Con el cliente se puede saber otros atributos como la ubicación geográfica del cliente.

Además se hacen campañas de mercadeo en las redes sociales Facebook y Twitter en donde los clientes también opinan sobre los productos especialmente cuando el departamento de mercadeo necesita saber la percepción sobre un tema en particular induce a la opinión a través de preguntas, encuestas o inclusive promociones.

2.2 Análisis de la Situación Actual

Analizar opiniones requiere de técnicas de minería de datos, específicamente de ***minería de opinión*** o análisis de sentimientos que es el campo de estudio que tien-e como objetivo extraer las opiniones y sentimientos de texto en lenguaje natural utilizando métodos computacionales [1].

Una ***opinión*** es una oración subjetiva que describe lo que una persona cree o piensa sobre algo [2]. Una representación básica de la opinión contiene:

a. Portador, es el individuo o grupo dueño de la opinión.

b. Objetivo, define de qué se trata la opinión.

c. Contenido, especifica que exactamente es la opinión.

Además una opinión enriquecida está representada por:

d. Contexto, detalla bajo qué situación la opinión es expresada, por ejemplo tiempo, ubicación, etc.

e. Sentimiento, nos dice los sentimientos del portador acerca de la opinión.

El ***léxico de sentimiento*** son las palabras o frases de sentimiento que representan un sentimiento negativo o positivo. Ejemplos de palabras positivas son, bueno, excelente, estupendo. Las palabras negativas pueden ser mala, pobre, terrible. Además de las palabras individuales las frases o modismos también forman parte del léxico de sentimiento, por ejemplo, “me costó un ojo de la cara”.

Las frases y palabras del léxico de sentimiento son uno de los indicadores más importantes del análisis pero aún están lejos de la precisión requerida para por un análisis de sentimiento. El problema es más complejo si consideramos los siguientes inconvenientes:

Una palabra o frase puede tener diferentes orientaciones o polaridades en diferentes contextos. Por ejemplo la palabra delgado puede ser positivo y negativo dependiendo del contexto:

a. Este teléfono inteligente es súper delgado

b. La cubierta de la computadora es muy delgada

Una opinión contiene palabras de sentimiento pero no expresa sentimiento alguno. Este fenómeno se da principalmente en preguntas y oraciones condicionales. Por ejemplo la palabra buena no representa sentimiento:

a. ¿Puedes decirme que cámara es buena?

b. Si encuentro una buena cámara la voy a comprar.

Las opiniones sarcásticas que con o sin palabras de sentimiento son difíciles de analizar. Por ejemplo:

a. ¡Qué lindo carro!, ni una semana de comprado dejo de funcionar.

Algunas opiniones sin palabras o frases de sentimiento pueden expresar un sentimiento positivo o negativo.

a. Esta lavadora usa demasiada agua.

b. Después de usar por primera vez los zapatos la suela se desprendió.

La minería de texto usa ***procesamiento del lenguaje natural*** (PLN) para extraer el sentido semántico de un bloque de texto. El análisis de sentimiento no necesita entender completamente cada oración o documento, sino comprender solo algunos aspectos como opiniones negativas o positivas y sus objetivos [1].

La solución propuesta requiere del diseño de un ***sistema de soporte de decisiones*** (DSS) orientado a los datos [3] que extraiga información útil de la tienda en línea, de las redes sociales y de sus procesos internos para analizarlos y determinar la polaridad de las opiniones de

productos de tal forma que los gerentes puedan tomar mejores decisiones sobre su plan de lealtad al cliente.

2.3 Diseño de la Solución

La solución propuesta es una aplicación WEB que en su primera versión utiliza el marco de trabajo Shiny [4] dado su simplicidad para publicar el resultado del análisis lo que nos permitirá obtener la retroalimentación de los usuarios en menor tiempo. La aplicación será distribuida y alojada usando el servicio de hosting de RStudio.

El modelo de minería de opinión esta implementado en el lenguaje R, en donde se utilizara los siguientes paquetes:

a. RTextTools, una librería de máquina de aprendizaje para clasificar texto [5].

b. e1071, implementa el clasificador naiveBayes [6].

c. tm, la infraestructura de minería de texto de R [7].

d. wordCloud dibuja una nube de palabras a través de múltiples documentos [8]

El repositorio central de los datos es una base de datos PostgreSQL. Además se extrae información de Twitter y de Facebook. Se utilizan los siguientes paquetes en R.

a. RPostgreSQL [9], es la interface para conectarse al repositorio central.

b. twitteR [10], provee la interface a la web API de Twitter desde R, para extraer el texto de opinión y llevarlo a la base de datos para su análisis.

c. Rfacebook [11], provee la interface a la API de Facebook desde R, para extraer el texto de opinión y llevarlo a la base de datos para su análisis.

Para el desarrollo de la aplicación se va a realizar en etapas o iteraciones de acuerdo a lo siguiente:

1. Extracción, transformación y carga de la Información
2. Definición del Modelo de Minería de Opinión
3. Plan de pruebas
4. Plan de Implementación

2.4 Extracción, transformación y carga.

Familiarizare con los datos es una de las primeras tareas en el proceso de minería de datos [12], por lo que es importante entender donde se

encuentran, en que forma están y cómo van a ser cargados para su posterior análisis.

Los datos de opinión estarán centralizados en una base de datos PostgreSQL en una tabla que además contiene información como la ubicación o lugar de la nota, el cliente que hace la nota, el producto al que se le crea la nota, la familia de producto a la que pertenece, la fuente de datos de donde proviene y la fecha/hora de creación de la nota.

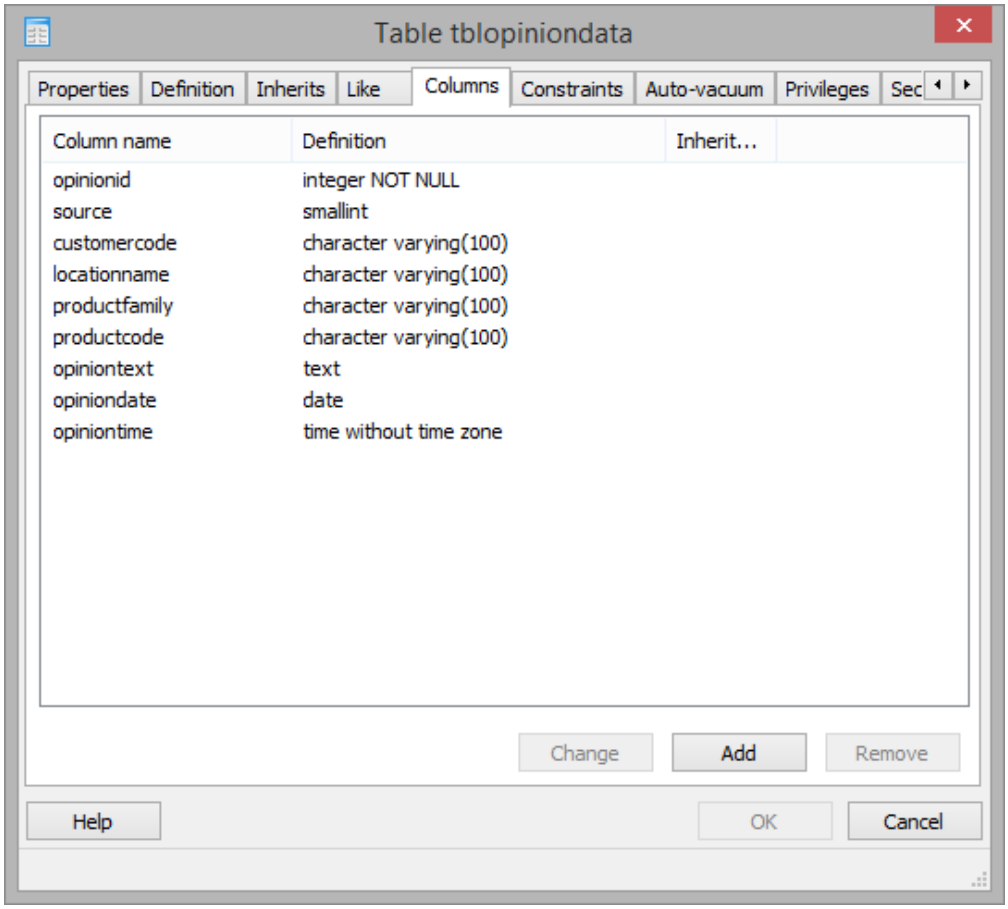

Figura 2.1Tabla de central de opiniones

La extracción de datos de Twitter se hace a través del paquete twitteR [9], para lo cual se requiere crear una cuenta de desarrollador y una aplicación con acceso de lectura y escritura. El código de autenticación en R es el siguiente:

loadTwitterData.R*

Source on Save

```
1  #Authenticate to Twitter
2  #Created by Juan Carlos Garcia on 19/07/2015
3
4  install.packages("twitteR")
5  library(twitteR)
6  ck = "xxxxx"
7  cs = "xxxxxxx"
8  at = "xxxxxxxxx"
9  as = "xxxxxxxxxxx"
10 setup_twitter_oauth(consumer_key = ck,consumer_secret = cs
11                     , access_token = at, access_secret = as)
```

Figura 2.2 Crear una conexión con twitteR

El formato que de los datos es una lista que será transformada y grabada en la base de datos.

```
> searchTwitter(searchString = "iPhone 6", lang="es",locale = "ec",
since = "2015-07-01",until = "2015-07-18",n = 4)
[[1]]
[1] "CataMujica_: RT @fefevallejoss: Quiero el iPhone 6"

[[2]]
[1] "fefevallejoss: Quiero el iPhone 6"

[[3]]
[1] "Eric9SC: RT @LuzuVlogs_: Juega conmigo GRATIS, puedes ganar un
a PS4, iPhone 6, iPad, PCs gamer, dinero.\n\nEntra ya: http://t.co/
vGXxCwl5Af  ← http://…"

[[4]]
[1] "IgnacioBJandRM: Necesito el Iphone 6 ya"
```

Figura 2.3 Datos de opinión extraídos de twitter para el IPhone 6

La extracción de datos de Facebook se utiliza el paquete RFacebook [10]. También requiere de una cuenta de desarrollador y la creación de una aplicación para obtener las claves de acceso y poder crear el token de autenticación.

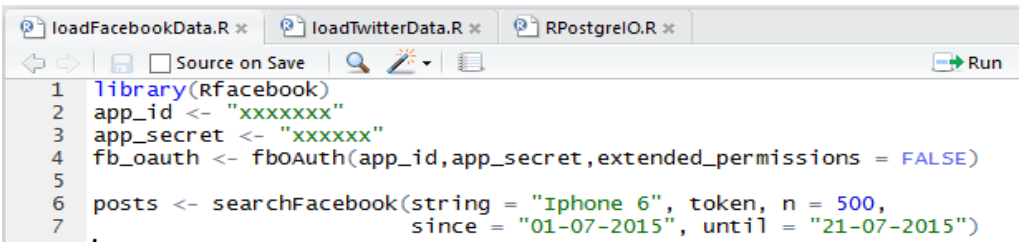

Figura 2.4 Conexión con la API de Facebook para extraer publicaciones

El proceso de extracción tiene una frecuencia diaria y se obtiene desde el sitio web, Twitter y Facebook. Después de varias pruebas durante el proceso de decisión del modelo se determinó que el proceso de transformación y limpieza de la información es el siguiente:

1. Remover los números.
2. Remover los símbolos.
3. Remover los espacios en blanco que están demás.
4. Reemplazar las vocales con tilde a vocales sin tilde.
5. Transformar las palabras a minúscula.

No se requiere modificaciones lingüísticas como reducir una palabra a su raíz (stemming) ni tampoco agrupar palabras en conjunto de sinónimos (WordNet) ya que de acuerdo a las pruebas estos degradaban los resultados del análisis.

El proceso de transformación y limpieza utiliza el paquete en R llamado “tm” (del inglés text mining) [6]. Una vez que la información está limpia se procede a grabarla en la base de datos PostgreSQL usando el paquete RPostgreSQL [8] para su posterior análisis.

2.5 Modelo de minería de opinión.

El análisis de sentimiento requiere clasificar las opiniones determinando si es positiva o negativa. El clasificador utilizado para la definición del modelo es Naive Bayes (Bayes ingenuo) debido a su rapidez de procesamiento y nivel de precisión adecuado. Se realizó la evaluación de otros algoritmos estándares de aprendizaje de maquina (del inglés Machine Learning) como Máxima entropía (del inglés Maximun Entropy) y Máquinas de vectores de soporte (del inglés Support Vector Machine), pero su nivel de precisión no se compensa con la capacidad de procesamiento requerida por lo que fueron descartados.

El clasificador Bayes ingenuo se basa en una característica específica de tal manera que contribuye de manera independiente a una

probabilidad. Este clasificador es una técnica eficiente de aprendizaje supervisado y lo vamos a implementar usando el paquete en R denominado e1071.

Estos algoritmos tienen filosofías un poco diferentes y han demostrado efectividad en estudios previos de categorización de texto [12]. La implementación de estos algoritmos se hizo con datos de opiniones en forma de unigramas usando técnicas de procesamiento natural del lenguaje. Bigramas, trigramas y n-gramas fueron excluidos del análisis ya que no tienen mejores resultados que los unigramas y en algunos casos tienen un desempeño menor.

Los datos se encuentran en la base de datos que es alimentada por opiniones para ciertos productos que requieren mayor análisis por el departamento de mercadeo. Las opiniones no están renqueadas ni cualificada solo tienen ciertos atributos como la fecha y la región desde donde fue creada, ya que se necesita conocer si una opinión tiene el mismo patrón es diferentes localidades. El análisis de sentimiento se realizó por producto que en promedio tenían 726 opiniones positivas y 627 opiniones negativas.

2.6 Prueba del modelo

El tamaño del conjunto de datos para entrenamiento es del 70% y el 30% adicional se usó como conjunto de prueba para evaluar el

desempeño del modelo. La evaluación del modelo empieza con el cálculo de los criterios, precisión y sensibilidad que se basan en la comparación entre los resultados del modelo con los resultados reales determinados por el usuario.

El indicador final de evaluación del modelo será la medida-F (del inglés F-score) que combina los resultados de la precisión y sensibilidad.

Tabla 1 Tabla de contingencias

Datos Reales	Resultado del Modelo	
	Sistema (S)	Sistema (N)
Usuario (S)	VP Verdadero Positivo	FN Falso Negativo
Usuario (N)	FP Falso Positivo	VN Verdadero Negativo

La precisión mide la probabilidad de opiniones positivas que son correctas y está dado por la división entre las decisiones correctas y la suma de las decisiones positivas correctas más las decisiones positivas incorrectas.

2.1

$$Precision = \frac{VP}{VP + FP}$$

La sensibilidad mide la proporción de las opiniones que son identificadas correctamente por el modelo. Esta dado por el número de

opiniones positivas correctas sobre la suma de las opiniones correctas más las opiniones negativas correctas.

2.2
$$Sensibilidad = \frac{VP}{VP + FN}$$

El indicador medida-F será el empleado para determinar el modelo a escoger para el análisis de sentimiento. El parámetro β es igual a 1.

2.3
$$medida - F = \frac{2 * Precision * Sensibilidad}{Precision + Sensibilidad}$$

El modelo también nos permite ver la relación entre los términos más frecuente donde podemos confirmar si el sentimiento es positivo o negativo. Por ejemplo con datos extraídos de Twitter sobre el IPhone durante el primer semestre del 2015 el modelo revela que el sentimiento es positivo y los términos más frecuentes son:

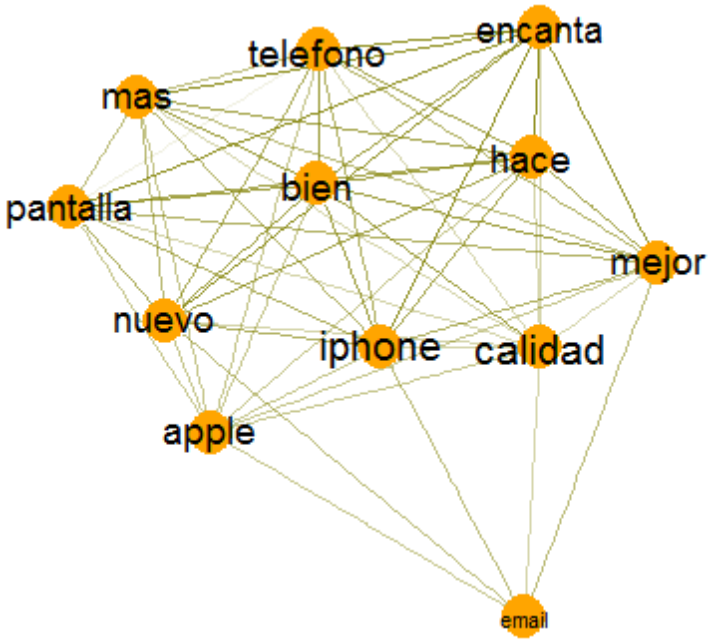

Figura 2.5 Análisis de Sentimiento para IPhone en Twitter 2015

2.7 Plan de implementación

El plan de implementación de la solución se realizó en base a un análisis de riesgos realizado en conjunto entre el equipo de desarrollo y los usuarios de la aplicación.

Tabla 2 Riesgos de la solución propuesta

Riesgo	Grado	Acción
El modelo no provee los resultados esperados	Alta	Realizar un plan piloto
La seguridad de la base de datos no es la adecuada	Alta	La base de datos estará en los servidores locales
El acceso no autorizado a la aplicación	Alta	La aplicación WEB requiere de credencial encriptada

Múltiples usuarios trabajando al mismo tiempo podría afectar a la aplicación.	Alta	Realizar una prueba de concurrencia
El volumen de información afecta el rendimiento	Media	Realizar pruebas de rendimiento
Desconocimiento de los usuarios	Media	Realizar el entrenamiento respectivo

Se requiere de una prueba piloto con el responsable del departamento de mercadeo que es el usuario principal de la aplicación. En este punto también se harán pruebas de concurrencia y rendimiento de la aplicación en general. Los tiempos están dados por la siguiente tabla:

Tabla 3 Plan de Implementación de la Solución

Tarea	Tiempo (días)	Responsable
Instalar la aplicación en Shiny	1	Desarrollador
Instalar la base de datos	1	Desarrollador
Realizar prueba piloto	3	Jefe de mercadeo
Realizar pruebas de concurrencia y rendimiento	2	Desarrolladores
Realizar los ajustes	3	Desarrollador
Entrenar a los usuarios	1	Jefe de mercadeo

Terminar implementación	1	Desarrollador
Total	12	

CAPÍTULO 3

ANÁLISIS DE LOS RESULTADOS

3.1 Efectividad del Modelo de Opinión

La precisión del modelo se determinó usando las medidas de precisión y sensibilidad, combinándolas al final para obtener el porcentaje de precisión total a través de la medida F. Para un conjunto de datos de 1353 opiniones con el 70% de datos de entrenamiento y el restante 30% de pruebas se obtuvieron los siguientes resultados.

Tabla 4 Resultados del sistema comparado con los datos reales

1353	Sistema negativo	Sistema positivo
Real negativo	561	66
Real positivo	231	495

El porcentaje de precisión del sistema es del 77% con una tendencia a mejorar dependiendo de la calidad del texto de opinión y mayor tamaño del conjunto de datos.

Tabla 5 Conjunto de datos y porcentaje de precisión

Opiniones	1353
Datos Entrenamiento	947
Datos Prueba	406
Precisión	88%
Sensibilidad	68%
Medida F (Total)	**77%**

3.2 Recursos requeridos

La aplicación tiene un rendimiento óptimo considerando el tiempo de respuesta y la capacidad de hardware requerida para procesar un conjunto de datos. Con un servidor básico de 16GB de memoria RAM, un procesador Intel i7 de 8 núcleos, se puede realizar el proceso de entrenamiento de 300.000 opiniones en un tiempo de 5-10 minutos. Para conjunto de datos de mayor capacidad se recomienda mayor memoria RAM.

El proceso de entrenamiento se hace bajo demanda, pero cuando el usuario desea usar el modelo para predecir un conjunto de opiniones el resultado es en segundos.

3.3 Análisis de costos

El proceso manual de análisis de sentimiento tiene una estimación de efectividad del 86% pero con una limitada cantidad de datos. En promedio se obtienen alrededor de 20.000 opiniones nuevas a la semana y se estima que una persona trabajando a tiempo completo puede procesar el 25% de ellas.

Para procesar el 100% de las opiniones se requieres de 4 personas a tiempo completo, sin considerar que el crecimiento de nuevos registros crece sosteniblemente a medida que se incorporan nuevos clientes.

Tabla 6 Costo primer año del proceso manual

Opiniones (mes)		80.000
Opiniones x persona		19.200
Personas requeridas		4
Sueldo	$	500
Costo mensual	$	2.083
Costo Manual	**$**	**25.000**
Precisión Manual		86%

Con el sistema de análisis de sentimiento los costos son considerablemente menor comparados al primer año de operación. Los costos fijos de desarrollo incluyen entrenamiento y hardware, ya que el costo de licencia es cero porque se usan herramientas open source. Se incluye un costo mensual de mantenimiento de hosting del sitio web.

Tabla 7 Costo primer año del proceso manual

Licencias	0
Entrenamiento	1000
Hardware	2500
Costo Mensual	1200
Costo Sistema	**$ 4.700,00**
Precisión Sistema	77%

Sin embargo la diferencia en precisión es 9% por lo que se justifica la inversión en el sistema, más aún si con el hardware y rendimiento mencionado anteriormente se asegura la escalabilidad.

3.4 Beneficios de la solución

Desde el punto de vista de negocios existen beneficios importantes que le permite al departamento de marketing tomar decisiones adecuadas en el tiempo oportuno al tener la capacidad de:

- Mejorar el servicio al cliente creando estrategias para generar un sentimiento positivo basado en información útil acerca de las preferencias de compra, gustos y disgustos sobre los productos.

- Cuantificar las percepciones sobre los productos que permiten crear estrategias que mejoren la reputación de la marca.

- Analizar desde otra perspectiva el impacto de las campañas de mercadeo validando su efectividad con el sentimiento generado.

- Cruzar la información de sentimiento con las ventas históricas para así establecer un pronóstico de mejor precisión.

- Conocer los sentimientos que rodean a los competidores comparándose con el desempeño de los demás.

- Crear nuevas oportunidades de negocio haciendo un buena análisis se puede detectar tendencias que explotadas correctamente podrían impactar el mercado en beneficio de la organización.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

1. El análisis de sentimiento constituye una herramienta importante de mercadeo que las empresas deben explotar para el mejoramiento del servicio al cliente, la consolidación de la lealtad de una marca y/o productos, y la detección de amenazas que a medida que pasa el tiempo será más costoso de mitigar.

2. El avance de la tecnología en áreas como minería de datos, lingüística computacional hacen posible crear sistemas de toma decisiones que analicen gran cantidad de información disponible en la web con un nivel de precisión aceptable.

3. Un sistema automático de análisis de sentimiento se justifica por su capacidad de analizar un gran volumen de información en tiempo real a un costo bajo.

4. La empresa puede utilizar el resultado de la información analizada para correlacionar con otras áreas como las ventas y así establecer mejores pronósticos incrementando la eficiencia.

5. Haciendo un buena análisis de la información obtenida se puede detectar tendencias que lleven a la organización a establecer nuevas oportunidades

Recomendaciones

1. Se recomienda hacer un mejoramiento continuo de la aplicación agregando nuevas funcionalidades como detección de ironía y opiniones subjetivas más complejas.

2. Crear léxicos de mejor calidad en idioma español ya que de esto depende el nivel de precisión de precisión y la variedad de los resultados.

3. Hacer énfasis en la calidad de los datos obtenidos desde distintas fuentes, se entiende que son datos no estructurados, pero agregando herramientas de autocompletar por ejemplo incrementaría considerablemente la claridad con que los clientes expresan sus opiniones.

BIBLIOGRAFÍA

[1] B. Liu, Sentiment Analysis, Estados Unidos: Cambridge University Press, 2015.

[2] C. Zhai, «Text Mining and Analytics,» Coursera, Chicago, 2015.

[3] K. C. Laudon y J. P. Laudon, «Sistemas de Información Gerencial,» Pearson Educación, Mexico, 2012.

[4] RStudio, «Shiny Web application framework for R,» [En línea]. Available: http://shiny.rstudio.com/. [Último acceso: 18 07 2015].

[5] T. P. Jurka, L. Collingwood, A. E. Boydstun, E. Grossman y W. v. Atteveldt, «RTextTools: Automatic Text Classification via Supervised Learning,» 2012. [En línea]. Available: http://CRAN.R-project.org/package=RTextTools. [Último acceso: 18 07 2015].

[6] D. Meyer, «Package e1071,» Misc Functions of the Department of Statistics, Probability Theory Group, 16 07 2015. [En línea]. [Último acceso: 01 08 2015].

[7] I. Feinerer y K. Hornik, «tm: Text Mining Package. R package version 0.6-2,» 2015. [En línea]. Available: http://CRAN.R-

project.org/package=tm. [Último acceso: 19 07 2015].

[8] I. Fellows, «Word Clouds,» [En línea]. Available: https://cran.r-project.org/web/packages/wordcloud/index.html. [Último acceso: 18 07 2015].

[9] D. E. Joe Conway, T. Nishiyam, S. K. P. (. 2. y N. Tiffin, «RPostgreSQL: R interface to the PostgreSQL database system,» 27 03 2013. [En línea]. Available: https://cran.r-project.org/web/packages/RPostgreSQL/index.html. [Último acceso: 19 07 2015].

[10] J. Gentry, « twitteR: R Based Twitter Client. R package version 1.1.8,» 11 02 2015. [En línea]. Available: http://CRAN.R-project.org/package=twitteR. [Último acceso: 2015 07 2015].

[11] P. Barbera y M. Piccirilli, «Rfacebook, Access to Facebook API via R,» 02 04 2014. [En línea]. Available: https://cran.r-project.org/web/packages/Rfacebook. [Último acceso: 19 07 2015].

[12] J. Han, M. Kamber y J. Pei, Data mining: concepts and techniques, Morgan Kaufmann, 2012.

[13] B. Pang y L. Lee, «Thumbs up? Sentiment Classification using Machine Learning Techniques,» 2002.