UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA

TESIS DE GRADO

"RECONOCIMIENTO DE PATRONES DE POLARIDAD EMOCIONAL PARA IDENTIFICAR NICHOS DE MERCADO BASADO EN MINERÍA DE TEXTOS"

PARA OPTAR AL TÍTULO DE LICENCIATURA EN INFORMÁTICA

MENCIÓN: INGENIERÍA DE SISTEMAS INFORMÁTICOS

POSTULANTE: LUIS ALFREDO MAMANI MOLLERICONA

TUTOR METODOLÓGICO: M.Sc. ALDO RAMIRO VALDEZ ALVARADO

ASESORA: M.Sc. ROSA FLORES MORALES

LA PAZ – BOLIVIA

2016

UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA

LA CARRERA DE INFORMÁTICA DE LA FACULTAD DE CIENCIAS PURAS Y NATURALES PERTENECIENTE A LA UNIVERSIDAD MAYOR DE SAN ANDRÉS AUTORIZA EL USO DE LA INFORMACIÓN CONTENIDA EN ESTE DOCUMENTO SI LOS PROPÓSITOS SON ESTRICTAMENTE ACADÉMICOS.

LICENCIA DE USO

El usuario está autorizado a:

a) visualizar el documento mediante el uso de un ordenador o dispositivo móvil.
b) copiar, almacenar o imprimir si ha de ser de uso exclusivamente personal y privado.
c) copiar textualmente parte(s) de su contenido mencionando la fuente y/o haciendo la referencia correspondiente respetando normas de redacción e investigación.

El usuario no puede publicar, distribuir o realizar emisión o exhibición alguna de este material, sin la autorización correspondiente.

Dedicatoria

Dedico la presente tesis:

A mis padres Andres Mamani y Modesta Mollericona que en el camino de mi vida son el faro que me ilumina y me guía.

A mis hermanos por el bello regalo de compartir momentos inolvidables.

A mi familia por apoyarme con paciencia, amor y concejos, por siempre confiar en mí, impulsándome para seguir adelante.

¡Con mucho cariño para ustedes!

Luis Alfredo Mamani Mollericona

AGRADECIMIENTOS

A Dios por darme el regalo de la vida, y con ella conocer personas que siempre han estado en los buenos y malos momentos.

A mi familia por todo el apoyo brindado para que pueda llegar a concluir mis metas.

A la M.Sc. Rosa Flores Morales por guiarme en este proceso de aprendizaje y brindarme su apoyo, gracias por revivir en mí las ganas de investigar, por la paciencia, la comprensión y por ser ejemplo de docente.

A la licenciada Menfy Morales Rios por la confianza depositada en mi persona, gracias por ser ejemplo de lucha y por la comprensión en las situaciones difíciles.

Al M.Sc. Aldo Valdez por ser un líder a seguir, gracias por guiarme en el proceso del desarrollo de la presente tesis, por la pasión hacia la carrera y por impulsarnos a ser mejores cada día.

A todos mi amigos del proyecto DUT y aquellos que me acompañaron y brindaron concejos, gracias por todo el apoyo, por las risas, por esos momentos que marcaron mi vida y por ser como son, “únicos”.

¡Gracias a todos, sin ustedes no sería la persona que soy!

RESUMEN

Las redes sociales digitales llegan a formar parte de nuestro diario vivir. Las conversaciones, las noticias que compartimos, las felicitaciones, los consejos que damos a nuestros amigos en estos medios de comunicación digital pueden develar ciertas características personales, gustos, preferencias que pueden ser usadas para identificar posibles nichos de mercado o incluso para modificar ciertos productos para que se adecuen a las necesidades de la clientela.

El marketing digital conocido tradicionalmente por colocar anuncios publicitarios en periódicos, canales de televisión, estaciones de radio, carteles o posicionados en la parte alta de algún edifico, hoy en día fueron sustituidos por el marketing realizado a través de internet, en redes sociales digitales y más específicamente en Facebook. El mismo brinda herramientas para identificar a qué público llegar, rango de edades, pero que sin embargo carece de herramientas que permitan conocer a nuestros posibles clientes e identificar preferencias hacia cierto tipo de características, patrones que puedan servir para llegar a ellos con los productos adecuados.

La minería de textos busca estos patrones, difíciles de encontrar con los métodos tradicionales, la misma que permitió desarrollar herramientas en español para colectar textos de la red social de Facebook, obtener información en archivos, que luego pasaron por un proceso de limpieza y eliminación de todo el ruido que ellas contienen, a través de las fases del procesamiento del lenguaje natural aplicadas al lenguaje español, además de la construcción de diccionarios de verbos y sinónimos. Lo que permitió visualizar características particulares de cada usuario. Posteriormente se procedió a categorizar cada uno de los comentarios de acuerdo a su polaridad, permitiendo identificar nichos de mercado para cierto tipo de productos de acuerdo a las características particulares de los usuarios. Los programas graficadores sirvieron para ver cierto tipo de tendencia del usuario y al mismo tiempo para elegir los productos de mayor aceptación para el público objetivo.

Palabras clave: Minería de textos, análisis de sentimiento, marketing digital

ABSTRACT

Digital social networks become part of our daily lives. The conversations, the news that we share, the congratulations, the advice we give to our friends in these digital media can reveal certain personal characteristics, tastes, preferences that can be used to identify possible market niches or even to modify certain products so they are adapted to the needs of the customers.

The digital marketing traditionally known for placing advertisements in newspapers, TV channels, radio stations, posters or positioned in the upper part of a building, nowadays they were replaced by marketing conducted through the Internet, in digital social networks and more specifically on Facebook. It provides us with tools to identify which audience to reach, range of ages, but nevertheless it lacks of tools that allow us to know our potential customers and to identify preferences towards certain types of characteristics, patterns that can serve us to reach them with the right products.

Text mining seeks these patterns, hard to find with traditional methods, the same that allowed us to develop tools in Spanish to collect texts from the social network of Facebook, to obtain information in archives, which then went through a process of cleaning and elimination of all the noise they contain, through the phases of natural language processing applied to the Spanish language, in addition the construction of dictionaries of verbs and synonyms. What allowed to visualize particular characteristics of each user. Subsequently each comment was categorized according to its polarity, allowing to identify market niches for certain types of products according to the user's particular characteristics. The graphics software served to see a certain type of trend of the user and at the same time to choose the products which could have the greater acceptance for the target public.

Keywords: Text Mining, sentiment analysis, digital marketing

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPITULO I

MARCO REFERENCIAL

1.1. INTRODUCCIÓN

La existencia de datos no estructurados en grandes cantidades y en diferentes formatos, generados por la expansión exponencial de internet en las últimas décadas, posibilitan tener una gran cantidad de información textual a nuestro alcance, para poder analizar, procesar y vislumbrar patrones en el contenido.

Las redes sociales llegan a ser el principal medio de comunicación y generación de datos en la actualidad. Este fenómeno se debe a la expansión del internet y la era de la telefonía móvil. Nuestro diario vivir ahora se ve afectado por un nuevo sistema de comunicación más libre y mucho menos estructurado como son las redes sociales. Existen muchas compañías dedicadas a redes sociales con distintos objetivos que las diferencian, ahí tenemos a Facebook una de las más grandes compañías con un perfil de red social, que permite compartir noticias, comentar sobre el estado de nuestros amigos, hablar con amigos a distancia, dar me gustas a los post de contactos, transmitir videos en vivo. Esta empresa desde su creación, el 4 de febrero de 2004 en Cambridge, Massachusetts hasta la fecha ha acumulado una cantidad exorbitante de 1590590 millones de usuarios (Moreno, 2016). Esta cantidad de datos permiten la actualidad analizar los comportamientos de los usuarios, saber lo que podrían pensar o quizá predecir posibles evento futuros basados en técnicas de minería de textos, considerando toda la información que tiene el usuario en las redes sociales digitales.

El marketing se realiza hoy a través de redes sociales como medio de difusión de las campañas publicitarias dirigidas a los consumidores activos de estas nuevas tecnologías. Estos consumidores interactúan, opinan, comparten con el productor y viceversa, las 24 horas del día. Cada usuario se convierte en el canal de transmisión de publicidad de manera más efectiva, además de generar nuevas ideas y ser promotores de la fidelidad hacia la marca a través de las redes personales que entre los usuarios comparten.

El *f-commerce* comienza un 8 de julio de 2009 como el complemento del e-commerce, con la primera compra a través de la plataforma de *Facebook* el 8 de julio de 2009 por 1800 flowers marcando el inicio de ventas de productos a través de esta red como canal de comunicación y publicidad efectiva. Facebook juego un papel importante en la evaluación de experiencias de post compra tanto negativa como positivamente, a través del intercambio de opiniones, así los consumidores co-crean junto a los productores la imagen de la marca, posicionando a los mercados como espacios de conversación en dos sentidos (Miranda Gonzaléz, Rubio Lacoba, Chamorro Mera, & Correia Loureiro, 2013).

1.2. ANTECEDENTES

La minería de texto es el proceso de tratar la información mediante métodos matemáticos/estadísticos para poder encontrar patrones difíciles de hallar por métodos tradicionales de búsqueda en texto debido a la complejidad de las relaciones de los datos, esta tecnología debe su formulación inicial a principios de la década de los noventa y su uso ha ido incrementándose con la generación masiva de información en redes sociales como Facebook, Twitter, LinkedIn, Google+, Youtube, Vimeo, blogs, foros, comunidades, periódicos digitales.

Revisadas las fuentes bibliográficas a continuación se describen trabajos relacionados con la temática:

"***Apply Magic Sauce***", de las interacciones con las redes sociales se muestra el tipo de personalidad y las características que la componen, midiendo así el grado de franqueza, escrupulosidad, extroversión, empatía, neurotismo de cada persona mediante minería de texto e inteligencia artificial, para tal efecto se analizan los likes, estados en Facebook, twitters, datos de navegación, datos abiertos, debido a que todo lo que se busca de cierta forma queda registrado en medios digitales. Con todo lo mencionado este sistema puede identificar entre cinco tipos de personalidades, tipos de inteligencia, satisfacción en la vida, orientación política, orientación religiosa, sexualidad, y profesión (The Psychometrics Centre - Apply Magic Souce, s.f).

*"**Social networks' text mining for consumer brand sentiments***, las redes sociales como *Facebook* y *Twitter* y los blogs actualmente han llegado a ser un gran recurso para el minado de textos en campos como en la relación del cliente, el rastreo de opiniones y filtrado de texto. Toda esta información ha sido analizada usando análisis de sentimientos básicamente, procesamiento del lenguaje natural (NLP) que usa lingüística computacional y minería de texto para identificar sentimientos y clasificarlos como positivos, negativos o neutros, esta técnica es conocida en la literatura de minería de textos como el análisis de polaridad emocional (EPA), minería de opinión, minería de revisión, o extracción de valoración del texto. El análisis de sentimientos permite encontrar patrones escondidos en grandes cantidades de textos. Para la determinación del valor de sentimiento, el texto es comparado con un diccionario para determinar el peso del sentimiento. Las compañías pueden usar toda esta información para diseminar información y mediante este procedimiento pueden identificar clientes insatisfechos. Por otra parte las campañas de publicidad podrían usar los *tweets* positivos para mejorar los esfuerzos de *marketing* de la compañía, pero no se debería obviar los *tweets* negativos que servirían para detectar lo que no está correcto en un producto o servicio. El análisis de sentimiento en *Tweets* puede ser usado para identificar las preferencias de los clientes, detectar la no satisfacción hacia un producto, analizar las tendencias de los clientes, manejar la retroalimentación del cliente y mejorar las campañas de publicidad (Mostafa, 2013).

"**El uso de las redes sociales digitales como herramienta de *marketing* en el desempeño empresarial**", las redes sociales han modificado la forma en que se comunican las personas a través de internet, lo que ha despertado el interés de los encargados de *marketing* de diferentes empresas. Es difícil medir y monetizar el impacto que estos medios ofrecen al desempeño de la empresa. Las redes sociales se han convertido en el medio de comunicación más influyente de las últimas décadas, debido a que dinamizan la comunicación entre los miembros ofreciendo nuevas formas de construir relaciones, de compartir información, de generar y editar contenidos, estas interacciones contienen video, música, fotos y otros medios digitales, por todo lo mencionado con anterioridad podemos ver la importancia de las redes

sociales en la creación de nuevas relaciones cliente – empresa. El uso actual de las redes sociales en el *marketing* se debe a su bajo coste y su alta popularidad, además de la generación de marca y la posibilidad de medir la reputación de la empresa basada en todo el texto generado. Además, todos los datos compartidos se pueden usar para explorar los patrones de amistad y comportamiento de los clientes para de esta forma acercarse a ellos con mejores ofertas permitiendo no solo a las grandes empresas aprovechar esta oportunidad sino que es una herramienta para las pequeñas empresas, pymes, a un muy bajo coste. Las investigaciones realizadas muestran que las empresas deben utilizar las redes sociales digitales como herramienta de *marketing*, con una participación activa y con la ayuda de un *community manager* (Administrador de comunidad) o una agencia de *marketing* digital especializada para poder aprovechar a todo el potencial de las redes sociales digitales y toda la información que la misma nos proporciona para el análisis (Uribe Saavedra, Rialp Criado, Llonch Andreu, 2013).

"**En Bolivia, el avance del marketing digital enfrenta 2 grandes obstáculos**", la situación del *marketing* digital en Bolivia se encuentra como la de México hace 5 ó 6 años, país rezagado en esta actividad. La principal razón por la cual surge este retraso pasa por la no aceptación del *marketing* digital y todo lo que implica con el mismo. El acercamiento al cliente por medios digitales como Facebook, Twitter, páginas web, micro sitios, aplicaciones permite visualizar, compartir y recibir retroalimentación del usuario hacia el producto ofrecido, como clave para la implementación de estrategias digitales. De esta forma visualizar que el ritmo del consumo es diferente al de años atrás, constantemente estamos observando nuestros perfiles, comentarios en noticias compartidas, *Tweets*, o teniendo reuniones por video. No obstante, las empresas Bolivianas se muestran temerosas aun a este nuevo campo, y buscan resultados rápidos sin darse cuenta de que el proceso de *marketing* digital conlleva un cierto tiempo hasta lograr poder extraer lo que la gente requiere y visualizar los patrones ahí ocultos, ver que contenidos comparten entre ellos que estén relacionados con los productos que se ofrece y que palabras están relacionadas las características del mismo (Imaña & Vásquez, 2015).

"**Facebook como herramienta de comunicación y venta: un análisis desde la oferta y la demanda**" se pretende conocer la intención de los usuarios de usar esta red social como instrumento de marketing, y por otro evaluar la presencia y el uso que actualmente hacen las empresas de sus páginas de Facebook. Para ello, en primer lugar hemos realizado una encuesta a usuarios españoles y portugueses de Facebook con el objeto de describir su intención de uso de Facebook en su proceso de compra. Se evaluó el perfil en Facebook de las 100 principales marcas a nivel internacional a través de un indicador especialmente diseñado a tal efecto. De esta forma, se ofrece un análisis tanto desde la oferta (evaluación del perfil de Facebook de las empresas) como desde la demanda (intención de uso de los consumidores) que nos permite establecer conclusiones y recomendaciones a la hora de señalar la estrategia de marketing en redes sociales de las empresas (Miranda Gonzaléz, Rubio Lacoba, Chamorro Mera, & Correia Loureiro, 2013).

1.3. PLANTEAMIENTO DEL PROBLEMA

Actualmente existen nuevos medios de comunicación conocidos como las redes sociales digitales como *Facebook*, *Twitter*, *LinkedIn*, tuvieron un crecimiento abrumador en la última década. Las redes sociales digitales se han convertido en un medio de comunicación entre el consumidor y el productor, pueden además ser utilizados como instrumento de marketing digital para mejorar la imagen de la empresa a la que representan, incluso para vender productos en línea y ofrecer servicios (Miranda Gonzaléz, Rubio Lacoba, Chamorro Mera, Correia Loureiro, 2013).

El consumidor 2.0 (Ferro MA, 2015) es un potencial consumidor que no está siendo tomado en cuenta a la hora de crear productos, el mismo es un usuario muy activo y brinda mucho feedback a la empresa y permitiría así mejorar los productos para adecuarse mejor a las necesidades del consumidor y llegar a él con una publicidad más certera y eficiente y con productos adecuados al consumidor. Existe una preocupación del productor para llegar a más clientes a través de herramientas y estrategias de marketing con resultados cuantificables y de manera rápida (Imaña y Vásquez, 2015).

La publicidad en redes sociales es el escenario más explotado en la actualidad, considerado ahora un medio de comunicación bidireccional entre marcas y compradores. Ahora los consumidores no solo son pasivos, ellos interactúan, crean contenido, opinan y comparten experiencias. Estos nuevos consumidores 2.0 considerados como *prosumer,* término originado de la unión de las palabras *producer* y *consumer*. Los consumidores pasan de su papel de divulgadores a generadores de nuevas ideas, promotores de la fidelidad e incluso el rechazo hacia ciertos productos. Ahora las redes sociales pueden influir en la proceso de compra en cuatro etapas: Búsqueda de la información, valoración de las alternativas, compra y evaluación de la decisión. Toda la información del producto ahora se encuentra disponible las 24 horas del día, todos los días del año, a través de diferentes dispositivos electrónicos con acceso a internet. La valoración de las alternativas de compra se ve favorecida a través de los comentarios que otros usuarios han hecho previamente en la página de Facebook de la marca o en el intercambio de opiniones que el consumidor puede tener en su Facebook con sus amigos. (Miranda Gonzaléz, Rubio Lacoba, Chamorro Mera, Correia Loureiro, 2013).

El poco uso de las redes sociales en nuestro medio ocasionado por nuestra escasa conexión a internet que sitúa a Bolivia como la penúltima a nivel latinoamericano detrás de Paraguay y encima de Venezuela (Mundo, 2015) ha ocasionado que las empresas no fijen su mirada en las nuevas tecnologías de comunicación y puedan así ver la importancia del marketing digital para promocionar sus productos. No obstante, se está haciendo esfuerzos por rebajar las tarifas en conexiones fijas y de móviles de parte de la empresa estatal de telecomunicaciones ENTEL S.A. (Roca, 2015), posibilitando así el acceso más universal. De esta manera las empresas podrían aprovechar el bajo coste de internet para empezar a crear estrategias de marketing digital para llegar a más personas y construir la relación con el consumidor.

1.3.1. PROBLEMA CENTRAL

¿Cómo identificar a los usuarios y asociar su grado de aceptación o rechazo hacia un determinado producto?

1.3.2. PROBLEMAS SECUNDARIOS

- La red social de Facebook brinda herramientas para dirigir nuestra publicidad en un público que responde a un rango de edades, ubicaciones, pero carece de herramientas de emparejamiento con las características de cualquier producto y la identificación de nuevas características.
- Los comentarios generados por los consumidores no son analizados por parte de los productores, siendo los mismos un recurso que serviría para incrementar la calidad de los productos.
- La falta de herramientas de análisis en español de documentos digitales generados por redes sociales, blogs y otros recursos de internet escritos en español para el análisis de mercado es una limitante para explotar estos datos.
- Se realiza un estudio casi nulo para identificar al posible comprador.
- La falta de atención en las preferencias de los usuarios.

1.4. DEFINICIÓN DE OBJETIVOS

1.4.1. OBJETIVO GENERAL

Determinar patrones de polaridad emocional con minería textual en redes sociales digitales para identificar nichos de mercado.

1.4.2. OBJETIVOS ESPECÍFICOS

- Aplicar un análisis del posible público objetivo del producto.
- Aplicar análisis de sentimiento a los comentarios.
- Desarrollar herramientas de minado de textos en español, para el procesamiento de textos.
- Identificar las preferencias del consumidor y relacionarlos con las características del producto.
- Identificar características relacionadas con el producto para la creación de publicidad especifica.

1.5. HIPÓTESIS

“La minería de textos, en datos no estructurados, permite identificar un patrón de polaridad emocional hacia un determinado producto relacionado con el grado de aceptación o rechazo del mismo”

1.5.1. OPERACIONALIZACIÓN DE VARIABLES

Una vez realizada la hipótesis de la investigación, se procedió a la identificación de las variables independiente y dependiente como muestra la tabla 1.1.

Variable dependiente: Aceptación, rechazo de un producto.

Variable independiente: Polaridad emocional.

Tabla 1.1: Variables dependiente e independiente

Fuente: [Elaboración propia]

Variable	Descripción
Polaridad emocional	Negativo Positivo Neutro
Aceptación Rechazo	Grado de aceptación Grado de rechazo

1.6. JUSTIFICACIÓN

Los programas desarrollados ayudaran a la recolección de textos, limpieza de textos, procesamiento de textos en español e identificación de patrones de polaridad en textos digitales escritos en español, programas de gráficos que permitirán visualizar los textos ya procesados.

1.6.1. JUSTIFICACIÓN ECONÓMICA

La investigación proveerá herramientas que procesen textos a un bajo coste de procesamiento de textos no estructurados para determinar si un producto tiene viabilidad de ser exitoso o de poder fracasar en una determinada población.

1.6.2. JUSTIFICACIÓN SOCIAL

Mediante esta investigación se pretende ayudar a las pequeñas empresas (PYMES), negocios o emprendimientos para que puedan analizar su posible mercado objetivo con productos que tengan un alto grado de aceptación de parte del consumidor potencial y retirar aquellos que no satisfacen las necesidades del público objetivó.

1.6.3. JUSTIFICACIÓN CIENTÍFICA

La investigación permite aplicar minería de textos en nuestro medio, desarrollando herramientas de minado en español, además de generar nuevos parámetros de evaluación de polaridad emocional a fin de determinar tendencias en el mercado Boliviano.

1.6.4. JUSTIFICACIÓN TECNOLÓGICA

Debido al bajo coste en hardware y software especializado, es posible crear herramientas de minería de texto con mucha velocidad de procesamiento debido a la tecnología actual.

1.7. ALCANCES Y LIMITES

1.7.1. ALCANCES

- Alcance espacial: Perfiles, grupos y páginas en la red social de Facebook.
- Alcance poblacional: Sin restricción y cumpliendo el alcance espacial.

1.7.2. LIMITES

- El procesamiento de texto se limita a la red social Facebook en la ciudad de La Paz Bolivia.

- La siguiente investigación exploratoria muestra la identificación de patrones de polaridad emocional con el grado de aceptación, rechazo o sentimiento neutro de un producto a la vez.
- Se asignara valores a las palabras según la polaridad emocional de las mismas, sean estas positivas, negativas.
- Se polariza los comentarios como positivos, negativos o neutros.

1.8. APORTES

La minería de textos generalmente tiene un coste elevado en cuanto a tiempo para desarrollar las herramientas y recursos de procesamiento se refiere, en este sentido los aportes más importantes serán:

1.8.1. PRACTICO

- Identificación de posibles nichos de mercado para un determinado producto.
- Identificación de patrones en los usuarios, grupos o páginas de la red social de Facebook.
- Desarrollo de programas para la recuperación de datos en la red social de Facebook.
- Construcción de un diccionario de verbos con sus conjugaciones respectivas.
- Construcción un listado de *stopwords* en base a los textos extraídos de Facebook.
- Construcción de un listado de sinónimos en español.
- Desarrollo de programas de limpieza de textos.
- Desarrollo de herramientas estadísticas.

1.9. METODOLOGÍA

La metodología de investigación será exploratoria basada en el proceso KDD (*Knowledge Discovery in Databases*) definido como el "proceso no trivial de identificar patrones válidos, nuevos, potencialmente útiles y en última instancia comprensibles en los datos" (Torres Silva, 2013).

CAPITULO II

MARCO TEÓRICO

2.1. MINERÍA DE TEXTOS

La minería de textos trata de la recolección de textos en documentos de las distintas fuentes en internet como ser blogs, redes sociales, periódicos digitales, foros, comunidades. La minería de textos, sigue los mismos métodos, algoritmos que la minería de datos, pero por el contrario de esta última la minería de textos analiza el conjunto de información contenida en documentos no estructurado y sin el uso de algún gestor de base de datos, lo que permite tener mayor diversidad de documentos.

La minería textual permite extraer nuevo conocimiento a partir del análisis de corpus textuales. La minería textual debe facilitar el análisis de textos que a priori nos resultarían inmanejables debido a su tamaño. El descubrimiento de nuevo conocimiento, el procesamiento y presentación de la información disponible en grandes colecciones de documentos en un formato que facilite su comprensión y análisis es el objetivo de la minería de textos. En numerosas ocasiones la minería textual se presenta como una actividad complementaria a la minería de datos, si bien no ha logrado el impacto de esta última. La minería de datos pretende obtener información a partir de los patrones y tendencias que pueden observarse en grandes volúmenes de información estructurada. Es decir, información disponible en bases de datos relacionales. Frente a esto, la minería textual busca un mismo objetivo en corpus textuales o información no estructurada. Existe una similitud entre minería textual y de datos, ya que ambas persiguen una misma finalidad: deducir nueva información a partir de la información ya existente. El cambio que se percibe es el tipo de información que se toma como base del análisis: datos estructurados en el primer caso, e información no estructurada (texto) en el segundo además de incorporar la lingüística computacional la cual agrupa una serie de técnicas para procesar textos y tratar de hacerlos comprensibles para un ordenador. La lingüística computacional permite el análisis sintáctico y gramatical de textos en formato electrónico, la alineación e identificación

de correspondencias entre textos escritos en diferentes idiomas, etc. Sus principales resultados se han materializado en los sistemas de traducción automática. La diferencia entre minería textual y recuperación de información se encuentra en que el objetivo de ésta última es identificar los documentos relevantes para un usuario dentro de una colección. La recuperación textual parte de una representación formal de los documentos sobre los que se realizará la búsqueda, y de la formulación de las necesidades de información del usuario mediante un sistema de representación equivalente. La categorización automática es la técnica se utiliza en la minería textual para clasificar documentos en una serie de categorías. La elaboración automática de resúmenes constituye un componente clave para facilitar el proceso de análisis, cuando éste se realiza a nivel de colección en lugar de hacerlo a nivel de documento individuales. Otro componente en un sistema de minería textual es la interface de usuario a través del cual se va a visualizar la información. La interfaz deberá mostrar los datos en un formato que haga posible su interpretación y permita al usuario moverse con facilidad entre los distintos textos analizados (Eíto Brun & A. Senso, 2004).

En los últimos años se han visto el incremento de los textos digitales, considerando las páginas web, los blogs, las redes sociales, nuestra capacidad de procesar, analizar y entender la información. Las páginas indexadas en *google* fueron alrededor de un millón en el año 1998 pero se llegó a un millón en el 2000 y excedió el trillón en el 2008. Con la llegada de las redes sociales este crecimiento se elevó, debido a la creación de datos a través de estos nuevos medios de comunicación. Además el llegado de los teléfonos celulares inteligentes llego a un número no imaginado antes (Fan & Bifet, 2012).

2.1.1. DEFINICIÓN

La minería de textos hereda los mismos objetivos de la minería de datos, que trabaja con documentos estructurados, a diferencia de la minería de textos que trabaja con conjunto de datos no estructurados, pero ambas siguen la misma línea de tratamiento de la información que implica agrupamiento de los datos o textos, categorización, clasificación y asociación conceptual, sea esta presentada en números o textos (Lanzarini, y otros, 2014).

La minería de textos busca identificar patrones en colecciones de documentos y datos no estructurados, las técnicas que sigue son las mismas que la minería de datos para la identificación de los patrones (Guevara López, 2011).

La minería de textos es útil en la extracción de patrones no triviales o complejos, con el propósito de extraer conocimiento de una colección de documentos, la misma se aplica en datos no estructurados al contrario con lo que sucede con la minería de datos (Santana Mansilla, Costaguta, & Missio, 2014).

2.2. PROCESADO DE LENGUAJE NATURAL

Los estudios en lingüística o el procesamiento del lenguaje natural antes del 2000 fueron minúsculos debido a que no se contaba con la cantidad de datos en formato digital. A partir de aquella fecha los estudios incrementaron con investigaciones relacionadas a la minería de datos, la minería web y la recuperación de la información que permitieron hacer un análisis del lenguaje natural (Liu, 2012).

En la época actual de información, el manejo eficiente de este conocimiento depende del uso de todos los demás recursos naturales, industriales y humanos. Durante toda la historia de humanidad el conocimiento, en su mayor parte se comunica, se guarda y se maneja en la forma de lenguaje natural –griego, latín, inglés, español, etc. La época actual no es ninguna excepción: el conocimiento sigue existiendo y creándose en la forma de documentos, libros, artículos, aunque éstos se guardan en forma electrónica, o sea digital. El gran avance es que en esta forma, las computadoras ya pueden ser una ayuda enorme en el procesamiento de este conocimiento. Una computadora puede copiar tal archivo, respaldarlo, transmitirlo, borrarlo, como un burócrata que pasa los papeles a otro burócrata sin leerlos. Esta ciencia, en función del enfoque práctico versus teórico, del grado en el cual se espera lograr la comprensión y de otros aspectos tiene varios nombres: procesamiento de lenguaje natural, procesamiento de texto, tecnologías de lenguaje, lingüística computacional. En todo caso, se trata de procesar el texto por su sentido y no como un archivo binaria (AMPLN, 2009)

2.2.1. NIVELES DE ANÁLISIS

El texto no se procesa directamente sino se identifica una representación formal que preserva sus características relevantes para la tarea o el método a seguir. Existen diferentes niveles de análisis de texto, de acuerdo al objetivo de la investigación, podemos mencionar el nivel de documento y el de oración.

2.2.1.1. DOCUMENTO

La tarea de análisis de sentimiento en documentos implica clasificar como positivo o negativo a todo el contenido del documento. Este tipo de análisis implica que el contenido del documento expresa la opinión acerca de una sola entidad, por consecuencia este tipo de análisis no se puede aplicar a los documentos que evalúan o comparar más de una entidad a la vez (Liu, 2012).

2.2.1.2. ORACIÓN

La tarea de análisis de sentimiento sobre la oración determina si la opinión sobre la misma es positiva, negativa o neutra. Lo opinión neutral refiere a la no emisión de opinión. La opinión positiva refiere a la oración un sentimiento positivo y de forma inversa la negativa Esta clasificación está relacionada con la subjetividad que separa las oraciones objetivas que expresan información objetiva de las oraciones subjetivas que expresan opiniones pero no la limita a solo ese análisis para determinar el sentimiento de la oración (Liu, 2012).

2.2.2. LA RECUPERACIÓN DE LA INFORMACIÓN

La aplicación del procesamiento de lenguaje natural más obvia y quizá más importante en el momento actual es la búsqueda de información (se llama también recuperación de información). Por un lado, en Internet y en las bibliotecas digitales se contiene una cantidad enorme de conocimiento que puede dar respuestas a muchísimas preguntas que tenemos. Entonces, la tarea se entiende cómo medir el grado de importancia para proporcionar al usuario primero el documento más relevante, si no le sirvió, el segundo más relevante. Este problema se complica ya que no existe un lenguaje formal en el cual el usuario podría

formular claramente su necesidad. La dirección más prometedora de resolver este problema es, nuevamente, el uso de lenguaje natural. Las técnicas más usadas actualmente para la recuperación de información involucran la búsqueda por palabras clave, se buscan los archivos que contengan las palabras que el usuario teclee. Las ideas más usadas son los modelos probabilísticos y los procedimientos iterativos e interactivos: tratar de adivinar qué necesita el usuario preguntándolo cuáles documentos le sirven. Coincidencia de las formas morfológicas de palabras, buscando pensar, encontrar piénsalo. Este problema es bastante simple a resolver en el lenguaje inglés, al cual se dedica la mayor parte de investigación en el mundo. Los métodos de la morfología computacional, la rama del procesamiento de lenguaje natural que se encarga del modelado de las formas morfológicas de palabras varían desde el uso de diccionarios que especifican las formas para cada palabra, hasta las heurísticas que ayudan a adivinarlas. Coincidencia de los sinónimos, conceptos más generales y más específicos: buscando cerdo, encontrar puerco, mascota, animal, etc. Este problema prácticamente no depende de lenguaje, aunque los diccionarios que se usan sí son específicos para cada lenguaje. Uno de los problemas que aún no recibieron una solución adecuada es medir las distancias en este árbol ¿qué tan parecida es la palabra cerdo a puerco? y a mascota, animal, objeto. Una generalización de esta idea son los diccionarios de las palabras conceptualmente relacionadas, por ejemplo, cerdo y tocino; sacerdote, Biblia, iglesia y rezar. Tomar en cuenta las relaciones entre las palabras en la petición del usuario y en el documento, buscando estudio de planes, rechazar como no relevante planes de estudio (AMPLN, 2009).

2.2.3. LA PILA DEL PROCESADO DEL LENGUAJE NATURAL

El procesamiento del lenguaje natural es el campo que combina las tecnologías de la ciencia computacional 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 a través de la siguiente pila de tareas.

2.2.3.1. TOKENIZACIÓN

La tokenización es el proceso de marcación y posiblemente clasificación de secciones de una cadena de caracteres de entrada. Las fichas resultantes se pasan entonces a alguna otra forma de procesamiento. La tokenización es frecuentemente definida por expresiones regulares, que son entendidas por un generador de análisis léxico. El proceso de tokenización conlleva la búsqueda de los espacios en blanco y otra simbología para dividir la frase en un arreglo en el texto y dividir partiendo de la anterior premisa a dividir la cadena en palabras, siendo este el algoritmo más utilizado. Además de ello tenemos al algoritmo *TreeBank* el cual funciona correctamente en la identificación de negaciones, pero falla al procesar datos desde una *URL* y *HTML* (Anjaria y Reddy Guddeti, 2014).

2.2.3.2. *STOPWORDS*

Los *Stopwords* son palabras como conjunciones, artículos, dirección web, palabras repetidas, caracteres repetidos nombre de usuarios que no muestran relevancia alguna para el procesamiento de dato (Anjaria, Reddy Guddeti, 2014). Las palabras extremadamente comunes parecen ser de poco valor para ayudar a la selección de documentos que coincidan con la investigación. Estas palabras se llaman *stopwords*. La estrategia general para determinar una lista de detención es clasificar los términos por frecuencia de recolección (el número total de veces que cada término aparece en la colección de documentos o textos) y luego tomar los términos más frecuentes, a menudo filtrado a mano para su contenido semántico relativo a el dominio de los documentos que se indexan, cuyos miembros se descartan durante la indexación (Manning, Raghavan, & Schütze, 2008).

La ley empírica de Zipf establece que la aparición de las palabras está relacionada con la frecuencia de las mismas. La aparición por ejemplo de la primera palabra más frecuente es el doble de la siguiente más frecuente y tres veces más que la siguiente (Moreno Sanchez, Font-Clos, Francesc, & Corral, 2016). Al analizar grandes cantidades de datos podemos observar que en el idioma ingles la palabra más frecuente es la THE seguida de OF, después viene AND y así van apareciendo más palabras (Harris, n.d.).

2.2.3.3. *STEAMMING*

El *steamming* es el método por el cual se eliminan las variantes de la palabra, el proceso consiste en limpiar de sufijos, inflexiones para tratar de llevar las palabras a sus respectivas raíces gramaticales y conseguir menor cantidad de datos para el procesamiento. Hay varios algoritmos de derivación que se han desarrollado para asegurar que las palabras se reducen a sus formas de raíz, reduciendo así el tamaño del diccionario de documentos. Esto se debe a que una raíz puede usarse para representar muchas variantes de términos usados en un lenguaje particular. Si bien este enfoque ayuda a recuperar documentos más relevantes, existe la posibilidad de subvertir donde dos palabras que pertenecen al mismo grupo conceptual se convierten en dos raíces o raíces diferentes, por ejemplo, una búsqueda de la palabra correr genera las palabras corriendo o corrido. Otro caso se presenta con dos palabras pertenecientes a grupos conceptuales diferentes se convierten en los mismos troncos o raíces, por ejemplo, cuando una búsqueda de la palabra juguete incluye un resultado de búsqueda que contiene La palabra jugue a lo que se puede denominar como *over-steeming* (Balakrishnan & Lloyd-Yemoh, 2014).

2.2.3.4. LEMATIZACIÓN

La lematización es el proceso mediante el cual las palabras de un texto que pertenecen a un mismo paradigma flexivo o derivativo son llevadas a una forma normal que representa a toda la clase. En este caso, consiste en encontrar el lexema de las palabras analizadas. Por otro lado utiliza el vocabulario y el análisis morfológico de la palabra e intenta eliminar las inflexiones, volviendo así las palabras a su forma de diccionario (Bassi A., n.d.). Comprueba para asegurarse de que las cosas se hacen correctamente analizando si las palabras de consulta se utilizan como verbos o sustantivos. También ayuda a hacer coincidir los sinónimos mediante el uso de un diccionario de sinónimos de modo que cuando uno busca "caliente" la palabra "cálido". En el mismo sentido una búsqueda para el "coche" producirá "coches" tan bien como "automóvil". La técnica se ha utilizado en varios idiomas para la recuperación de la información (Balakrishnan & Lloyd-Yemoh, 2014).

2.3. ANÁLISIS DE SENTIMIENTO

El análisis de sentimiento es un estudio computacional que ayuda a identificar emociones, actitudes en opiniones, permitiendo saber cómo son percibidos los productos y sus características (Anjaria, Reddy Guddeti, 2014).

El análisis de sentimiento también llamado análisis de opinión, es aquel campo que estudia los sentimientos, opiniones, actitudes, emociones, hacia cierto tipo de entidades como productos, servicios, organizaciones, temas personas y todo lo que los caracteriza.

Existen muchos términos como minería de opiniones, extracción de opinión, análisis subjetivo, análisis de afecto, análisis de emoción, minería de revisión, que denotan diferentes tareas pero forman el análisis de sentimiento. Ciertamente el análisis de sentimiento lleva una estrecha relación con NLP (*Natural Language Processing*), además el análisis de sentimiento tiene un profundo impacto en las ciencias sociales, economía, política, aunque esta recién habrá visto la vida en 2000, hubieron algunos trabajos previos sobre las estructuras gramaticales como el sentimiento de los adjetivos, los puntos de vista y la interpretación de las metáforas (Liu, 2012).

Las opiniones soy muy importantes para el ser humano, son puntos claves en nuestro comportamiento. Cada vez que se presenta la oportunidad de tomar una decisión, se requiere de saber la opinión de personas cercanas a uno. Así pues los posibles compradores de un servicio quieren saber la opinión de otros usuarios, antes de comprar algún producto, como también ocurre con las elecciones, organizaciones, colegios. Instituciones del gobierno, cada vez que se pueda tomar una decisión la opinión de los más cercanos o de personas que ya habrán consumido algún servicio, es muy valiosa para el usuario. Con la crecida abrumadora de los medios sociales digitales como ser revistas digitales, foros, *blogs*, redes sociales, las compañías están usando esta información generada para campañas electorales, marketing. Para una organización ya no será necesario realizar encuestas los usuarios debido a que ahora contamos con mucha información públicamente disponible (Liu, 2012).

No obstante, monitorear toda esta información, clasificarla y procesarla hasta lograr obtener la información requerida es una tarea ya por si tediosa. Cada uno de estos sitios (Páginas web) cuentan con una cantidad grande de datos que en muchas ocasiones es difícil de discriminar e identificar en ellos la información relevante. La persona promedio tendrá muchas dificultades en poder extraer toda la información de páginas, resumirlas y categorizarlas. Por esta razón es que los sistemas de recolección y procesamiento son necesarias. En los últimos años se ha observado que las opiniones en las redes sociales han reformado negocios, además de haber movilizado grandes cantidades de personas en el medio oriente. Llega a ser por consecuente una necesidad colectar y estudiar las opiniones en la *Web* (Liu, 2012).

Es cierto que la información no solo está en la web, también existen bancos de datos generados por empresas. Por estas razones es que el análisis de sentimiento se ha expandido mucho más allá de los productos, servicios, salud, servicios de finanzas, eventos sociales y la política (Liu, 2012).

2.3.1. EL CLASIFICADOR NAIVE BAYES

Es un método basado en la teoría de probabilidad, usa frecuencias para calcular probabilidades condicionales para calcular predicciones sobre nuevos casos. Es una técnica tanto predictiva como descriptiva. Sean E y F eventos. Podemos expresar a E como:

$$E = EF \cup EF^C$$

Es decir que para que ocurra un evento E, deben suceder E y F o bien debe suceder E y no F. Debido a que EF y EF^c son mutuamente excluyentes, tenemos que:

$$P(E) = P(EF) + P(EF^C)$$

$$= P(E\,|F\,)P(F) + P(E|F^C)P(F^C)$$

$$= P(E\,|F)P(F) + P(E\,|F^C)\big(1 - P(F)\big) \quad (2)$$

La ecuación (2) establece que la probabilidad del evento E es una ponderación de la probabilidad condicional de E dado que F ha ocurrido y la probabilidad condicional del evento E dado que F no ha ocurrido. Cada probabilidad condicional proporciona tanta ponderación como el evento condicionado tiende a ocurrir.

La ecuación (2) puede generalizarse de la siguiente manera: supongamos que los eventos $F_1, F_{2\,,\ldots}\, F_n$ son mutuamente excluyentes tal que $\cup_i^n F_i = S$, donde S es el espacion muestral. En otras palabras, exactamente uno de los eventos ocurrirá. Podemos escribir lo anterior como:

$$E = \bigcup_{i=1}^{n} E_i$$

De la definición de probabilidad condicional tenemos que

$$P(EF_i) = P(\,E\,|F_i)\,P(F_i) \quad (3)$$

Además, usando el hecho de que los eventos $EF_i, i = 1\,, \ldots.\,, n$ son mutuamente excluyentes, obtenemos que

$$P(E) = \sum_{i=1}^{p} P(EF_i)$$

$$= \sum_{i=1}^{n} P(E\,|F_i)P(F_i) \quad (4)$$

Así, la ecuación (4) muestra cómo, para eventos dados $F_1, F_{2\,,\ldots}\, F_n$ de los cuales uno y solamente uno puede ocurrir, se puede calcular P(E) condicionado a que ocurra F_1. Esto es, se establece que P(E) es igual al promedio de las ponderaciones de $P(E|F_i)$ y cada termino es ponderado por la probabilidad del evento en el cual es condicionado. Supóngase ahora que E ha ocurrido y que se quiere determinar la probabilidad de que el evento $F_j$haya ocurrido.

Por la ecuación (4) tenemos que:

$$P(F_j|E) = \frac{P(EF_j)}{P(E)}$$

$$= \frac{P(E\,|F_j)P(F_j)}{\sum_{i=1}^{n} P(E\,|F_i)P(F_i)} \quad (5)$$

La ecuación (5) es conocida como la fórmula de Bayes (Pacheco Leal, Días Ortiz, & Rodolfo, 2005).

2.4. REDES SOCIALES

Los servicios de *microblogging* como *Facebook, Twitter, Google+* generan diariamente una cantidad exorbitante de comentarios. Los usuarios de internet han cambiado los *blogs* tradicionales, *chats* por estos nuevos servicios de redes sociales que permiten expresar opiniones sobre variados temas, comentarios. La mayoría de estos sitios desde su creación han capturado millones de usuarios.

Dada la novedad del fenómeno y su popularidad, muchas empresas han comenzado a utilizar las redes sociales digitales como una herramienta de marketing, algunas incluso sin ningún tipo de estrategia. Aún no está claro si el marketing con redes sociales digitales puede ayudar a las empresas a obtener mejores resultados, por lo que el objetivo de este artículo es contribuir con el creciente pero aún limitado estudio sobre el uso de las redes sociales digitales como herramienta de marketing y su impacto sobre el desempeño (Uribe Saavedra, Rialp Criado, Llonch Andreu, 2013).

Las redes sociales se han convertido en un fenómeno pero solo algunas han alcanzado un nivel grande de afiliados, muchas de las empresas viendo el crecimiento y el bajo coste de uso están empezando a utilizarlas para la construcción de marca (Uribe Saavedra, Rialp Criado, Llonch Andreu, 2013), para mejorar sus productos y atender a su clientela las 24 horas.

2.4.1. DEFINICIÓN

Definimos los sitios de redes sociales como servicios basados en las web que permiten a los individuos construir a los individuos un perfil público o semi-publico dentro un sistema, articular una lista de usuarios con quien comparten conexión, ver y traspasar sus listas de conexiones. La naturaleza y nomenclatura de estas conexiones puede variar de sitios a sitio. Lo que hace único a las redes sociales no es que permitan conocer individuos extraños, sino que permite a los usuarios articular y hacer visible sus redes sociales. Esto puede llevar a hacer conexiones entre individuos que de otra forma no lo podrían haber hecho, pero esto muchas veces no es el objetivo y estos encuentros se realizan entre personas afines.

2.5. MARKETING DIGITAL

La red social Facebook, como medio de comunicación y venta. Desde la creación de la empresa en 2006, se ha convertido en el medio de interrelación entre sus usuarios, representando además un medio de marketing que las empresas usan actualmente. Los usuarios en estos nuevos medios de comunicación no solo son usuarios pasivos sino que participan, opinan, son más activos, crean contenidos. Estos consumidores 2.0 son catalogados como *prosumer* término acuñado de las palabras productor y consumidor que hace elocuencia de que el productor, también pasa a ser consumidor y el consumidor pasa a ser productor con la retroalimentación que brinda a través de las redes sociales como Facebook (Miranda Gonzaléz, Rubio Lacoba, Chamorro Mera, & Correia Loureiro, 2013).

Facebook y el resto de las redes sociales, pueden influir en cuatro etapas el proceso de decisión de compra que siguen al reconocimiento de la necesidad: búsqueda de la información, valoración de las alternativas, compra y evaluación de la decisión. Así, los habituales canales de búsqueda de información se complementan ahora con la información que las empresas colocan de sus productos en sus páginas de Facebook, un medio accesible a cualquier comprador potencial las 24 horas del día, todos los días del año, a través de cualquier dispositivo con conexión a Internet: ordenador, *Tablet*, móvil o *Smart TV*. Por su parte, la valoración de las alternativas de compra se ve favorecida a través de los comentarios

que otros usuarios han hecho previamente en la página de *Facebook* de la marca o en el intercambio de opiniones que el consumidor puede tener en su Facebook con sus "amigos" (Miranda Gonzaléz, Rubio Lacoba, Chamorro Mera, & Correia Loureiro, 2013).

Las empresas pueden crear en Facebook una página que les permita vender y cerrar una transacción comercial sin necesidad de remitir a la página web de la empresa o a una tienda física. En este sentido entendemos el *f-commerce* como el uso de Facebook como plataforma para potenciar e incluso materializar cualquier tipo de transacción electrónica, por lo que el *f-commerce* es un complemento del *e-commerce*. El inicio del *f-commerce* puede fijarse con la primera compra que se realizó en la tienda en Facebook, 1-800-Flowers, el 8 de julio de 2009. Facebook juega un papel importante en la evaluación post-compra, puesto que facilita la difusión de las experiencias de compra, tanto las positivas como las negativas. (Miranda Gonzaléz, Rubio Lacoba, Chamorro Mera, & Correia Loureiro, 2013).

2.6. *WEB SCRAPING*

Como el internet evoluciono, la tarea de extraer información de las páginas se hacía más ardua y complicada. Los formatos formaron parte de la solución para entonces como XML que conjunto con convenciones llegaron a formar XHTML, un conjunto de etiquetas que se abren y cierran, bajo una estructura de árbol, los mismos que pasaron a ser bien formado *XHTML*. El intercambio de información por internet llego a ser más fluido pero los datos compartidos eran cada vez más complicados de extraer, debido a que venía acompañada por anuncios, videos, imágenes y otros recursos multimedia. (Russell, 2013)

2.6.1. EXTRACCIÓN

La extracción de datos a través del uso de lenguajes de programación o *web scraping* puede observarse como una tarea no muy complicada. La verdad recae en el contenido de los datos extraídos, como ser documentos con formato XML que son proporcionados como fuente de datos desde un servidor, HTML que es el lenguaje de marcado de hipertexto con el cual compartimos contenidos a través del internet, así como los RSS o recursos que proveen

formatos para compartir información. El texto contenido en estos formatos está acompañado por anuncios, cabeceras, *banners*, imágenes, videos y otros recursos en los que no se está interesado. Procediendo a la eliminación de las etiquetas de *HTML* pareciese que el problema quedara resuelto, nada más alejado de la realidad, ya que existen muchos más datos que deben ser descartadas antes de seleccionar la información en texto requerida por el usuario. (Russell, 2013)

2.7. JAVASCRIPT

Javascript es un lenguaje de programación liviano, interpretado, con funciones de primera clase. Mientras es bien conocido como el lenguaje de *scripting* para las páginas web, muchos ambientes que no son navegadores también lo usan, tales como node.js y Apache CouchDB. Javascript esta basado en prototipos, multiparadigma, lenguaje dinámico de *scripting*, suporte de orientación a objetos, imperativo y declarativo (jsx, 2015).

2.8. PYTHON

Python es simple y aun así un lenguaje de programación poderoso con excelente funcionalidad para el procesado de datos lingüísticos. Se elige Python porque tiene una curva de aprendizaje poco profunda, su sintaxis y semántica son transparentes y tiene un buen manejo de cadenas. Como un lenguaje interpretado facilita la exploración interactiva. Como un lenguaje orientado a objetos permite que los datos y métodos sean encapsulados para su reusó fácilmente. Como lenguaje dinámico, facilita un desarrollo rápido.

2.8.1. NLTK

Originalmente creada en 2001 como parte del curso de lingüística computacional del departamento de ciencia de la información y computación de la universidad de Pensilvania. Desde aquel entonces ha sido desarrollado y extendido con la ayuda de docenas de contribuidores. Ha sido adoptado en muchos cursos en universidades y servido de base de muchos proyectos de investigación (Bird, Klein, & Loper, 2009).

NLTK ha sido diseñado con cuatro objetivos primarios:

2.8.1.1. SIMPLICIDAD

Proveer un *framework* intuitivo conjunto con bloques de construcción substanciales, dando a los usuarios un conocimiento práctico de NLP sin la tediosa tarea asociada con el procesado del lenguaje anotado.

2.8.1.2. CONSISTENCIA

Proveer un framework uniforme con interfaces consistentes y estructura de datos y nombres de métodos fáciles de adivinar.

2.8.1.3 EXTENSIBILIDAD

Proveer una estructura en la que los nuevos módulos puedan ser fácilmente acomodados, incluyendo implementaciones alternativas.

2.8.1.4 MODULARIDAD

Proveer componentes que pueden ser usados independientemente sin necesidad de entender el resto del paquete de herramientas.

2.9. EL PROCESO KDD

Encontrar patrones útiles en los datos conocido con diferentes nombres incluyendo la minería de datos, extracción de conocimiento, descubrimiento de información, recolección de información, arqueología de los datos y procesado de patrones. El termino minería de datos es más usado por los estadísticos. El termino KDD se usa para referirse al proceso completo de descubrimiento de información útil en los datos. El minado de datos es un paso más en este proceso de descubrimiento de información útil. Los pasos adicionales en el proceso KDD como la preparación de datos, la selección de datos, la limpieza de los datos, la incorporación apropiada de datos y la interpretación de los resultados de la minería de los datos son pasos requeridos para el descubrimiento de información útil. KDD ha evolucionado y continua

evolucionando de las investigaciones en los campos de las bases de datos, las máquinas de aprendizaje, reconocimiento de patrones, la estadística, la inteligencia artificial, adquisición de conocimiento por sistemas expertos, la visualización de los datos, recopilación de datos y computación de alto rendimiento. Los sistemas de software de KDD incorporan teoría, algoritmos y métodos de todos estos campos.

En la gran variedad de campos de estudio, los datos son colectados y acumulados a un paso dramático. Existe una necesidad por una nueva generación de teorías y herramientas para asistir a las personas al extraer información útil o conociendo de los volúmenes de rápido crecimiento de datos (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

2.9.1. ¿POR QUÉ NECESITAMOS KDD?

Los métodos tradicionales de transformar datos en conocimiento recaen en el análisis manual y su interpretación. Por ejemplo en la industria de la salud, es común para los especialistas analizar periódicamente las tendencias y cambios en los datos de la salud de sus pacientes y los colocan en los reportes, los mismos llegan a ser la base para las futuras decisiones y para los planes del manejo de la salud del paciente. Así como en las ciencias, *marketing*, finanzas, salud o cualquier otro campo el objetivo clásico de análisis de datos recae fundamentalmente en uno o más analistas llegando a ser íntimamente familiares con los datos y sirviendo como una interfaz entre los datos, los usuarios y los productos. Debido a este tipo de aplicaciones, probar manualmente el conjunto de datos llega a ser bastante un proceso lento. De hecho, mientras los datos vayan creciendo, el análisis manual llega a ser impráctico (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

2.9.2. ETAPAS DEL PROCESO KDD

El proceso KDD es interactivo e iterativo, involucrando numerosos pasos con muchas decisiones hechas por el usuario. Las etapas del proceso KDD muestran como desarrollar minería de datos en fases que permitan identificar y procesar de mejor manera el contenido a investigar, el mismo consta de las siguientes fases:

2.9.2.1. SELECCIÓN

Esta etapa consiste en crear un conjunto de datos o enfocarse en un subconjunto de variables o ejemplos de datos en los cuales se descubrirá el conocimiento. Además de desarrollar un entendimiento del dominio de la aplicación e identificar el objetivo para aplicar el proceso KDD. En este paso distintas fuentes de datos pueden ser combinados.

Luego se procede a la creación del conjunto de datos, selección de datos o enfoque hacia un conjunto de variables en los que se buscara el conocimiento.

2.9.2.2 PRE PROCESAMIENTO

Esta etapa consiste en el limpiado de los datos obtenidos y el pre-procesamiento para obtener información consistente a nuestra investigación. El limpiado de datos y pre-procesamiento incluye eliminar el ruido si es necesario y estrategias para el manejo de los campos de datos vacíos.

2.9.2.3. TRANSFORMACIÓN

Este paso consiste en la transformación de los datos usando reducción de dimensionalidad o métodos de transformación. Los datos son transformados en formas apropiadas para la minería de datos y/o se seleccionan los atributos más útiles capaces de representar los datos dependiendo de las metas propuestas.

2.9.2.4. DATA MINING

Esta etapa consiste en la búsqueda de patrones de interés, dependiendo del objetivo de la minería de datos en una forma particular de representación, por ejemplo clasificación o *clustering*.

2.9.2.5. INTERPRETACIÓN Y EVALUACIÓN

Esta etapa consiste en la interpretación y evaluación de los patrones encontrados. Se identifican los patrones realmente interesantes basados en alguna medida de interés.

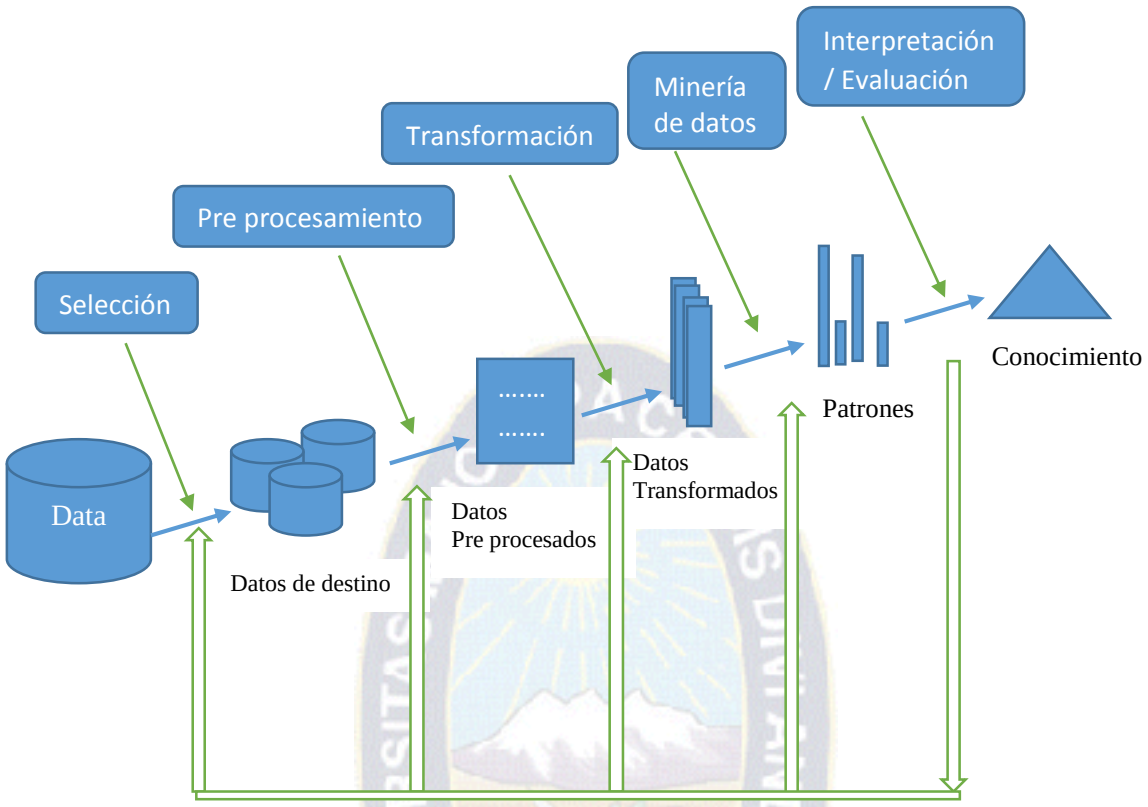


Figura 2.1: Los pasos que componen el proceso KDD

Fuente: (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

CAPITULO III

DESARROLLO DE HERRAMIENTAS DE MINADO DE TEXTO

Se desarrolló las bases para el minado de textos en la red social de Facebook basado en el proceso KDD descrito por Fayyad (Fayyad, Piatetsky-Shapiro y Smyth, 1996) con los parámetros descritos en la figura 3.1.

Figura 3.1: Proceso KDD con tareas específicas

Fuente: [Proceso KDD]

Se procedió según los pasos especificados en la figura 3.1 del proceso KDD. Las siguientes herramientas se desarrollaron en el lenguaje interpretado de programación *Python* exceptuando el colector de datos que fue desarrollado en Javascript.

3.1. SELECCIÓN

Para la selección de perfiles se procedió con el uso de código Javascript inyectado a la página del perfil de las personas figura 3.1 para la colección automática de las publicaciones y comentarios hechos en su línea de tiempo. El código inyectado tiene un parámetro x que especifica cuantas veces la página se deslizará hacia la parte baja de la página para cargar más publicaciones y así obtener más comentarios.

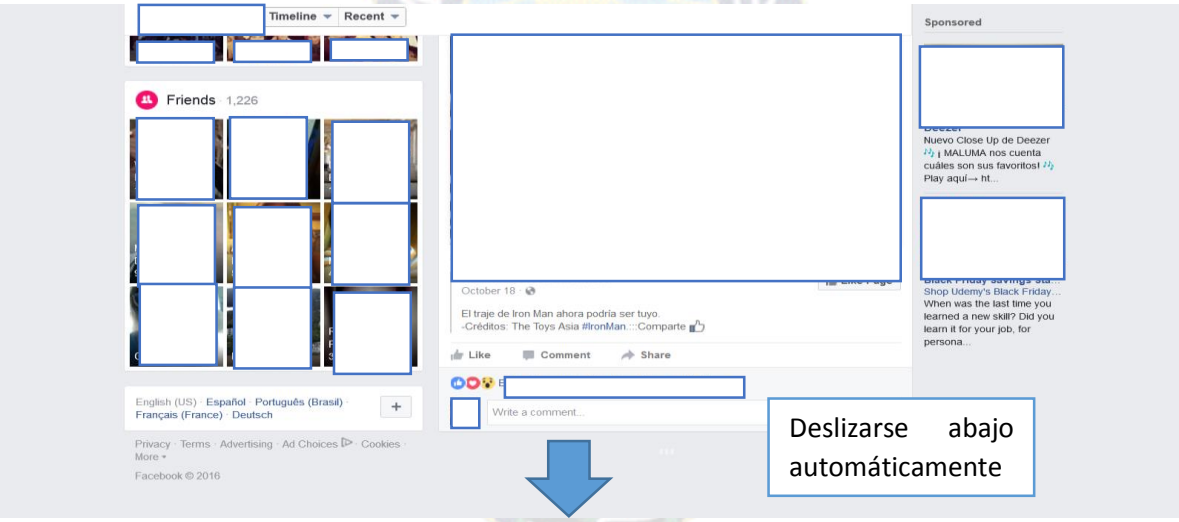


Figura 3.2: Deslizamiento automático de perfil

Fuente: [https://www.facebook.com/]

El script inyectado recolecta información del navegador presionando automáticamente las secciones de "ver más comentarios" o "ver más", al finalizar la cantidad establecida de recolección se genera un archivo csv separado por dos columnas como muestra la figura 3.3 la primera columna con el nombre del autor del post y la segunda columna con el comentario eliminando los saltos de línea al guardar el archivo con nombre del autor de perfil y guarda todos los comentarios realizados en el muro de la persona.

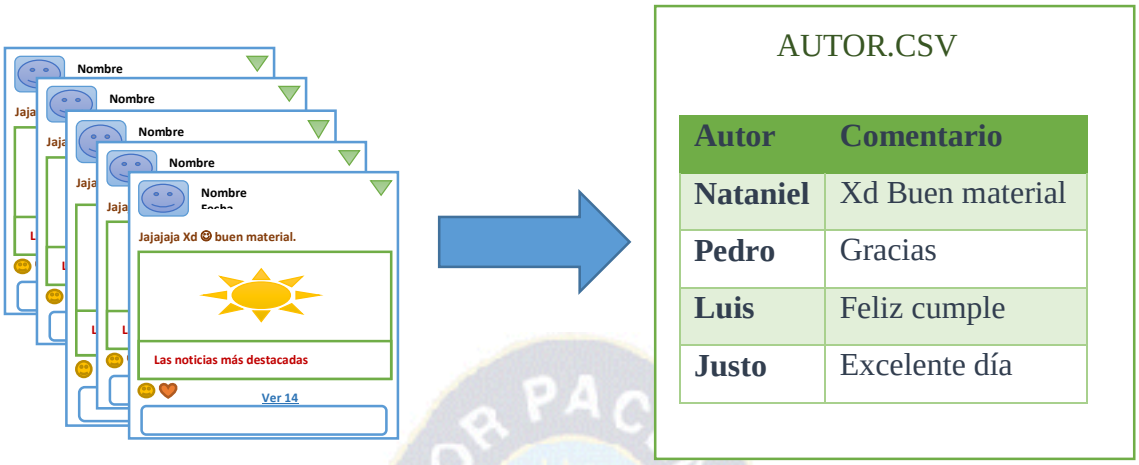


Figura 3.3: Proceso de colección de datos con Javascript a archivo CSV

Fuente: [elaboración propia]

El script recolector de páginas de Facebook acepta como entrada el identificador de la página, el *token* de acceso de Facebook que se asigna cuando se crea una aplicación en https://developers.facebook.com/ y a través de peticiones por *GET* hacia el API de grafos y con la debida autorización del *token* de acceso se pasa a procesar la información y al final se guarda los resultados en un archivo CSV con el id de estado, el nombre de enlace, el contenido del estado, el enlace permanente, como muestra la figura 3.4.

Figura 3.4. Recolección de publicaciones con *Python* de una página

Fuente: [elaboración propia]

El script recolector de datos de grupos de Facebook recoge las publicaciones del grupo público con identificador como dato de entrada, y *token* de acceso para hacer una petición al API de grafos de Facebook previamente autenticado por el *token* de acceso, luego se procesa todas las publicaciones que se realizaron en el grupo y se pasa a generar un archivo CSV con el id de estado, el contenido del estado (texto del post), el nombre de enlace, el enlace permanente, como muestra la figura 3.5.

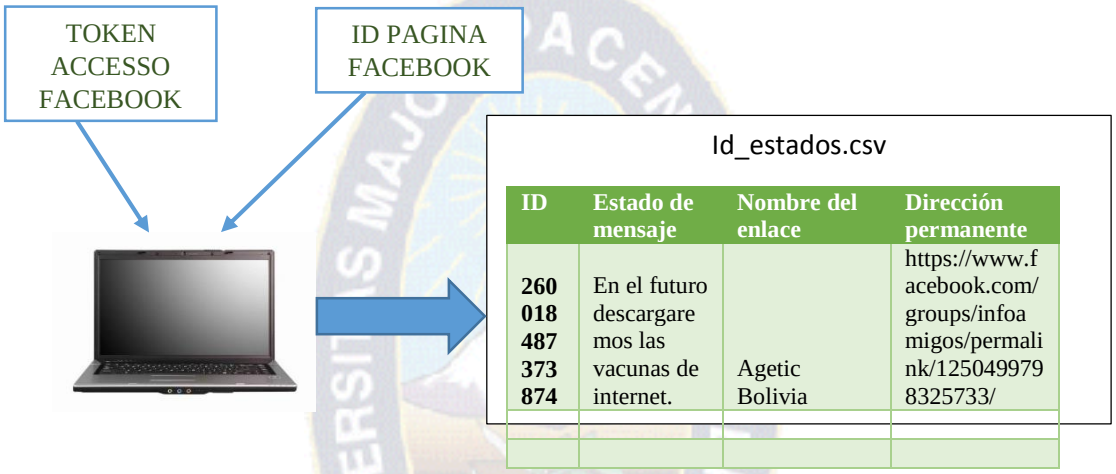


Figura 3.5: Proceso de recolección de las publicaciones de grupo público

Fuente: [Elaboración propia]

La recolección de comentarios de los post de las páginas y de los grupos públicos se los realiza con el script recolector de comentarios, el que requiere del identificador de la página o del grupo y el *token* de acceso del API de *Facebook*, una vez con esta información se busca el archivo generado con los dos programas anteriores con el nombre introducido, toma el contenido de los archivos generados y gracias a la dirección web permanente de la publicación contenido en los archivos de estado, se procede a buscar los comentarios de cada uno de las publicaciones y guardarlos en un archivo nuevo con el mismo identificador seguido de la palabra comentario, con todo el contenido de los mismos descrito en los siguientes campos: id de comentario, contenido de la publicación, autor del comentario, fecha de publicación, como muestra la figura 3.6.

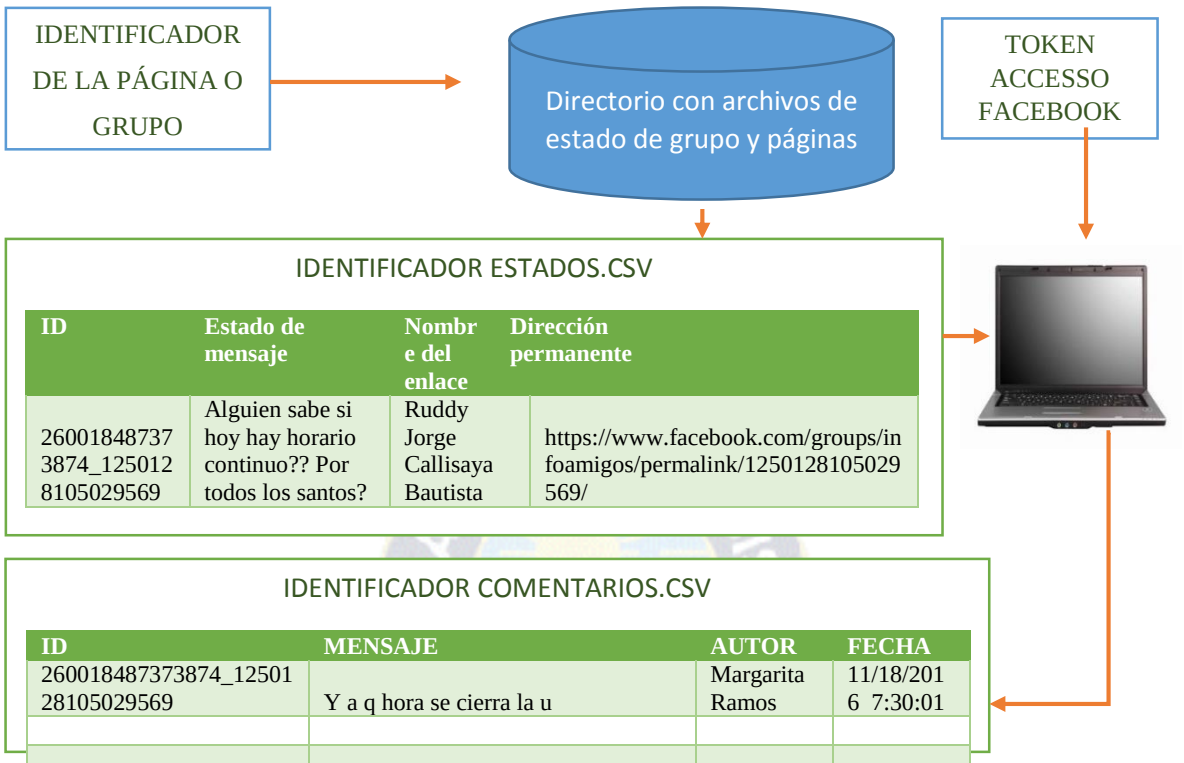


Figura 3.6: Proceso de colección de comentarios de páginas y grupos

Fuente: [Elaboración propia]

3.2. PRE-PROCESAMIENTO

Se desarrolló herramientas para uniformar el contenido de los textos, para tener datos consistentes para su posterior análisis, el pre procesamiento se realiza como tarea previa a la transformación para evitar el ruido del texto adquirido y se toma estrategias para el manejo de información el mismo fue realizado en los pasos siguientes:

3.2.1. DIVISIÓN DE PERFILES

Se procedió a tomar todos los archivos csv generados en la figura 3.3 uno a la vez, del archivo seleccionado se toma el primer comentario y su autor y se crea el archivo con el nombre del autor y se adiciona al final del archivo el comentario, lo mismo ocurre con el segundo comentario y en adelante. Así se obtiene archivos personales únicos de comentarios realizados por la persona que lleva el nombre del archivo como muestra la figura 3.7.

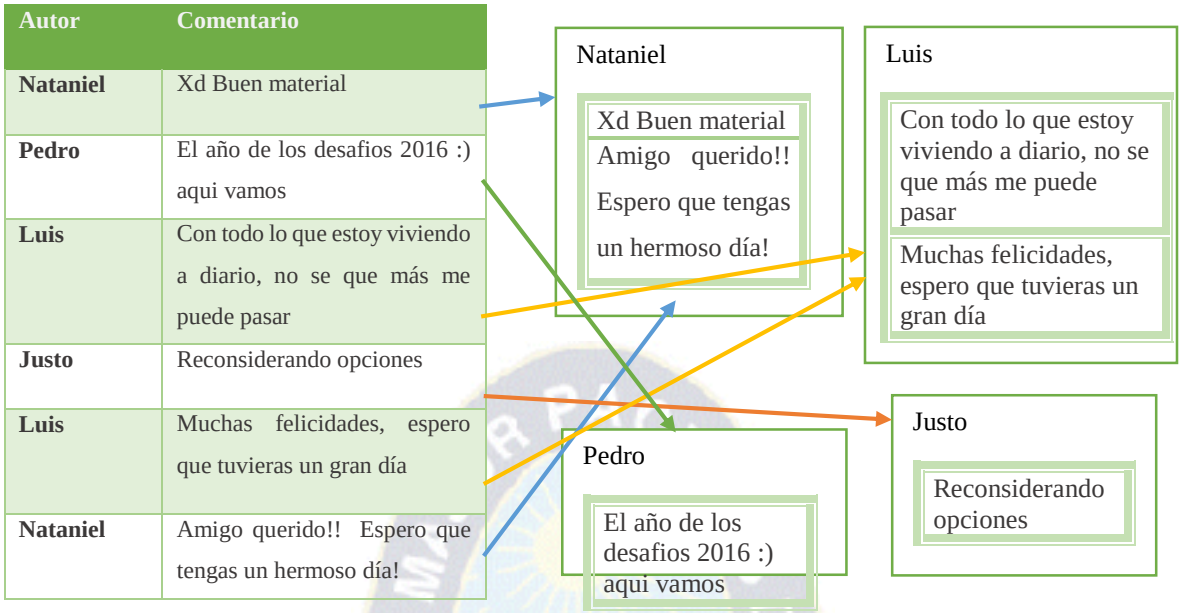


Figura 3.7: División de perfiles

Fuente: [Elaboración propia]

3.2.2. UNIÓN DE ARCHIVOS DE ESTADOS Y COMENTARIOS

Se procedió a la unión de los archivos de estado de las páginas y grupos de Facebook con los archivos de comentarios de las mismas, el resultado se combina en un archivo único, extrayendo solo el id de autor y el comentario, el texto resultante tendrá el nombre del identificador de la página o grupo público, de esta manera podremos tratarlo como uno solo en la minería de textos en los siguientes pasos como se muestra en la figura 3.8.

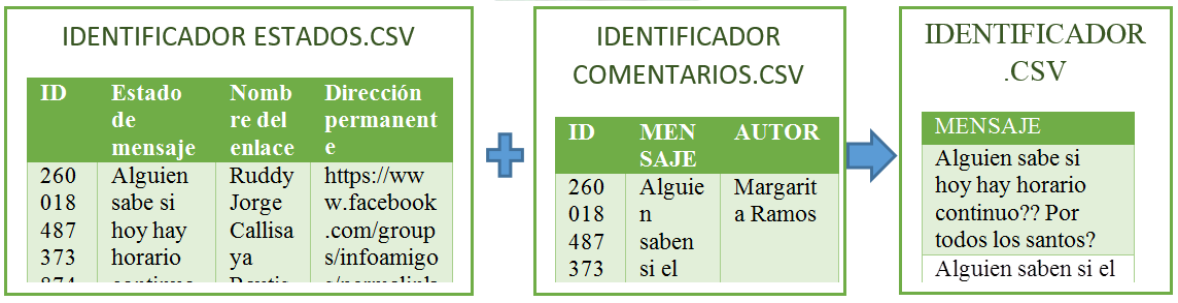


Figura 3.8: Unión de archivos recolectados de grupos y páginas de Facebook

Fuente: [Elaboración propia]

3.2.3. ELIMINACIÓN DE SIGNOS DE PUNTUACIÓN

Para un mejor procesamiento de los textos se procede a llevar todo el contenido de cada documento a minúsculas. La eliminación de signos de puntuación se realizó con la expresión regular [^\w\s] que identifica todos los caracteres de puntuación como muestra la figura 3.9 y luego se procede a reemplazarlos con una cadena en blanco para el análisis en el siguiente paso.

. , ¿ ? [ ] { } ( ) ¡ ! ` ~ + * / , ^ % $ # @ = ; < >

Figura 3.9: Signos de puntuación a eliminar

Fuente: [elaboración propia]

3.2.4. ELIMINACIÓN DE ACENTOS

Se creó un método que recibe como entrada un texto y reemplaza toda vocal con acento por la vocal semejante sin acento con las equivalencias que muestra la tabla 3.1 además se reemplaza las comas y saltos de línea por el espacio en blanco para el mejor tratamiento de la información en los siguientes pasos.

Tabla 3.1: Equivalencias de acentos a sus representaciones sin acento

Fuente: [Elaboración propia]

Representación UNICODE	Representación UTF-8	Carácter a reemplazar
,	,	“ ”
\n	Salto de línea	“ ”
\xe3\xb1	ñ	n
\xe3\xad	í	i
\xe3\xa1	á	a
\xe3\xa9	é	e
\xe3\xb3	ó	o
\xe3\xba	ú	u

3.2.5. ELIMINACIÓN DE DIRECCIONES WEB

Los comentarios realizados en la red social de Facebook llevan en su contenido enlaces a páginas web, vídeos, foros, blogs y otros recursos, que para el presente estudio no presentan relevancia, debido a esto se creó el método de eliminación de enlaces que limpia de un texto todo el contenido de enlaces a recursos web. Para la identificación de direcciones web se hace uso de la expresión regular que se muestra en la figura 3.10 se pueden identificar y después eliminarlos.

```
(https?:\/\/(?:www\.|(?!www))[^\s\.]+\.[^\s]{2,}|www\.[^\s]+\.[^\s]{2,})
```

Figura 3.10: Expresión regular que identifica todos las direcciones web

Fuente: [Elaboración propia]

3.3. TRANSFORMACIÓN

Se procede a la transformación de los textos obtenidos en el proceso de selección de perfiles. Mediante la pila de procesamiento del lenguaje natural se procedió a la implementación de los siguientes pasos:

3.3.1. TOKENIZACIÓN

Se introduce un comentario y se divide su contenido en palabras con el uso del método *Word_tokenize* de la librería NLTK de *Python*, que devuelve el contenido en un vector unidimensional para poder iterar sobre el mismo.

3.3.2. *STOPWORDS*

La librería NLTK de *Python* tiene un módulo llamado *stopword* con múltiples idiomas, que lleva las palabras que menos significado aportan al análisis de textos, como ser los artículos, determinantes, conjunciones, en este estudio usaremos los *stopwords* para el idioma español, que observaremos en la figura 3.11.

de , la , que , el , en , y , a , los , del , se , las , por , un , para , con , no , una , su , al , lo , como , más , pero , sus , le , ya , o , este , sí , porque , esta , entre , cuando , muy , sin , sobre , también , me , hasta , hay , donde , quien , desde , todo , nos , durante , todos , uno , les , ni , contra , otros , ese , eso , ante , ellos , e , esto , mí , antes , algunos , qué , unos , yo , otro , otras , otra , él , tanto , esa , estos , mucho , quienes , nada , muchos , cual , poco , ella , estar , estas , algunas , algo , nosotros , mi , mis , tú , te , ti , tu , tus , ellas , nosotras , vosostros , vosostras , os , mío , mía , míos , mías , tuyo , tuya , tuyos , tuyas , suyo , suya , suyos , suyas , nuestro , nuestra , nuestros , nuestras , vuestro , vuestra , vuestros , vuestras , esos , esas , estoy , estás , está , estamos , estáis , están , esté , estés , estemos , estéis , estén , estaré , estarás , estará , estaremos , estaréis , estarán , estaría , estarías , estaríamos , estaríais , estarían , estaba , estabas , estábamos , estabais , estaban , estuve , estuviste , estuvo , estuvimos , estuvisteis , estuvieron , estuviera , estuvieras , estuviéramos , estuvierais , estuvieran , estuviese , estuvieses , estuviésemos , estuvieseis , estuviesen , estando , estado , estada , estados , estadas , estad , he , has , ha , hemos , habéis , han , haya , hayas , hayamos , hayáis , hayan , habré , habrás , habrá , habremos , habréis , habrán , habría , habrías , habríamos , habríais , habrían , había , habías , habíamos , habíais , habían , hube , hubiste , hubo , hubimos , hubisteis , hubieron , hubiera , hubieras , hubiéramos , hubierais , hubieran , hubiese , hubieses , hubiésemos , hubieseis , hubiesen , habiendo , habido , habida , habidos , habidas , soy , eres , es , somos , sois , son , sea , seas , seamos , seáis , sean , seré , serás , será , seremos , seréis , serán , sería , serías , seríamos , seríais , serían , era , eras , éramos , erais , eran , fui , fuiste , fue , fuimos , fuisteis , fueron , fuera , fueras , fuéramos , fuerais , fueran , fuese , fueses , fuésemos , fueseis , fuesen , sintiendo , sentido , sentida , sentidos , sentidas , siente , sentid , tengo , tienes , tiene , tenemos , tenéis , tienen , tenga , tengas , tengamos , tengáis , tengan , tendré , tendrás , tendrá , tendremos , tendréis , tendrán , tendría , tendrías , tendríamos , tendríais , tendrían , tenía , tenías , teníamos , teníais , tenían , tuve , tuviste , tuvo , tuvimos , tuvisteis , tuvieron , tuviera , tuvieras , tuviéramos , tuvierais , tuvieran , tuviese , tuvieses , tuviésemos , tuvieseis , tuviesen , teniendo , tenido , tenida , tenidos , tenidas , tened

Figura 3.11: *Stopwords* del idioma español de la librería NLTK de PYTHON

Fuente: [Modulo *Stopwords* de la librería NLTK]

Además de las palabras incorporadas desde la librería NLTK se procede a crear una nueva lista de palabras más frecuentes que no presentan sentido para el análisis de texto en la red social de Facebook, entonces se visualiza las palabras en graficas de frecuencias como muestran las figuras 3.12 y 3.13 y a través de la observación se pueden identificar las que no representen valor semántico para análisis, procedimiento se realiza con todos los textos.

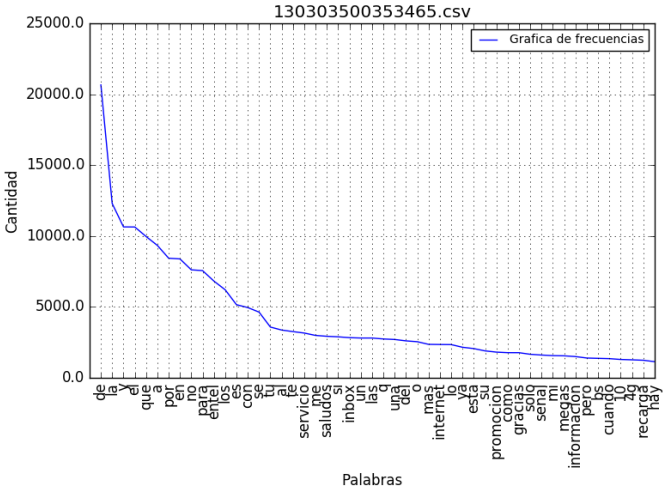


Figura 3.12: Selección de *stopwords* página 130303500353465

Fuente: [Elaboración propia]

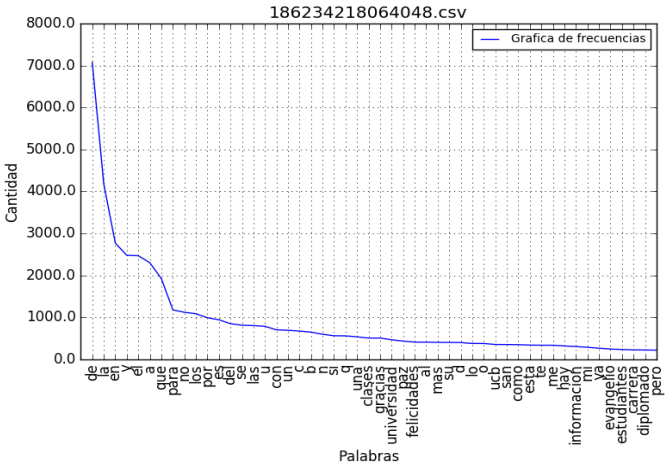


Figura 3.13: Selección de *stopwords* página 186234218064048

Fuente: [Elaboración propia]

Al Finalizar con la observación de todos los gráficos de frecuencia se identificó las palabras que muestra la figura 3.14 las cuales fueron adicionadas a nuestro archivo de *stopwords* para la red social de Facebook.

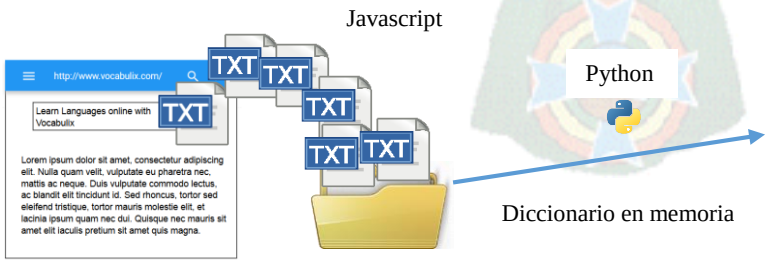
_ , all , and , are , aww , at , by , d , for , gg , ggg , gggg , ggggg , hahahaha , i , in , is , it , ja , jaja , jajaj , jajaja , jajaja , jajajaja , jajajajaja , jajajajajaja , je , jee , jee , jeee , jeeee , jejej , jejeje , jua , love , m , mmm , much , my , n , of , oh , oooooh , or , oww , p , q , si , sii , t , that , the , this , to , to , u , us , v , we , who , with , x , xd , you

Figura 3.14: *Stopwords* identificados mediante observación exhaustiva

Fuente: [Elaboración propia]

3.3.3. RAÍCES

En el proceso de minado de textos se requiere solo la representación del verbo en forma infinitiva, debido a esto es que se crea una carpeta con todas las conjugaciones de 2230 verbos, cada verbo con su respectivo archivo de conjugaciones correspondientes tomados de la página web http://www.vocabulix.com/ obtenidos gracias a un programa en Javascript que recolecta las conjugaciones de un verbo a la vez y que después pasan a construir un diccionario en memoria con Python como muestra la figura 3.15 para reemplazar todas las palabras de los archivos de los perfiles, grupos y páginas que contengan alguna ocurrencia de las mismas, como resultado tendremos todos los archivos con los verbos en infinitivo para su posterior procesamiento.



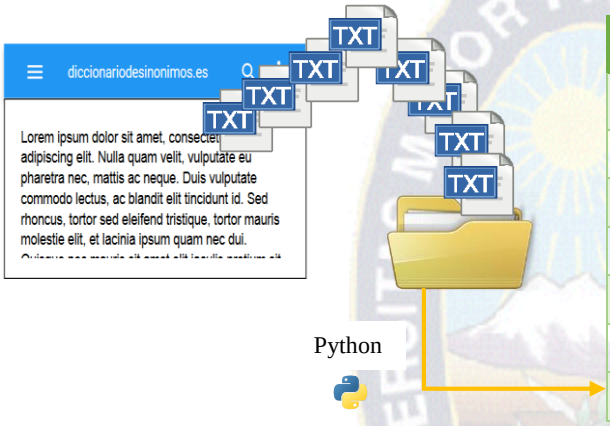
Otra forma	Verbo infinitivo
aborto	abortar
abunda	abundar
enfadaba	enfadar
enojaremos	enojar
espires	espirar
estafan	estafar
evites	evitar

Figura 3.15: Creación de diccionario de otras formas verbales a verbos en infinitivo

Fuente: [Elaboración propia]

3.3.4. SINÓNIMOS

Para la descripción del producto se procedió a la creación de un archivo de sinónimos del español para poder tener una representación del producto más detallada, estos sinónimos fueron extraídos de la página http://www.diccionariodesinonimos.es/ a través de un script de Python de la misma manera que se procedió con los verbos, se construye un diccionario en memoria como muestra la figura 3.16.



Palabra	Sinonimos
abastecer	aprovisionar,suministrar,proveer
expresivo	efusivo, entusiasta, vehemente
grupo	faccion, bando, banderia
economico	moderado, barato, modico
efectuar	ejecutar, proceder, realizar
lujuria	lascivia, erotismo, sensualidad
madera	lena, tronco, palos

Figura 3.16: Creación diccionario de Sinónimos

Fuente: [Elaboración propia]

3.4. TEXT MINING

Se hará una adaptación de KDD en esta sección, debido a que se realizara una minería de textos en vez de la minería de datos, mediante los siguientes pasos:

3.4.1. ANÁLISIS DE SENTIMIENTO

Se procedió al etiquetado de palabras con sus respectivas polaridades obtenidas de la traducción del corpus de palabras positivas, negativas de la University of Illinois en Chicago el cual fue compilado por muchos años desde la investigación de (Hu & Liu, 2004) https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html como muestra la figura 3.17 y la figura 3.18 para analizar el sentimiento de los comentarios en base a la ecuación.

abolir, abominable, abominablemente, abominar, abominación, abortar, abortado, aborta, desgastar, abrasivo, abrupto, abruptamente, fugarse, ausencia, despistado, ausente, absurdo, absurdo, absurdamente, absurdo, abuso, abusado, abusos, abusivo, abismal, abismalmente, abismo, accidental, molestia, molestias, irritado, molesto, molesto, molesta, anómalo, anomalía

Figura 3.17: Fragmento de Palabras negativas

Fuente: [https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html]

asombroso, impresionantemente, brisa, brillante, aclarar, más brillante, brillantez, brillantes, brillante, brillantemente, enérgico, fraternal, engatusar, calma, calmante, calma, capacidad, capaz, dote, cautivar, cautivador, despreocupado, devolución de dinero, devoluciones, pegadizo, celebrar, celebrado, celebración, campeón, carisma, carismático caritativo, encanto, encantador, encantadoramente

Figura 3.18: Fragmento de Palabras positivas

Fuente: [https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html]

Para calcular la polaridad de un comentario se procedió a calificar la oración de acuerdo a la cantidad de palabras negativas y positivas que contenga, la misma fue calculada bajo la ecuación, que permite ver la polaridad de un comentario:

$$P(x) = \frac{\sum PP}{\sum PN}$$

Dónde: x es el comentario introducido

$\sum PP$, es la suma total de palabras positivas

$\sum PN$, es la suma total de palabras negativas

La ecuación demostró ser eficaz con en un 34% de los casos probados, como siguiente paso se implementó un clasificador Bayesiano en el lenguaje Python con unos 637 casos previamente clasificados por diferentes personas, se muestra un pequeño ejemplo de los comentarios polarizados en la Tabla 3.2.

Tabla 3.2: Comentarios clasificados

Fuente: [Elaboración propia]

Comentario	Clasificador
la comida esta rica.	pos
nos sentimos felices con tu presencia.	pos
me siento muy bien.	pos
este es el mejor trabajo.	pos
que vista tan maravillosa	pos
estoy muy feliz de conocerte	pos
no me gusta este restaurante	neg
estoy cansado de esta tarea.	neg
no puedo soportar esto.	neg
no me gusta caminar.	neg
mi jefe es horrible.	neg
la vida no es justa	neg
la comida esta rica.	pos
nos sentimos felices con tu presencia.	pos
me siento muy bien.	pos
este es el mejor trabajo.	pos
que vista tan maravillosa	pos
estoy muy feliz de conocerte	pos
no me gusta este restaurante	neg
estoy cansado de esta tarea.	neg

Como resultado obtuvimos un 50% de aciertos en la clasificación de polaridades emocionales de los comentarios. Debido a los resultados mostrados por el uso de las palabras categorizadas y la implementación de la red Bayesiana, se eligió el servicio de *meaningcloud* que incluye entre sus herramientas un clasificador de sentimiento entrenado con un corpus mucho más amplio, el servicio trabaja con entidades, conceptos, y clasifica en forma estándar de acuerdo a la configuración hecha al hacer la petición al API. La clasificación se la realiza a todo el conjunto del comentario, también se puede observar la subjetividad y objetividad del comentario. Después del análisis de los resultados obtenidos con los dos métodos anteriores para etiquetar los comentarios, se procedió a la construcción del módulo de integración con el API de *meaningCloud*, que permitió categorizar según la polaridad todos los textos obtenidos, como muestra la tabla 3.3.

Tabla 3.3: Corpus polarizado de uno de los perfiles

Fuente: [Elaboración propia]

Opinión	polaridad
donde hacemos el intercambio v	NONE
explicacion grafica	NONE
ir a mexico para revisar mis dientes jajaja xd algun dia xd invitacion	P
uhhhhh una verguenza total y luego quieren respeto	N
si daja lo dice pues daja tiene razon	P
primo perdido gracias vamo a trabajar por todas las metas y deseos saludos a la familia espero verlos pronto	N
3 una nina de 5anos saco la foto se veian muy bien mis preciosos uno ya no esta conmigo solo me quedo con hermosos recuerdos	P
tu fanatico esto es cierto xd xd	P
uhhhhhh jajaja su majestad su merced asi lo haremos v xd	P
siempre fui maldita xd	NEU
tushe 3	NONE
ja ja ja algo mucho mejor je je je xd	P
gracias 3 yoo nancy	P
y asi nacio safedriving 3 nuestrosinicios	P
hay que conseguir una	P
gracias jorge	P
traumatis ugudis xd xd jajaja	P
y no dejar de pensar en pokemon s jajaja	P
oh no donde	NONE
diosito mi codigo que o o nooooo no puede ser aun puedo arrepentirme verdad v no se por que lo lei con voz de alvaro garcia linea cuando estaba de meteorologo xd v	NEU
vamos arruinando algunas cosas	N+

Una vez obtenidos los corpus polarizados de todos los textos recolectados, se procede a la identificación de características de los usuarios, las palabras más frecuentes, el grado de dispersión de las mimas. La identificación de sentimiento en los comentarios que se relacionen con los descriptores de los productos para los cuales se hizo el análisis de nicho de mercado.

3.4.2. NUBE DE PALABRAS

El módulo de nube de palabras permite visualizar las palabras de acuerdo a sus frecuencias y mostrarlas en una gráfica donde las que son más repetidas tienen un tamaño más alto que las que no lo son. En la figura 3.19 podemos observar al perfil 14 y ver cuáles son las palabras más recurrentes, evitando observar los *stopwords*, y realizado la lematización, para tener un panorama más detallado de aquello realmente importante.

Figura 3.19: Nube de palabras del perfil 14

Fuente: [Elaboración propia]

3.4.3. DIAGRAMA DE FRECUENCIAS ENLAZADAS

El diagrama de frecuencias enlazadas de la figura 3.20 muestra las frecuencias de las palabras desde el más frecuente hacia el menos frecuente del perfil 14. Esta grafica permite ver con mucha más claridad las frecuencias individuales de las palabras. Por la parte izquierda podemos observar la cantidad y en la parte inferior vemos las palabras.

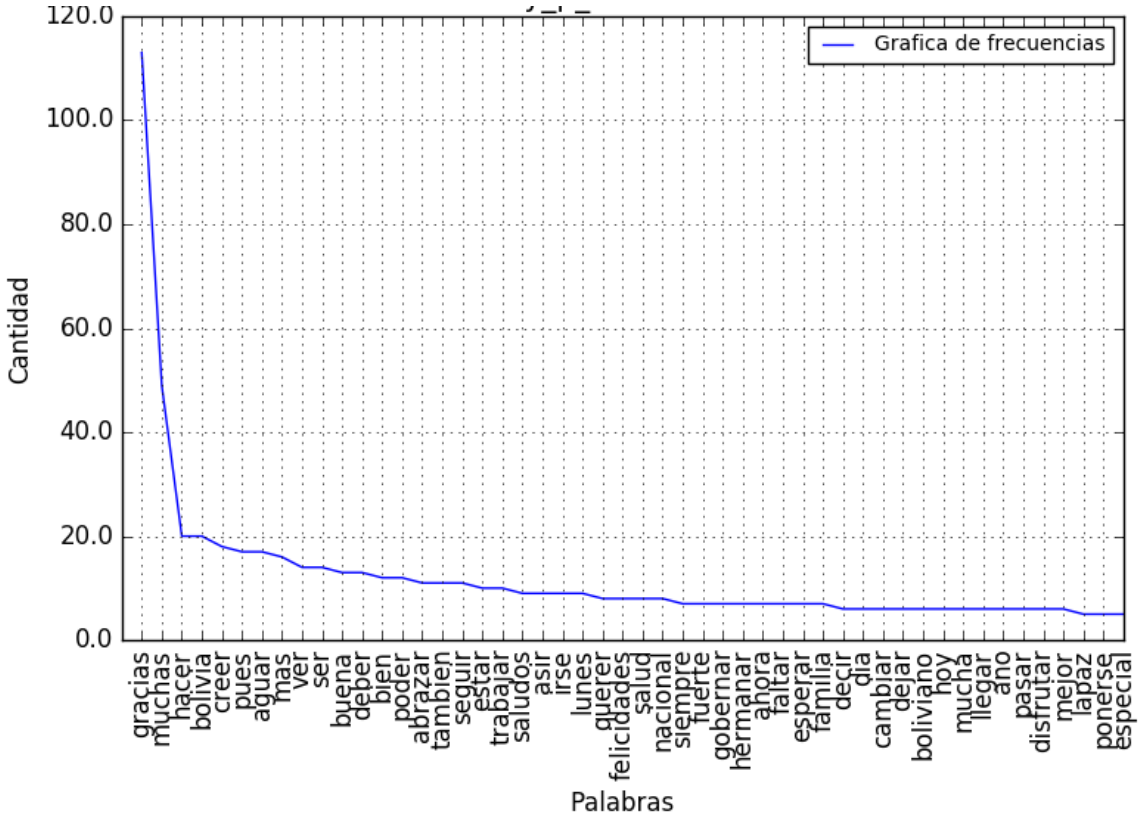


Figura 3.20: Diagrama de frecuencias perfil 14

Fuente: [Elaboración propia]

3.4.4. DIAGRAMA DE FRECUENCIAS NO ENLAZADAS

El diagrama de frecuencias no enlazadas a diferencia de la anterior grafica no muestra los datos a través de una línea, al contrario los datos se muestran conectados a sus frecuencias por puntos, como lo muestra la figura 3.21.

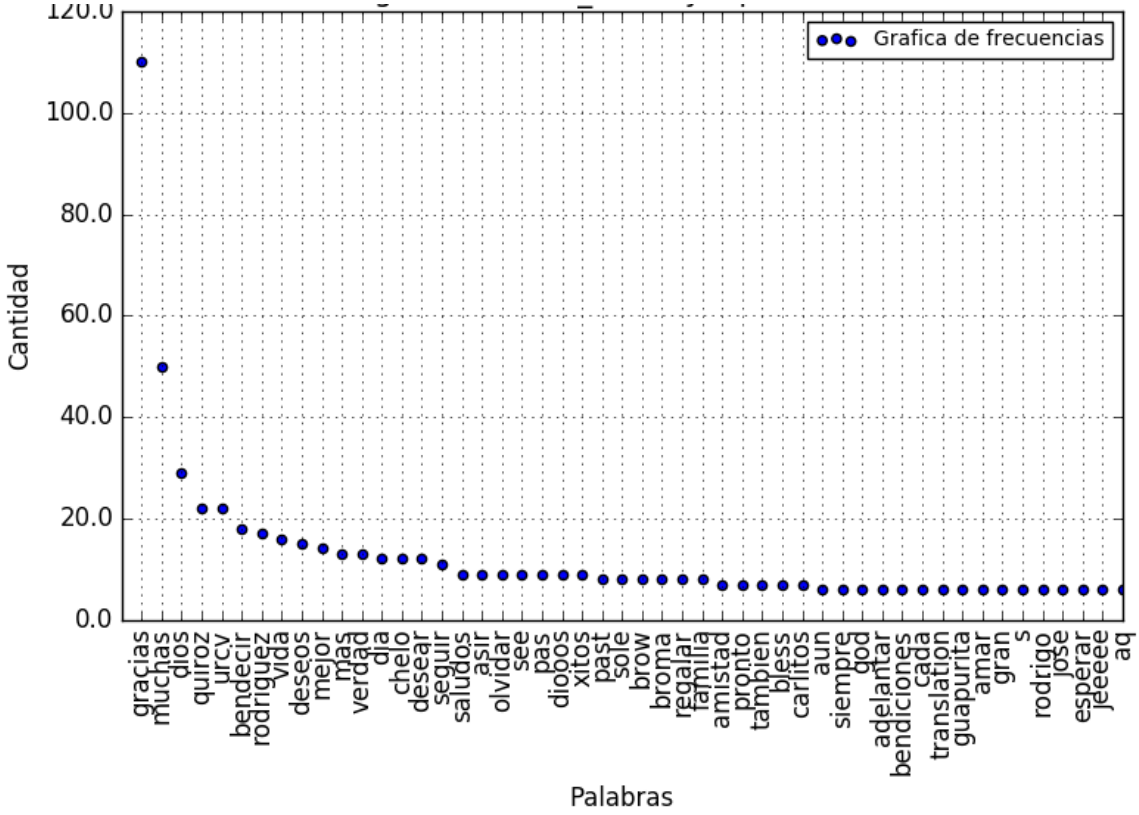


Figura 3.22: Diagrama de frecuencias no enlazadas del perfil 3

Fuente: [Elaboración propia]

3.4.5. DIAGRAMA DE DISPERSIÓN

El diagrama de dispersión permite analizar si existe algún tipo de relación entre dos variables. Por ejemplo, puede ocurrir que dos variables estén relacionadas de manera que al aumentar el valor de una, se incremente el de la otra. En este caso hablaríamos de la existencia de una correlación positiva. También podría ocurrir que al producirse una en un sentido, la otra derive en el sentido contrario; por ejemplo, al aumentar el valor de la variable x, se reduzca el de la variable y. Entonces, se estaría ante una correlación negativa. Si los valores de ambas variable se revelan independientes entre sí, se afirmaría que no existe correlación. La figura 3.23 permite ver una dispersión casi uniforme de las palabras, en este perfil se tomó 6000 palabras del texto obtenido para graficarlo.

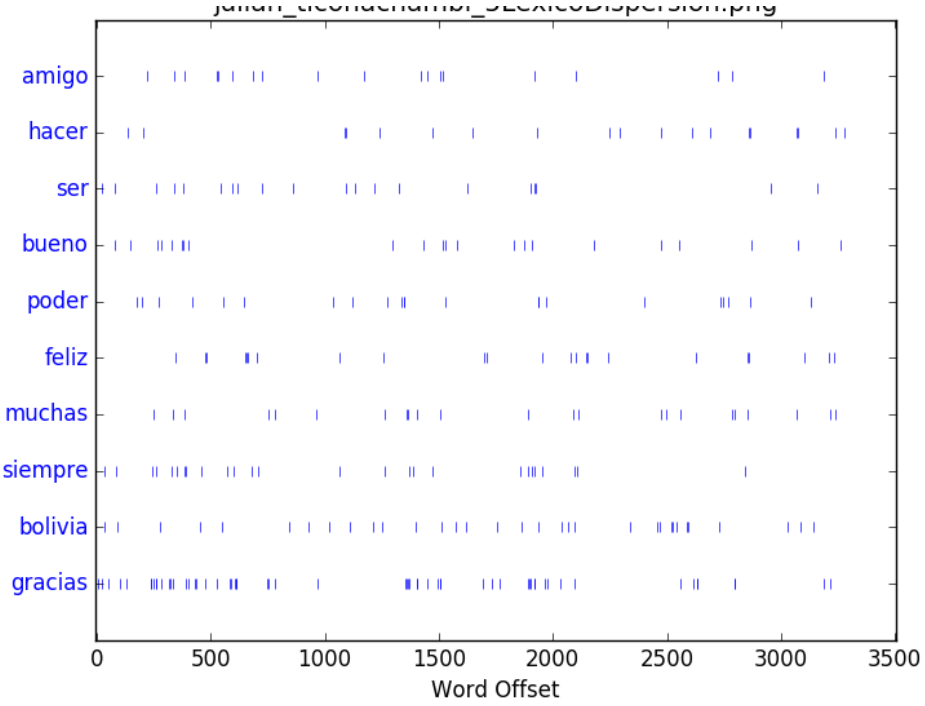


Figura 3.23: Gráfico de dispersión de las 10 palabras más mencionadas por el perfil 1

Fuente: [Elaboración propia]

3.5. INTERPRETACIÓN / EVALUACIÓN

Una vez desarrolladas estas herramientas se procedió a describir los productos para la interpretación de los datos obtenidos con el corpus polarizado de los perfiles y la evaluación de la polaridad de compra por parte de los dueños de los perfiles, en la tabla 3.4 se muestra descrito el producto 4.

Tabla 3.4: Descripción del producto 4

Fuente: [Elaboración propia]

Producto 4	Características
	linux, blanco, azul, pinguino, geek, 1996, debian, ubuntu

Seguido este paso se procede a buscar las características en los perfiles recolectados, en este caso identificaremos al usuario 1 y observamos su nube de palabras en la figura 3.24 en donde no se observa aun alguna relación con el producto.

Figura 3.24: Nube de palabras del perfil 1

Fuente: [Elaboración propia]

El siguiente paso que se tomo fue emparejar las características del producto con todo el perfil del usuario 1, extrayendo los comentarios donde alguna de estas palabras aparecía, como ya se polarizo los mismos en los pasos anteriores, extraemos el comentario y su polaridad, además de contar las ocurrencias de las características por cada comentario, para al final proceder con la fórmula de polaridad:

$$Pos = \sum_{1}^{n} Pp * Fq - \sum_{1}^{n} Pn * Fq$$

Donde:

Pos, es la polaridad de compra con tres posibles valores neg (<=1), pos (>=1) y neutro (=0).

Pp, son todas los comentarios con polaridad positiva.

Pn, son todos los comentarios con polaridad negativa.

Fd, es la frecuencia de las características en cada uno de los comentarios.

Obtendremos las polaridades además de las palabras que se mencionan en cada comentario con su respectiva frecuencia de palabras emparejadas con el producto 4 del perfil 1, como muestra la tabla 3.5.

Tabla 3.5: Polaridades de los comentarios del perfil 1 relacionados con el producto 4

Fuente: [Elaboración propia]

Polaridad	Características	Frecuencia características	Polaridad	Características	Frecuencia características
N	linux	1	P	azul	1
N	azul	1	P	linux	1
N	linux	1	P	linux	1
N	ubuntu ubuntu	2	P	linux	1
N	linux azul	2	P	linux	1
P	linux	1	P	ubuntu	1
P	linux ubuntu	2	P	linux	1
P	linux	1	P	linux	1
P	azul	1	P	linux	1
P	linux	1	P	blanco	1

La anterior tabla no muestra todos los comentarios, ni sus polaridades debido a que se quiere identificar la polaridad de compra solo del producto 4, entonces solo se seleccionaron los comentarios relacionados con el ítem 4. Seguido esto se procedió a aplicar la formula, de la siguiente manera:

$$Pos = \sum_{1}^{n} Pp * Fq - \sum_{1}^{n} Pn * Fq$$

Tomamos cada comentario positivo como 1 (un comentario) y lo multiplicamos por su respectiva frecuencia de las palabras emparejadas con el producto 4, procedemos de la misma forma con los comentarios negativos, lo que resulta en:

$$Pos = 16 - 7$$

$$Pos = 9$$

Como la polaridad de compra es mayor a cero, se concluye que el perfil 1 es un posible comprador del producto 4. Ahora podemos ver el diagrama de frecuencias, solo de los comentarios donde se mencionó alguna de las características del producto en la figura 3.25.

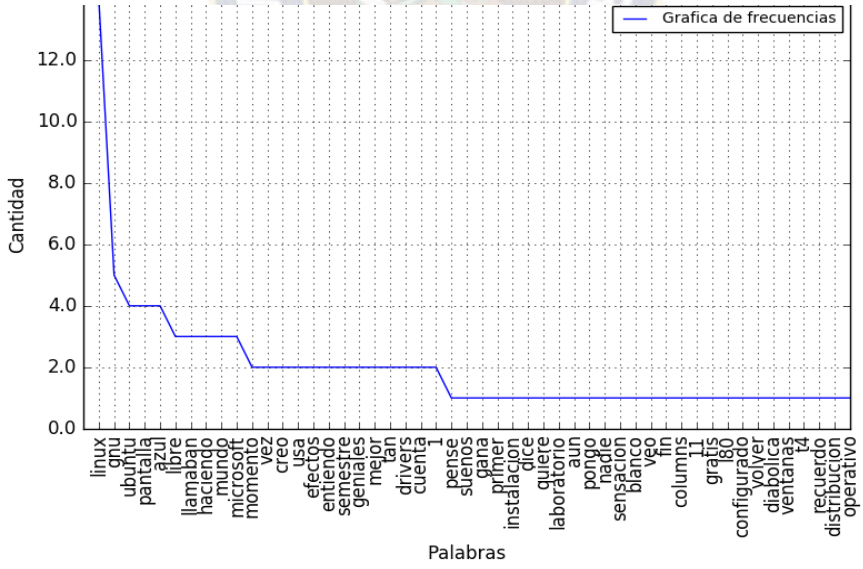


Figura 3.25: Diagrama de frecuencia de perfil 1, emparejados con el producto 4

Fuente: [Elaboración propia]

Se pudo observar la alta frecuencia de palabras relacionadas con la descripción del producto 4. Como siguiente paso pudo apreciar la nube de palabras aislando solo los comentarios que contengan algunas de las descripciones del producto 4 (Figura 3.26).

Figura 3.26: Nube de palabras perfil 1 emparejado con descripción del producto 4

Fuente: [Elaboración propia]

Se pudo observar de la figura 3.26 algunos otros conceptos relacionados con la descripción del producto 4 y evidenciar que cuando se habló de alguna de las características se ve una alta frecuencia, y se pudo observar la gráfica de dispersión solo de las características del producto 4 con los comentarios del perfil 1 que contenían alguna ocurrencia de las mismas, como se puede apreciar en la figura 3.27.

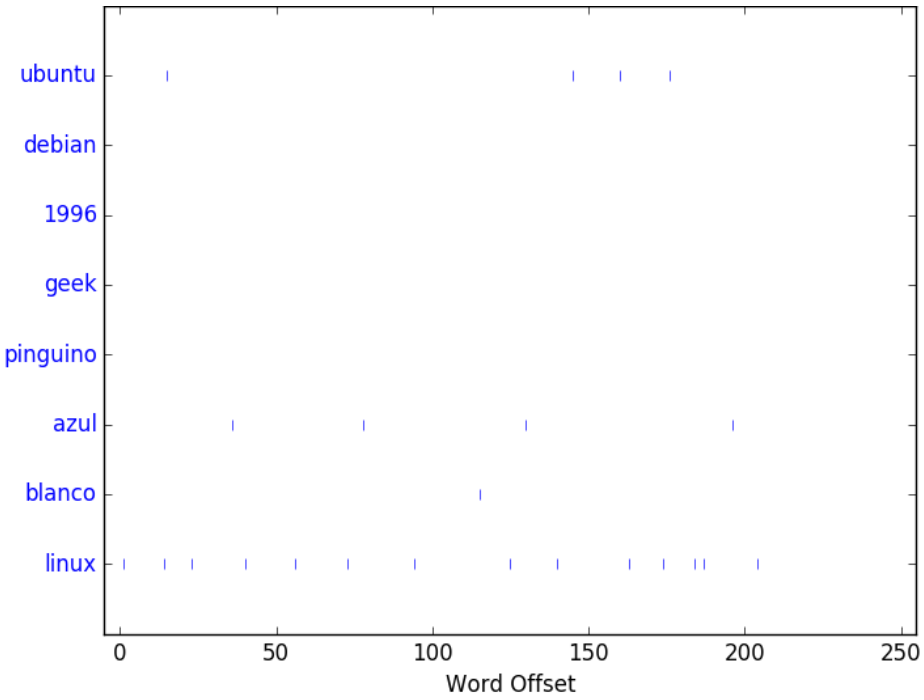


Figura 3.27: Diagrama de dispersión perfil 1, comentarios emparejados con producto 4

Fuente: [Elaboración propia]

Al final de todo este análisis se pudo concluir que el perfil 1 muestra una polaridad más positiva hacia las características del producto 4 además de ver una alta frecuencia de las mismas. Se concluye que el perfil 1 es un posible comprador del producto 4.

CAPITULO IV

RESULTADOS

4.1. PRUEBA DE HIPÓTESIS

Para la prueba de hipótesis se trabajó con la fórmula de cálculo de polaridad descrito en el capítulo anterior, la misma que se describe a continuación:

$$Pos = \sum_{1}^{n} Pp * Fq - \sum_{1}^{n} Pn * Fq$$

Donde:

Pos, es la polaridad de compra e indica cuan positivo, negativo o neutro se muestra la persona cuando menciona alguna de las características del producto.

Pp, son todas los comentarios con polaridad positiva.

Pn, son todos los comentarios con polaridad negativa.

Fd, es la frecuencia de las características en cada uno de los comentarios.

Pos puede tomar uno de estos valores: negativo (<=1), positivo (>=1) ó neutro (=0)

Con esta fórmula podemos calcular la posibilidad de compra por parte de un usuario.

Con la fórmula se minaron las publicaciones de varios usuarios los mismos que se exponen a continuación eligiendo tres casos representativos, de aceptación del producto, rechazo y sin conclusión sobre si compraría el producto o no. El proceso seguido en cada uno de ellos es el siguiente: Primero se obtiene una nube de palabras como resultado del minado en la página de cada usuario. Identificada la tendencia en las expresiones se le presentan artículos descritos con palabras cercanas a la tendencia y a continuación se aplica una encuesta consultando la preferencia de compra.

4.1.1. ESTUDIO DE CASOS

Caso 1: (caso aceptación del producto 6)

Perfil 16

La figura 4.1 mostró la tendencia del perfil 16, así que se le proporciono el producto 6 que coincidía con alguna de las características del producto 6.

Figura 4.1: Nube de palabras perfil 16

Fuente: [Elaboración propia]

Ahora describiremos el producto 6 en la tabla 4.1.

Tabla 4.1: Descripción del producto 4.1

Fuente: [Elaboración propia]

Producto 4	Características
	dota, negro, héroe, videojuego, videojuegos, game, games, héroes, warcraft, wow

Después de la descripción del producto se procede a la identificación de los comentarios donde se mencionó las características del producto, como se muestra en la tabla 4.2.

Tabla 4.2: Polaridades mostradas por el perfil 16 hacia el producto 6 con frecuencias

Fuente: [Elaboración propia]

Polaridad	Características	Frecuencia características	Polaridad	Características	Frecuencia características
N	dota	1	P	dota dota	2
P	dota	1	N	dota	1
P	dota	1	P	dota	1
N	dota	1	N	dota	1
P	dota	1	P	dota	1
P	dota	1	P	dota	1
P	dota	1	P	dota	1
P	dota	1	P	dota	1
P	dota	1	N	dota	1
P	dota	1	N	dota	1
P	dota	1	P	dota	1
P	dota	1	P	dota	1
P	dota	1	P	dota	1
P	dota	1	P	dota	1
P	dota dota	2	P	dota	1
P	dota	1	P	dota	1
P	dota	1	P	dota	1
P	dota	1	P	dota	1
P	dota	1			

Aplicando la fórmula de la polaridad de compra, y sumando los datos con la tabla 4.2.

$$Pos = 33 - 6$$

$$Pos = 27$$

La nube de palabras asociadas con el producto en los comentarios se visualiza en la figura 4.2 en el cual se observa una alta frecuencia de las características del producto.

Figura 4.2: Nube de palabras perfil 16 emparejadas con el producto 6

Fuente: [Elaboración propia]

Observaremos el diagrama de frecuencias de la figura 4.3 con los comentarios que contienen alguna de las descripciones del producto.

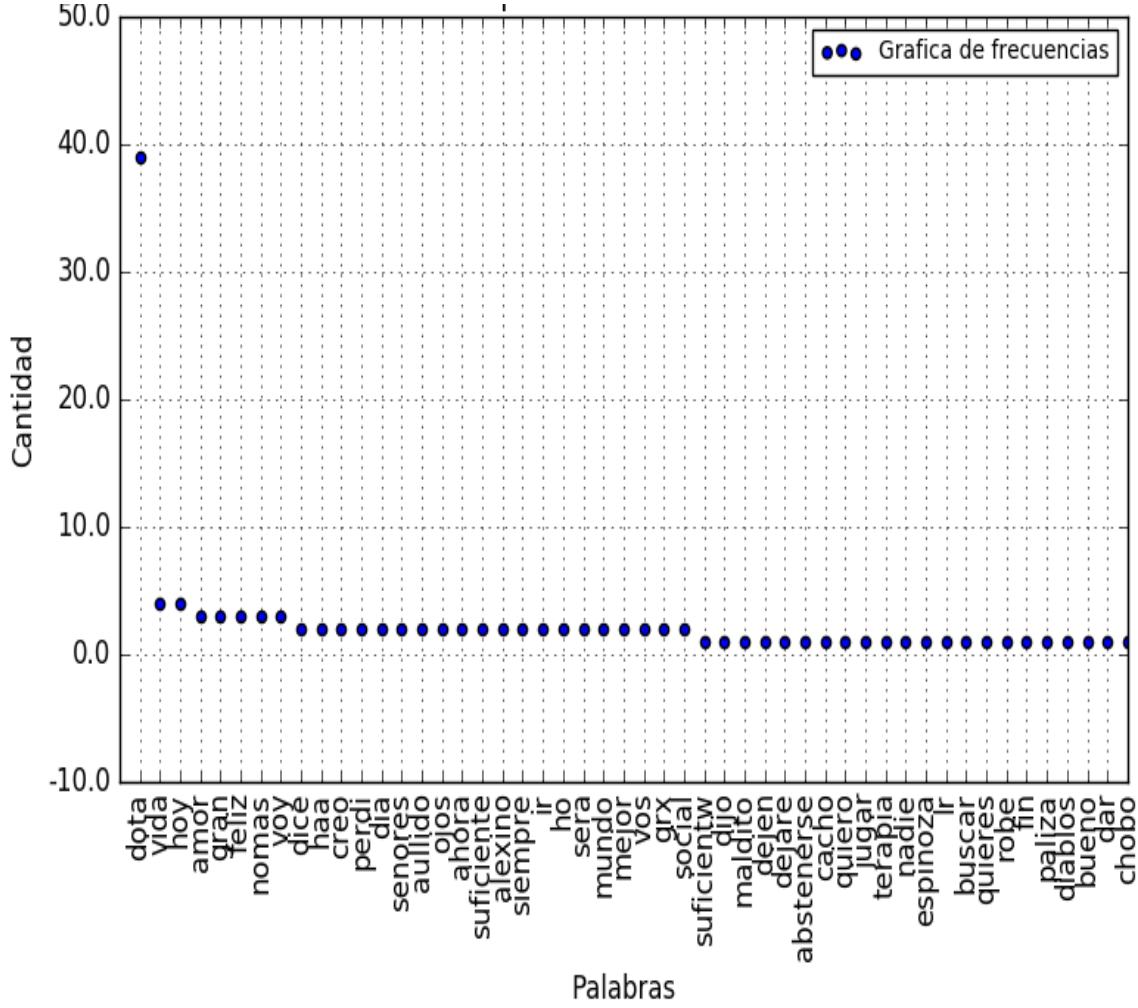


Figura 4.3: Nube de palabras perfil 16

Fuente: [Elaboración propia]

La grafica 4.3 muestra de otra forma las frecuencias de las palabras en cuanto se refiere al producto 6. El resultado del análisis realizado con las técnicas de minería de textos permite afirmar que el perfil 16 es un posible comprador del producto 6, descrito con anterioridad, pues el mismo se muestra positivo a características del producto. En la tabla 4.3 se puede observar como el perfil 16 contestó a la pregunta sobre si compraría el producto.

Tabla 4.3: Encuesta realizada al usuario 16

Fuente: [Elaboración propia]

Encuesta lista de productos	
Producto 1	Si lo compraría
Producto 2	No me decido
Producto 3	Si lo compraría
Producto 4	No me decido
Producto 5	No me decido
Producto 6	Si lo compraría
Producto 7	Si lo compraría
Producto 8	No lo compraría
Producto 9	Si lo compraría
Producto 10	No lo compraría

Comparando el resultado obtenido a través de la minería de textos y la respuesta del perfil 16 se concluye que el caso fue de éxito en la afirmación de compra.

Caso 2: (caso rechazo producto 2)

Perfil 15

Figura 4.4: Nube de palabras perfil 15

Fuente: [Elaboración propia]

La figura 4.4 permitió observar que producto no podría ser aceptado por el perfil 15, el mismo que se describe en la tabla 4.4.

Tabla 4.4: Descripción del producto 2

Fuente: [Elaboración propia]

Producto 2	Características
	azul, windows, cuadrado, microsoft

Después de la descripción del producto se procede a la identificación de los comentarios donde se mencionó las características del producto, como se muestra en la tabla 4.5.

Tabla 4.5: Polaridades mostradas por el perfil 15 hacia el producto 2 con frecuencias

Fuente: [Elaboración propia]

Polaridad	Características	Frecuencia características
N	windows	1
P	windows	1
N	windows	1

Aplicando la fórmula de la polaridad de compra, y sumando los datos con la tabla 4.5.

$$Pos = 1 - 2$$

$$Pos = -1$$

La nube de palabras asociadas con el producto en los comentarios se visualiza en la figura 4.5. en el cual se observa la frecuencia de las características del producto descrito en la tabla 4.4. que es relativamente alta pero que viene asociada con términos que se podría considerarse opuestos, lo cual llevo la balanza hacia otra dirección.

Figura 4.5: Nube de palabras perfil 15 emparejadas con el producto 2

Fuente: [Elaboración propia]

Como la polaridad de compra en este caso es negativa, se puede concluir que cuando se habla de alguna de las características del producto 2 se habla negativamente, además podemos observar en la nube de palabras, conceptos adversos a una de las características del producto. La grafica de dispersión de la figura 4.6 permite observar la dispersión de los descriptores del producto 2.

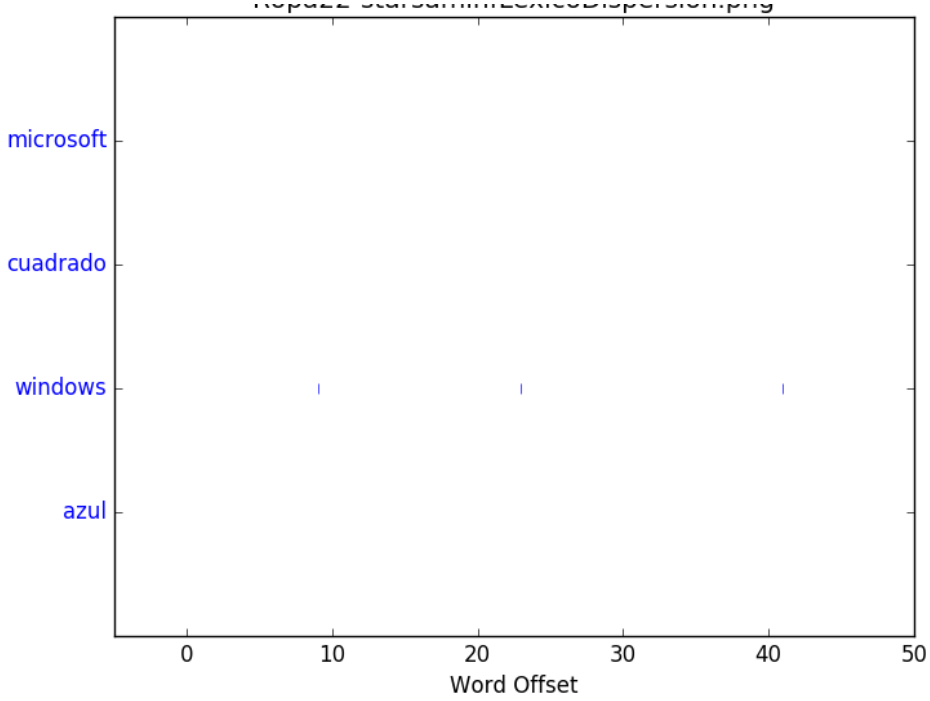


Figura 4.6: Grafica de dispersión perfil 15, emparejado con producto 2

Fuente: [Elaboración propia]

La tabla 4.6 muestra las respuestas del perfil 15 a la encuesta, y podemos observar que el usuario respondió rechazando el producto 2 concordando con el análisis realizado.

Tabla 4.6: Encuesta realizada al perfil 15

Fuente: [Elaboración propia]

Encuesta lista de productos	
Producto 1	No me decido
Producto 2	No lo compraría
Producto 3	No me decido
Producto 4	Si lo compraría
Producto 5	Si lo compraría
Producto 6	Si lo compraría
Producto 7	No lo compraría
Producto 8	No lo compraría
Producto 9	No lo compraría
Producto 10	No me decido

Caso 3: (caso neutro producto 5)

Perfil 14:

La figura 4.7 mostró la tendencia muy imprecisa del perfil 14, así que se le proporciono el producto 6 que coincidía con alguna de las características del producto 6.

Figura 4.7: Nube de palabras perfil 14

Fuente: [Elaboración propia]

La figura 4.7 no permitió asegurar la compra de algún producto para el perfil 14, se eligió el producto 5 descrito en la tabla 4.7.

Tabla 4.7: Descripción del producto 5

Fuente: [Elaboración propia]

Producto 5	Características
	gnome, linux, debian, ubuntu, pie, negro, entorno

Después de la descripción del producto se procede a la identificación de los comentarios donde se mencionó las características del producto, como se muestra en la tabla 4.8.

Tabla 4.8: Polaridades mostradas por el perfil 14 hacia el producto 5 con frecuencias

Fuente: [Elaboración propia]

Polaridad	Características	Frecuencia características
N	negro	1
P	pie	1

Aplicando la fórmula de la polaridad de compra, y sumando los datos con la tabla 4.8.

$$Pos = 1 - 1$$

$$Pos = 0$$

La figura 4.8 muestra la nube de palabras con muy poca información acerca del producto y sus características. No podemos afirmar la presencia de algún descriptor del producto que muestre alta relevancia.

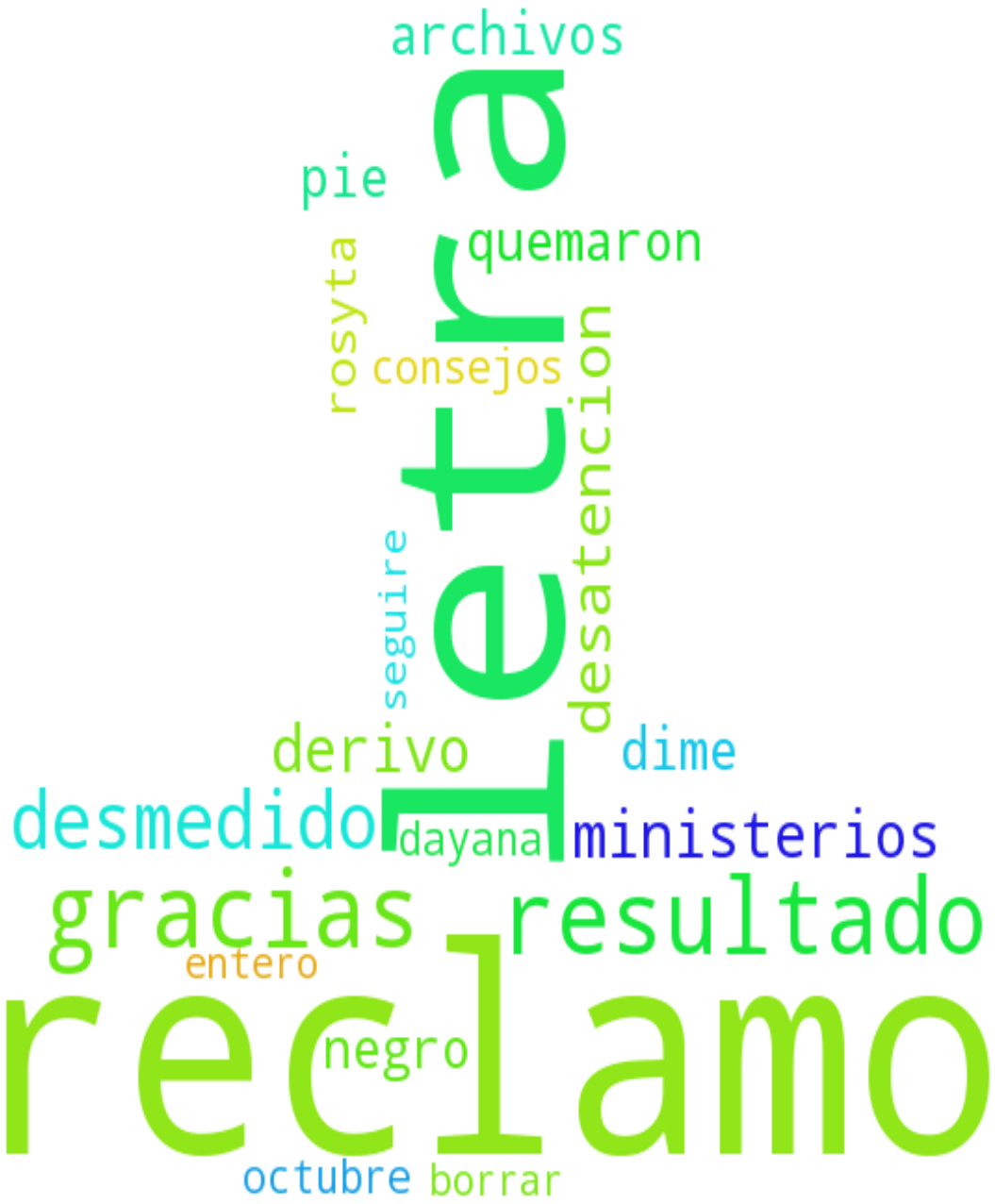


Figura 4.8: Nube de palabras perfil 14

Fuente: [Elaboración propia]

La figura 4.9 muestra la dispersión de las palabras, y muestra una uniformidad en la aparición de las palabras.

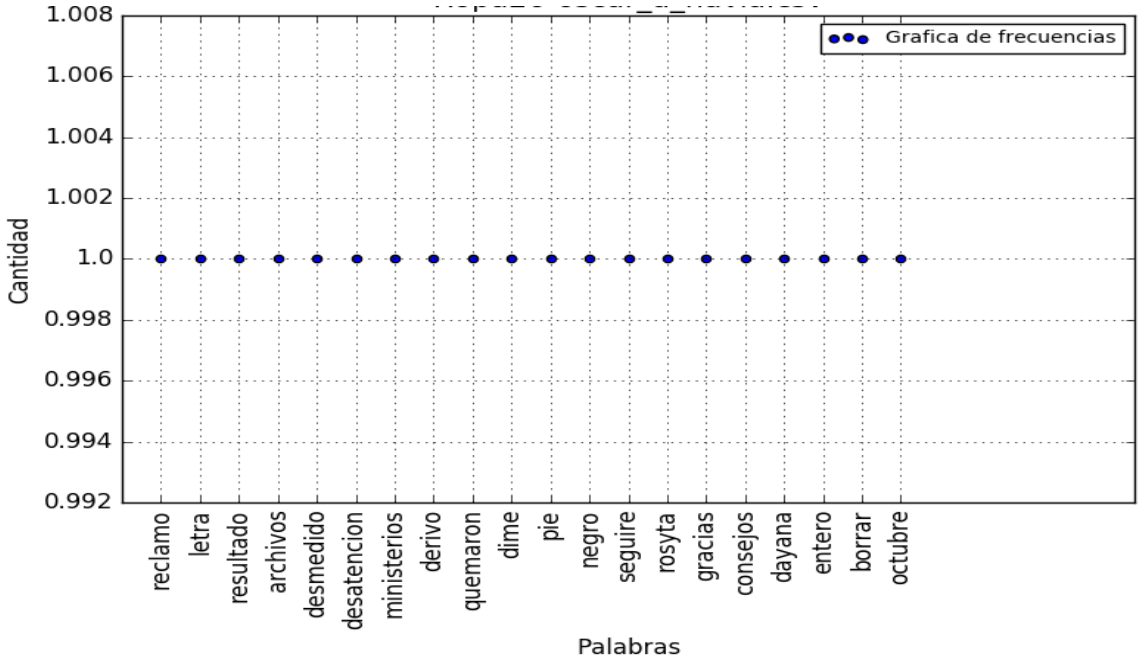


Figura 4.9: Diagrama de frecuencias perfil 14

Fuente: [Elaboración propia]

La tabla 4.9 (La encuesta realizada al perfil 14) afirma lo indicado por el análisis realizado, que refuerza el desconocimiento de que el perfil 14 se decida a comprar o no el producto 5.

Tabla 4.9: Encuesta realizada al perfil 14

Fuente: [Elaboración propia]

Encuesta lista de productos	
Producto 1	No me decido
Producto 2	No lo compraría
Producto 3	Si lo compraría
Producto 4	No lo compraría
Producto 5	No me decido
Producto 6	Si lo compraría
Producto 7	No me decido
Producto 8	No lo compraría
Producto 9	No me decido
Producto 10	No lo compraría

4.2. RESULTADOS

El estudio se hizo con 16 páginas de Facebook. En promedio un usuario tiene 8000 palabras para analizar. La tabla 4.10 muestra los resultados obtenidos gracias a la encuesta realizada a las 16 personas dueñas de los perfiles.

Tabla 4.10: Resultado de la encuesta realizada a cada usuario sobre los 10 productos

Fuente: [Elaboración propia]

Perfiles\Productos	1	2	3	4	5	6	7	8	9	10
perfil 1	si	si	no	si	si	duda	si	si	si	no
perfil 2	si	duda	no	si	duda	no	si	duda	no	duda
perfil 3	si	no	duda	si	no	no	si	no	si	no
perfil 4	si	duda	si	si	si	no	duda	no	duda	no
perfil 5	si	duda	no	no	si	no	no	si	si	no
perfil 6	no	si	no	si	no	si	si	si	si	no
perfil 7	si	duda	duda	si	no	si	no	no	si	no
perfil 8	no	no	no	si	si	duda	si	no	no	no
perfil 9	si	no	si	si	si	duda	no	no	no	duda
perfil 10	duda	no	si	si	no	no	no	no	si	no
perfil 11	duda	si	duda	si	no	no	si	no	si	no
perfil 12	no	no	no	si	no	no	no	si	si	no
perfil 13	duda	no	si	si	no	no	si	si	no	no
perfil 14	duda	no	si	no	duda	si	duda	no	duda	no
perfil 15	duda	no	duda	si	si	si	no	no	no	duda
perfil 16	si	duda	si	duda	duda	si	si	no	si	no

La respuesta si refiere a que el usuario compraría el producto, la respuesta no indica que el no compraría el producto y la duda da a entender que no se decide si comprar el producto o no hacerlo. En contraste con la tabla 4.10 se elaboró la tabla 4.11 que indica que respuestas coincidieron con las indicadas por el análisis.

Tabla 4.11: Resultados acertados

Fuente: [Elaboración propia]

Perfiles	1	2	3	4	5	6	7	8	9	10
perfil 1		si		si	si		si			no
perfil 2	si	duda	no		duda	no		duda		duda
perfil 3	si			si					si	
perfil 4	si	duda	si						duda	
perfil 5	si					no	no	si	si	
perfil 6	no					si	si			
perfil 7	si	duda							si	
perfil 8				si	si	duda				
perfil 9			si	si	si	duda				duda
perfil 10			si	si					si	
perfil 11	duda						si		si	
perfil 12	no	no						si		
perfil 13			si					si		
perfil 14			si		duda	si	duda	no		
perfil 15		no	duda			si				duda
perfil 16	si	duda		duda	duda	si	si			

La figura 4.10 muestra el grado de aciertos y desaciertos obtenidos en la presente investigación que permitió identificar al posible comprador en un 40%.

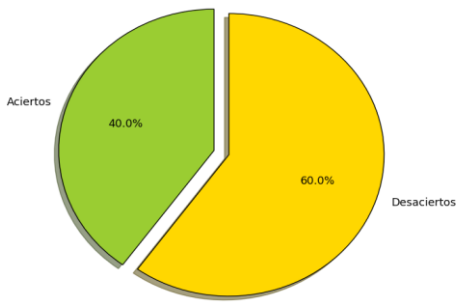


Figura 4.10: Gráfico de aciertos y desaciertos de la investigación

Fuente: [Elaboración propia]

CAPITULO V

5.1. CONCLUSIONES Y RECOMENDACIONES

El objetivo general fue cumplido en su totalidad, al identificar posibles compradores que componen un nicho de mercado. Se analizó al posible público. Los comentarios quedaron clasificados según la polaridad que cada uno tiene. Además de desarrollar herramientas para el idioma español y se pudo identificar características en cada usuario para poder construir publicidad específica para cada grupo.

5.1.1. CONCLUSIONES

Los perfiles con poca actividad no revelan ningún patrón para asociarlo con preferencias de compra o rechazo, debido a la poca cantidad de textos analizados.

La investigación muestra ser 40% efectiva en la identificación de posibles compradores un porcentaje alto en comparación a otras estrategias de marketing digital que logran menores resultados.

El clasificador bayesiano mostró ser eficiente en un 53% en la identificación de la polaridad emocional con una muestra de 637 comentarios clasificados en positivos, negativos y neutros. Se pudo evidenciar que se requiere de muchos más ejemplos para la polarización de un comentario, además de ser necesario que la clasificación sea realizada por diferentes especialistas como lingüistas, sociólogos y otros.

Las herramientas para la minería de textos están disponibles siempre y cuando se dirijan al estudio en el idioma Ingles, siendo esto una limitación para documentos escritos en español, razón por la cual se desarrollaron herramientas de limpieza de texto en español, con la construcción de *stopwords* para la red social de Facebook. La aplicación de estrategias de manejo de la información, los programas de diagramas de dispersión, frecuencia y nube de palabras.

La utilidad del diccionario de verbos con sus respectivas conjugaciones verbales demostró ser de mucha ayuda en la limpieza del texto, y de ayudar a la visualización de los mismos.

No se puede identificar la polaridad emocional de un comentario solo por el hecho de contener palabras negativas o positivas, fue necesario utilizar un diccionario de palabras negativas y positivas desarrollado por la Universidad de Illinois en Chicago, el mismo que fue traducido para el análisis de polaridad.

5.1.2. RECOMENDACIONES

Es necesario que se tomen en cuenta ciertas recomendaciones para mejorar este trabajo o futuras investigaciones que tengan algún tipo de relación con esta.

Se recomienda el procesado de imágenes en los perfiles, grupos o páginas para la obtención de textos que coadyuven en la minería, ya que hoy en día los usuarios se expresan mediante imágenes (memes).

Se encontraron palabras asociadas a las características de los productos con la nube de palabras, que servirían para modificar y hacer más aceptable el mismo.

Se debe tener cuidado en el manejo de los caracteres especiales y de acento, aplicar estrategias de manejo de este tipo de información.

Se puede usar videos para obtener mayor información, los mismos que son compartidos por los usuarios.

BIBLIOGRAFÍA

AMPLN, A. M. (30 de Octubre de 2009). Obtenido de http://www.cicling.org/ampln/index-tmp.htm

Anjaria, M., Reddy Guddeti, R. M. (12 de febrero de 2014). A novel sentiment analysis of social networks using supervised learning. Springer-Verlag Wien.

Balakrishnan, V., Lloyd-Yemoh, E. (Agosto de 2014). Stemming and Lemmatization: A comparison of Retrival Performances.

Bassi A., A. (s.f.). *Lematización basada en análisis no supervisado de corpus*. Obtenido de http://users.dcc.uchile.cl/~abassi/ecos/lema.html

BBC MUNDO. (19 de Agosto de 2015). Obtenido de BBC Website: http://www.bbc.com/mundo/noticias/2015/08/150819_difusion_internet_america_latina_cepal_ac

Bird, S., Klein, E., Loper, E. (2009). *Natural Language Processing with Python.* O'Really Media Inc.

Eíto Brun, R., A. Senso, J. (2004). Minería Textual. *El profesional de la información*, 11-27.

Fan, W., Bifet, W. (2 de diciembre de 2012). Mining Big Data: Current Statu, and Forecast to the Future.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 37-54.

Ferro MA, X. (2015). El poder del clic: el consumidor y las nuevas formas. *Revista Internacional de Tecnología, Conocimiento y Sociedad*, 157-167.

Guevara López, R. (Octubre de 2011). Minería de textos en la red social Twitter. México.

Harris, J. (s.f.). *wordcount*. Obtenido de http://www.wordcount.org

Henriquez Miranda, C., Guzman, J., Salcedo, D. (marzo de 2016). Mineria de Opiniones basado en adaptacion al español de ANEW sobre opiniones acerca de hoteles. *Procesamiento del lenguaje natural*, 25-32.

Hu, M., , Liu, B. (22-25 de Agosto de 2004). Mining and Summarizing Customer Reviews. Seattle, Washington, Estados Unidos de Norteamerica.

Imaña, G., , Vásquez, W. (19 de Julio de 2015). *En Bolivia, el avance del marketing digital enfrenta 2 grandes obstáculos: La Razón*. Obtenido de http://www.la-

razon.com/index.php?_url=/suplementos/financiero/Bolivia-marketing-obstaculos-financiero_0_2309169161.html

jsx, g. P. (30 de Octubre de 2015). *Mozilla Foundation*. Obtenido de https://developer.mozilla.org/en-US/docs/Web/Javascript/About_Javascript

Lanzarini, L., Hasperué, W., Estrebou, C., Formia, S., Corbalan, L., Ronchetti, F., . . . Quiroga, F. (2014). Metaheurísticas aplicadas a Procesamiento de Señales y Minería de Datos. *WICC 2014 XVI Workshop de Investigadores en Ciencias de la Computación*, (págs. 203-207).

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan , Claypool Publishers.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to information retrival.* Cambridge University Press. Obtenido de http://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

Miranda Gonzaléz, F., Rubio Lacoba, S., Chamorro Mera, A., , Correia Loureiro, S. M. (2013). Facebook como herramienta de comunicación y venta: un análisis desde la oferta y la demanda*. *Cuadernos de CC.EE. y EE.*, 81-98.

Moreno Sanchez, I., Font-Clos, Francesc, , Corral, A. (22 de Enero de 2016). Large-Scale Analysis of Zipf's Law in English. Argentina.

Moreno, M. (28 de enero de 2016). *TreceBits*. Obtenido de http://www.trecebits.com/2016/01/28/facebook-ya-tiene-1-590-millones-de-usuarios/

Mostafa, M. M. (2013). Social networks' text mining for customer brand sentiments. *ELSEVIER*, 11.

Pacheco Leal, S. D., Días Ortiz, L. G., , Rodolfo, G. F. (2005). El clasificador Naive Bayes en la extraccion de conocimiento de base de datos. *Ingenierías Volumen VIII*, 24-33.

Roca, M. Á. (23 de Diciembre de 2015). *El Deber*. Obtenido de http://www.eldeber.com.bo/economia/entel-baja-tarifas-internet-ahora.html

Russell, M. A. (2013). *Mining the Social Web.* O'Reilly Media.

Santana Mansilla, P., Costaguta, R., , Missio, D. (2014). Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de e-tutores colaborativos. *Revista Iberoamericana de Inteligencia Artificial*, 57-67.

Stavrianou, A., Andritsos, P., , Nicoloyannis, N. (2007). Overview and Semantic Issues of Text Mining. En *SIGMOD Record* (págs. 23-34).

The Psychometrics Centre - Apply Magic Souce. (s.f). *Acerca de nosotros: Apply Magic Sauce*. Obtenido de http://applymagicsauce.com/

Torres Silva, D. A. (Junio de 2013). Diseño y aplicación de una metodología para el análisis de noticias policiales utilizando mineria de textos. Santiago de Chile, Chile.

Uribe Saavedra, F., Rialp Criado, J., , Llonch Andreu, J. (2013). El uso de las redes sociales digitales como herramienta de marketing en el desempeño empresarial. *Cuadernos de administracion Bogota (Colombia)*, 205-231.

ANEXO

Anexo A – Descripción de los productos

Productos	Características
	azul, tejiendo, lazos, diversidad, embajada, estados unidos, feria, libro, bandera, EE.UU., americana, americano
	azul, Windows, cuadrado, Microsoft
	rojo, caminando, caminata, vida, juntos, unidos, acompañados, cáncer, sol
	Linux, blanco, azul, pingüino, geek, 1996, debian, ubuntu
	gnome, linux, debian, ubuntu, pie, negro
	dota, negro, héroe, videojuego, videojuegos, game, games, héroes, warcraft, wow
	dota, negro, rojo, videojuegos, videojuego, warcraft, game, games
	bolivar, choli, bolivarista, entel, coca cola, samsung, celeste, cholis
	Bolivia, verde, escudo, cóndor, blanco, liso, brillo
	tinku, naranja, amarillo, verde, puros, montera, tinkus

DOCUMENTACIÓN