ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

PROGRAMA DE MAESTRIA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA:

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS DE SENTIMIENTOS EN LAS REDES SOCIALES SOBRE PRODUCTOS DE LA MARCA FIDEOS CAYAMBE

AUTOR: ING. NARVAEZ MONTOYA MARIA SALOME

SANGOLQUÍ

2017

ii

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

CERTIFICACIÓN

Certifico que el trabajo de titulación, denominado **"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS DE SENTIMIENTOS EN LAS REDES SOCIALES SOBRE PRODUCTOS DE LA MARCA FIDEOS CAYAMBE"**, realizado por la Señorita **MARÍA SALOMÉ NARVÁEZ MONTOYA**, ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, por lo tanto me permito acreditarlo y autorizar a la señorita **MARÍA SALOMÉ NARVÁEZ MONTOYA** para que lo sustente públicamente.

Ciudad, 04 de septiembre del 2015

MGs. PAUL M. DIAZ ZUÑIGA
DIRECTOR

iii

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

ESCUELA POLITÉCNICA DEL EJÉRCITO
PROGRAMA DE MAESTRIA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORÍA DE RESPONSABILIDAD

Yo, **MARIA SALOME NARVAEZ MONTOYA**, con cédula de identidad N° 0104161914, declaro que este trabajo de titulación **"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS DE SENTIMIENTOS EN LAS REDES SOCIALES SOBRE PRODUCTOS DE LA MARCA FIDEOS CAYAMBE"**, ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas.

Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Ciudad, 04 de septiembre del 2017

MARIA SALOME NARVAEZ MONTOYA
C.C. 0104161914

iv

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

ESCUELA POLITÉCNICA DEL EJÉRCITO
PROGRAMA DE MAESTRIA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORIZACIÓN

Yo, **MARIA SALOME NARVAEZ MONTOYA**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar en la biblioteca Virtual de la institución el presente trabajo de titulación **"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA EL ANÁLISIS DE SENTIMIENTOS EN LAS REDES SOCIALES SOBRE PRODUCTOS DE LA MARCA FIDEOS CAYAMBE"**, cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Ciudad, 04 de septiembre del 2017

MARIA SALOME NARVAEZ MONTOYA
C.C. 0104161914

DEDICATORIA

Esta tesis de graduación va dedicada a mis Padres ya que ellos han sido la mayor motivación y motor para mi superación profesional y personal ya que con su ejemplo, sabiduría y entrega me han enseñado a luchar por mis ideales y sueños

De manera especial a mis Hermanos Aldo, Trotsky y Joffre que han demostrado siempre que ante los obstáculos que nos pone la vida, no siempre se tomará el camino más fácil si no el que nos lleve hacia nuestros objetivos.

AGRADECIMIENTOS

En primera instancia a ese ser superior llamado Dios que hace que cada circunstancia y vivencia tenga su razón de ser, gracias de manera muy especial a mi hermana mayor Camila Valarezo Montoya y Samuelito, que me han demostrado una vez más su apoyo incondicional con su ejemplo y motivación, a mi Tutor Ing. Paúl Díaz por su paciencia y dedicación, por la información y material otorgado Armando Puyol y con cariño a Nicolás Aliaga por ser ese clic en el momento indicado.

INDICE DE CONTENIDO

INDICE DE TABLAS

INDICE DE FIGURAS

Resumen

Se ha establecido una nueva estrategia de marketing para la Marca Fideos Cayambe, en la que se analiza el posicionamiento a través de redes sociales, para obtener un porcentaje de incremento en las ventas de marca a nivel nacional. Es necesario contar con información precisa, que nos permita conocer cuáles son las opiniones de los consumidores acerca de la marca a nivel de la red social Facebook. Según varias investigaciones, existen diferentes técnicas y herramientas con las que se puede realizar la extracción, preparación y procesamiento de datos, aplicando algoritmos de minería de datos. Este proyecto consiste en el desarrollo de la metodología CRISP-DM, mediante la cual, se ha analizado la información recolectada de la red social Facebook; con la finalidad de evaluar y determinar las opiniones de las personas, dejando como resultado una percepción positiva, en base a los comentarios publicados por los seguidores de la marca a nivel nacional.
Las técnicas aplicadas fueron: árboles de clasificación, clustering y regresión lineal. Se ha establecido como la técnica más adecuada, según el análisis realizado, la técnica de árboles de clasificación, por su mayor porcentaje de correlación entre las variables establecidas. WEKA y Qlick View fueron las herramientas que han sido evaluadas y usadas para el procesamiento de dichas técnicas

Palabras Clave:

- **MINERIA DE DATOS**
- **ANALISIS DE SENTIMIENTOS**
- **INTELIGENCIA DE NEGOCIOS**

Abstract

A new marketing strategy has been established for the Cayambe Noodle Brand, which analyzes the positioning through social networks, to obtain a percentage increase in brand sales nationwide. It is necessary to have accurate information, which allows us to know what are the opinions of consumers about the brand at the level of the social network Facebook. According to several investigations, there are different techniques and tools that can be used to extract, prepare and process data, applying data mining algorithms. This project consists of the development of the CRISP-DM methodology, through which the information collected from the social network Facebook has been analyzed; with the purpose of evaluating and determining the opinions of the people, leaving as a result a positive perception, based on the comments published by the followers of the brand at the national level.

The applied techniques were: classification trees, clustering and linear regression. According to the analysis carried out, the technique of classification trees has been established as the most adequate technique, due to its higher percentage of correlation between the established variables. WEKA and Qlick View were the tools that have been evaluated and used for the processing of these techniques

Key words:

- **DATA MINING**
- **FEELING ANALYSIS**
- **BUSINESS INTELLIGENCE**

1. Introducción

En la presente investigación, se cumplirá el siguiente objetivo: Investigar las diferentes técnicas de minerías de datos: árboles de clasificación, clustering y Regresión lineal, que permitirá el análisis de sentimientos de la marca fideos Cayambe, cuya percepción deberá ser extraída de la Red Social Facebook, cuyos resultados facilitará la medición del grado de participación, relación y opinión que tienen los seguidores con respecto a la marca de dicho producto. La metodología de minería de datos CRIPS-DM cuenta con las etapas necesarias para el cumplimiento de los procesos, que nos permitirán llegar de manera eficaz a resultados concretos, de tal manera que cada etapa será desarrollada y detallada para la compresión de los resultados.

Para la aplicación de las técnicas seleccionadas, que procesarán los algoritmos, serán usadas herramientas de libre acceso, que cuentan con las características necesarias para el uso requerido. Estas técnicas serán analizadas y comparadas para obtener los mejores resultados alineados a los objetivos planteados para la marca Fideos Cayambe, y así brindar nuevas oportunidades de mejora y cambios con resultados a corto y largo plazo.

El universo al cual será aplicada esta investigación serán los seguidores de la página de Fideos Cayambe en la red social Facebook de los últimos dos años.

.

2. CAPITULO I

EL PROBLEMA

2.1 Antecedentes

En la actualidad gracias a los medios sociales, las empresas que ofertan productos alimenticios han visto una oportunidad para implementar sus estrategias de marketing, donde pueden incrementar sus publicaciones, dando a conocer las bondades y beneficios que su producto contiene, esto pasa en los medios de comunicación, donde se emite un mensaje pero no se cuenta con la opinión o el criterio con el cual recibe el mensaje el posible consumidor, las redes sociales como Facebook, nos permiten crear esta relación, donde el consumidor o seguidor de la marca, puede expresar su opinión, ya sea positiva, negativa o neutra. Para realizar un estudio de opiniones, existen técnicas de minería de datos adecuadas para la explotación de la información, que almacenada en los medios sociales, en las cuales aplicando una metodología de desarrollo de minería de datos, se determina cada uno de los procesos que nos llevan a los objetivos planteados.

2.2 Planteamiento del problema

Las empresas han tomado como estrategia de negocio, publicitar sus productos en las llamadas comunidades digitales, y a su vez los usuarios generan opiniones sobre los productos. Estas comunidades crecen diariamente y generan grandes volúmenes de información, por lo que a las empresas se les dificulta analizar las opiniones de sus seguidores, que podrían ser sus clientes. Es por esto que se ha visto la necesidad de crear

mecanismos mediante los cuales se pueda analizar las opiniones de los usuarios aplicando técnicas de minería de datos.

Le empresa Fideos Cayambe se encuentra lanzando su nueva campaña publicitaría a través de las redes sociales y su página web; donde requiere saber cuál es el nivel de relación de las personas con respecto a su producto, enfocada en un análisis de sentimiento de las opiniones de sus clientes en la red social Facebook. A partir de esto se espera definir nuevas estrategias para poder llegar de mejor manera a su segmento de clientes, lo afirma Armando Puyol jefe de la marca.

1.1 Justificación

Hoy en día las redes sociales se han transformado en un medio, donde las empresas pueden optar por aplicar sus estrategias de marketing, para lograr una mayor relación entre el consumidor o posible consumidor con respecto al producto ofertado, para esto es necesario saber qué opinan las personas al momento de tener una experiencia con la marca, si esta experiencia es positiva, darán paso a una relación constante en el medio social en el que se le permita expresar este sentimiento; esta investigación se basa en determinar la técnica más adecuada para definir este tipo de sentimientos, para lograr una clasificación de opiniones positivas y negativas, utilizando los conocimientos adquiridos para la aplicación de técnicas de minería de datos.

1.2 Objetivo General

Investigar las diferentes técnicas de minería de datos orientadas al análisis de sentimientos dentro de la marca Fideos Cayambe.

1.3 Objetivos específicos

- Elegir tres técnicas de análisis de sentimientos más relevantes relacionadas con el tema de investigación propuesto, en base al estado del arte.
- Recolectar información de una red social de la marca Fideos Cayambe.

- Extraer, transformar y analizar la información recolectada aplicando tres técnicas de análisis de sentimientos
- Establecer la mejor técnica de análisis de sentimiento mediante la cual se pueda llegar al mayor porcentaje de precisión para definir los sentimientos positivos y negativos de los consumidores de Fideos Cayambe a nivel nacional.

1.4 Hipótesis / proposición

La aplicación de las técnicas de minería de datos permitirá el análisis de sentimientos en las redes sociales sobre productos de la marca Fideos Cayambe.

2. CAPITULO II

MARCO REFERENCIAL

2.1 Redes sociales

Las Redes son formas de interacción social, definida como un intercambio dinámico entre personas, grupos e instituciones en contextos de complejidad. Un sistema abierto y en construcción permanente que involucra a conjuntos que se identifican en las mismas necesidades y problemáticas y que se organizan para potenciar sus recursos. (Ponce,2016).

Analizando este concepto, una red social se podría decir que es el intercambio de actividades o eventos usando una tecnología para identificar a las personas en una sociedad, con mismas preferencias, gustos, profesiones, objetivos o hobbies, en cualquier lugar del mundo dependiendo la interfaz con la que haga uso de esta herramienta. En la actualidad este intercambio de acciones se representa con imágenes, videos, fotografías, post y comentarios.

Típicamente en un medio social las personas comparten sus historias y sus experiencias con otros, de manera natural. Existen medios de comunicación como son los periódicos, canales de televisión y emisoras de radio, una característica que comparten los medios sociales y los Medios de comunicación de masas, es la capacidad de llegar a un público grande, por ejemplo, una publicación o un programa de TV de un medio tradicional pueden llegar a millones de personas en muchas partes del mundo, pero no podemos palpar cual es el criterio de esas personas al recibir el mensaje de

la publicación realizada, a diferencia de las redes sociales, donde sí, existe esta interacción de manera más directa.

Las redes sociales a su vez tienden a tener sus ventajas y desventajas, ya que como bien es cierto retomamos amistades abandonadas, nos enteramos de la vida de los demás, consultamos sitios de trabajo e interés, pero a la vez que hacemos todo, esto, existe lo opuesto, que es pasar horas de horas buscando navegando sin ningún interés en especial. Existen problemas por la actividad obsoleta en los trabajos por este tipo de distracciones, se ha investigado el uso de las redes sociales y su influencia, para poder determinar el buen uso de las mismas (Caldevilla ,2010); si las redes sociales tiene gran influencia en los seres humanos, las empresas como Fideos Cayambe deberían explotar esta ventaja para influenciar de manera positiva con sus productos a sus seguidores.

En España se ha decretado que los más famosos social media son: Facebook,Twiter,Linkedin,Tuenti,Google+,Pinterest, Instagram, Flickr, YouTube, SlideShare y WordPress, entre otros y el 82% de los navegadores están entre los 18 y 55 años utilizando estas redes sociales (mpc,2015). Si este gran porcentaje de personas es sólo en España nos podemos imaginar la gran cantidad de personas que se encuentran en la red social y que pueden interactuar, consultar, publicar e investigar en estas redes a nivel mundial.

Una red social es el acto de interacción entre dos usuarios en un social media, estos se relacionan por que comparten intereses, pensamientos o gustos; o simplemente se conocen físicamente y quieren mantener ese contacto. Visto de otra manera, es la retroalimentación entre dos o más usuarios los cuales pueden ser un nombre personal o en representación de una empresa, marca o evento lo que hace que exista una red comunicativa entre usuarios a través de los medios sociales.

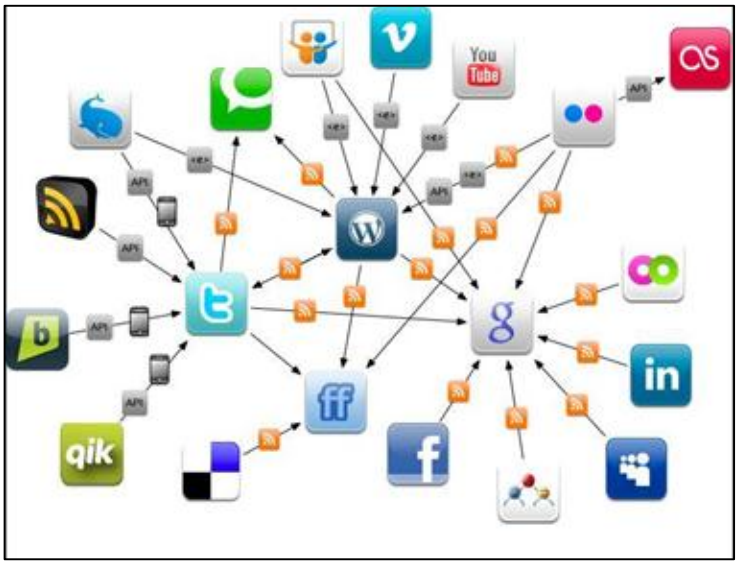


Fig. 1 Redes sociales

Fuente: (Helmod.2009)

3.2 Metodología CRISP-DM

Para la exploración de información que requiere este trabajo, se ha seleccionado la metodología CRISP-DM, la cual ha sido analizada y estructurada y consiste en 4 niveles de abstracción jerárquicos que van desde una visión macro hacia lo más detallado (Chapman,1999). De manera general el proceso de CRISP-DM esta detallado en 6 fases y cada fase tiene un grupo de tareas que describen ciertas acciones a realizarse para completar el proceso de cada una de estas fases.

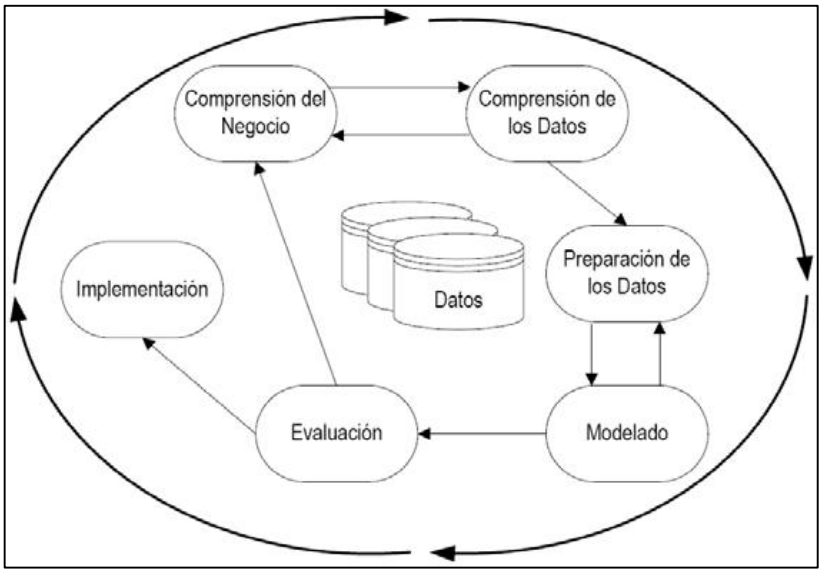


Fig. 2 Fases de la Metodología CRISP-DM

Fuente: (Britos, 2008).

Según el análisis desarrollado para la comparación de varias metodologías (Britos,2008) nos detalla cada una de ellas de la siguiente manera:

Comprensión del negocio

En esta etapa se requiere un análisis de cuáles son los objetivos que tiene el negocio, y qué es lo que quiere descubrir mediante este análisis de información, todo esto desde el punto de vista empresarial, donde se convertirá los objetivos de negocio en objetivos técnicos.

Compresión de los datos

En esta etapa se realiza la recolección de los datos a ser analizados, donde será el primer contacto con ellos para su exploración y manipulación, lo cual nos permitirá determinar su calidad y proyectar una base sólida para el procesamiento.

Preparación de los datos

Esta etapa consiste en seleccionar los datos que se van a utilizar, la agregación de nuevas variables derivadas de las existentes, limpieza de datos, estructuración e integración de los mismos.

Modelado

En esta etapa se va a seleccionar las técnicas más adecuadas para la necesidad del proyecto a desarrollar, tomando en cuenta el problema, la data requerida y el conocimiento de la técnica a aplicar.

Evaluación

Para esta etapa es necesario haber realizado las pruebas correspondientes al modelo aplicado para dar un criterio más preciso de las técnicas a evaluar, presentando de manera clara y entendible los resultados del procesamiento de la información para llegar a las conclusiones deseadas.

A continuación se presenta en la Figura 3 las etapas a desarrollar dentro de la metodología CRISP-DM para el análisis de información de la marca Fideos Cayambe.

FASE	TAREAS COMPONENTES	ACTIVIDADES ASOCIADAS
Comprensión del negocio	Determinar los objetivos del negocio	▪ Background ▪ Objetivos del negocio ▪ Criterios de éxito del negocio
	Evaluar la situación	▪ Inventarios de recursos ▪ Requisitos, supuestos y requerimientos ▪ Riesgos y contingencias ▪ Terminología ▪ Costos y beneficios
	Determinar objetivos del proyecto de Explotación de Información	▪ Las metas del Proyecto de Explotación de Información ▪ Criterios de éxito del Proyecto de Explotación de Información
	Realizar el Plan del Proyecto	▪ Plan de proyecto ▪ Valoración inicial de herramientas
Comprensión de los datos	Recolectar los datos iniciales	▪ Reporte de recolección de datos iniciales
	Descubrir datos	▪ Reporte de descripción de los datos
	Explorar los datos	▪ Reporte de exploración de datos
	Verificar la calidad de datos	▪ Reporte de calidad de datos
Preparación de los datos	Caracterizar el conjunto de datos	▪ Conjunto de Datos ▪ Descripción del Conjunto de Datos
	Seleccionar los datos	▪ Inclusión / exclusión de datos
	Limpiar los datos	▪ Reporte de calidad de datos limpios
	Estructurar los datos	▪ Derivación de atributos ▪ Generación de registros
	Integrar los datos	▪ Unificación de datos
	Caracterizar el formato de los datos	▪ Reporte de calidad de los datos
Modelado	Seleccionar una técnica de modelado	▪ La técnica modelada ▪ Supuestos del modelo
	Generar el plan de pruebas	▪ Plan de pruebas
	Construir el modelo	▪ Configuración de parámetros ▪ Modelo ▪ Descripción del modelo
	Evaluar el modelo	▪ Evaluar el modelo ▪ Revisación de la configuración de parámetros
Evaluación	Evaluar Resultado	▪ Valoración de resultados mineros con respecto al éxito del negocio ▪ Modelos aprobados
	Revisar	▪ Revisión del proceso
	Determinar próximos pasos	▪ Listar posibles acciones

Fig. 3 Etapas de la metodología CRISP-DM

Fuente:(Britos,2008).

3.3 Procesamiento de Lenguaje Natural(PLN)

Para esta investigación nos hemos basado en realizar un análisis de sentimiento donde se realizará lo bien llamado procesamiento de lenguaje natural que es la combinación de la ciencias de la computación, inteligencia artificial y lingüística que estudia la interacción del leguaje humano con las computadoras.

En el artículo Procesamiento de Lenguaje Natural Cortez et. al (2015). Nos dicen que la arquitectura de un sistema de PLN se sustenta en una definición del lenguaje natural por niveles y estos son:

Nivel Fonológico: trata de cómo las palabras se relacionan con los sonidos que representan.

Nivel Morfológico: trata de cómo las palabras se construyen a partir de unas unidades de significado más pequeñas llamadas morfemas.

Nivel Sintáctico: trata de cómo las palabras pueden unirse para formar oraciones, fijando el papel estructural que cada palabra juega en la oración y que sintagmas son parte de otros sintagmas.

Nivel Semántico: trata del significado de las palabras y de cómo los significados se unen para dar significado a una oración, también se refiere al significado independiente del contexto, es decir de la oración aislada.

Nivel Pragmático: trata de cómo las oraciones se usan en distintas situaciones y de cómo el uso afecta al significado de las oraciones. Se reconoce un subnivel recursivo: discursivo, que trata de cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.

En el artículo, El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines Vallez et.al (2007), nos dice que a nivel morfológico de una misma palabra puede adoptar diferentes roles morfo-sintácticos en función del contexto en el que aparece, ocasionando problemas de ambigüedad, por ejemplo:

Deja la comida que sobre, sobre la mesa de la cocina, dijo llevando el sobre en la mano.

La palabra sobre es ambigua morfológicamente ya que puede ser un sustantivo masculino singular, una preposición, y también la primera o tercera persona del presente de subjuntivo del verbo sobrar.

María vio a un niño con un telescopio en la ventana.

La interpretación de la dependencia de los dos sintagmas preposicionales, con un telescopio y en la ventana, otorga diferentes significados a la frase:

María vio a un niño que estaba en la ventana y que tenía un telescopio.

María estaba en la ventana, desde donde vio a un niño que tenía un telescopio

María estaba en la ventana, desde donde miraba con un telescopio, y vio a un niño.

A nivel semántico, donde se estudia el significado de una palabra y el de una frase a partir de los significados de cada una de las palabras que la componen. La ambigüedad se produce porque una palabra puede tener uno o varios sentidos, es el caso conocido como polisemia.

Luís dejó el periódico en el banco.

El término banco puede tener dos significados en esta frase.

Entidad bancaria y asiento. La interpretación de esa frase va más allá del análisis de los componentes que forman la frase, se realiza a partir del contexto en que es formulada.

Y también hay que tener en cuenta la variación léxica que hace referencia a la posibilidad de utilizar términos distintos a la hora de representar un mismo significado, es decir el fenómeno conocido como sinonimia.

Coche / Vehículo / Automóvil.

A nivel pragmático, basado en la relación del lenguaje con el contexto en que es utilizado, en muchos casos no puede realizarse una interpretación literal y automatizada de los términos utilizados. En determinadas circunstancias, el sentido de las palabras que forman una frase tiene que

interpretarse a un nivel superior recurriendo al contexto en que es formulada la frase.

Se moría de risa.

En esta frase no puede interpretarse literalmente el verbo morirse si no que debe entenderse en un sentido figurado.

3.4 Minería de datos

La minería de datos es un proceso que nos permite descubrir que los datos son o no procesables ya que hace referencia a grandes cantidades de datos, dado esto se generan patrones y tendencias, se pueden definir como modelo de minería de datos, donde se agrupan técnicas para este análisis de información. (Perez,2008).

Las herramientas como DSS(Decision Support Systems), OLAP(On-Line Analytical Processing) y los sistemas de informes facilitan el acceso a la información para que el análisis sea más efectivo, son instrumentos de apoyo para la minería de datos, y como tal necesitan de un repositorio o Data Warehouse para su almacenamiento, donde los datos se encuentren debidamente organizados para poder analizarlos y describirlos en los resultados que queremos obtener. Por tanto para lograr un KDD (extracción de conocimiento a partir de datos) es necesario saber que la minería de datos forma parte de este proceso y a su vez tiene que cumplir con las etapas antes descritas en la metodología CRISP-DM para conseguir los objetivos planteados. Entonces el proceso viene definido de esta manera SELECCIÓN ->EXPLORACION -> LIMPIEZA -> TRANSFORMACION -> MINERIA DE DATOS -> EVALUACION -> DIFUSION. (Perez,2008).

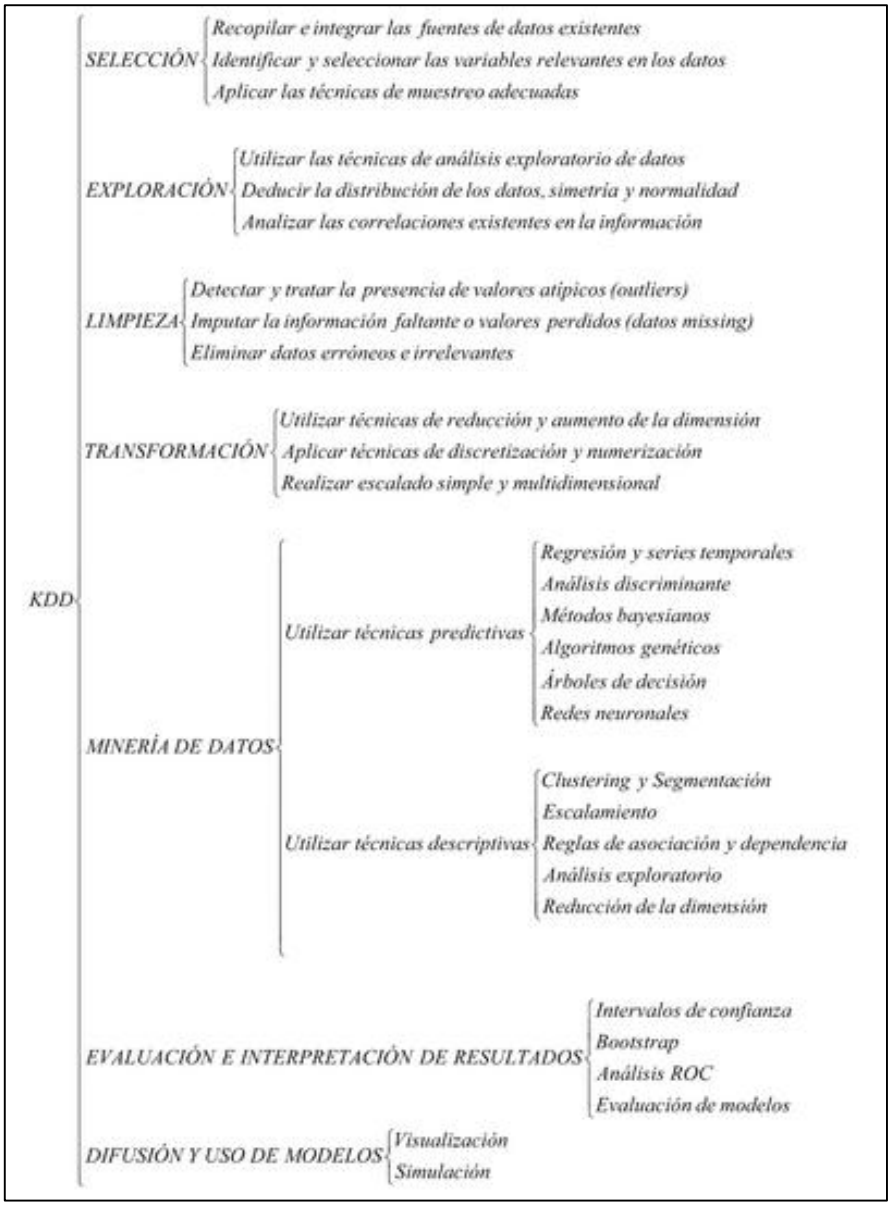


Fig. 4 Etapas KDD

Fuente: (Perez,2008).

Según la Fig.4 se pude observar la clasificación de las técnicas de minería de datos en predictivas y descriptivas.

Las técnicas predictivas especifican el modelo para datos en base a un conocimiento previo, este modelo consta de fases como: identificación

objetiva, donde a partir de los datos se aplican reglas que sean las más adecuadas para elegir el mejor modelo, estimación, esta fase es un proceso de cálculo de parámetros del modelo seleccionado, diagnosis, es el proceso de validación del modelo estimado y la etapa de predicción donde se utiliza el modelo seleccionado, validado y estimado para predecir los valores en cuanto a las variables dependientes .

En este tipo de técnicas se puede mencionar a todos los tipos de regresión: series temporales, análisis de la varianza y covarianza, análisis discriminante, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas. Tanto los árboles de decisión, como las redes neuronales y el análisis discriminante son a su vez técnicas de clasificación, donde se pueden extraer perfiles de comportamiento o clases.

En las técnicas descriptivas no se da importancia a las variables como en las técnicas antes descritas, estas técnicas se basan en el reconocimiento de patrones, las técnicas que pertenecen a este grupo son: clustering y segmentación, las técnicas de asociación y dependencia, las de análisis exploratorio de datos y las de reducción de la dimensión y de escalamiento multidimensional (Perez,2008).

Árboles de decisión

Un árbol de decisión es la imposición de una regla sobre el comportamiento mayoritario de un conjunto de datos, y estas están dadas por las condicionantes IF y THEN.

La precisión de las reglas se dan por la matriz de confusión que generan los árboles de decisión, donde nos muestra el porcentaje de aciertos y el porcentaje de casos que no cumplen con la condicionante.

El valor esperado representa el promedio a obtener a largo plazo bajo el principio del muestreo repetido. Se asume que hay una medida de probabilidad P(X) que permite establecer varios escenarios cuyo resultado es caracterizado por una realización X que puede tomar k posibles valores

x1,...,xk. El valor esperadoE(X) es:

$$E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

En las evaluaciones se utiliza el razonamiento presente, en la programación dinámica "backward induction" (inducción hacia atrás), es decir: comenzando en un nodo final se regresa al nodo inicial.

Clustering

Consiste en la división de datos en grupos de objetos que se asemejan, usando la información que dan las variables que pertenecen a cada objeto, se mide la similitud entre los mismos y de esta manera se agrupan por cada clase.

Simple K-Means

En este algoritmo se debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar. Los pasos básicos para aplicar el algoritmo son muy simples. Primeramente se determina la cantidad de clusters en los que se quiere agrupar la información, en este caso las simulaciones. Luego se asume de forma aleatoria los centros por cada clusters. Una vez encontrados los primeros centroides, el algoritmo hará los tres pasos siguientes:

Determina las coordenadas del centroide.

Determina la distancia de cada objeto a los centroides.

Agrupa los objetos basados en la menor distancia. Finalmente quedarán agrupados por clusters, y los grupos de simulaciones según la cantidad de clusters se haya definido (Ecured.cu,2017).

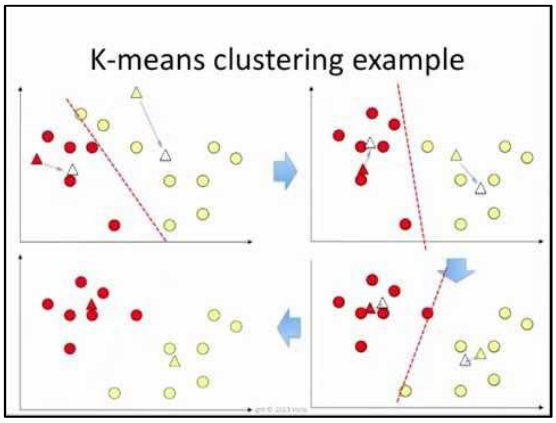


Fig. 5 Ejemplo K-means.

Dado un conjunto de observaciones (x1, x2, …, xn), donde cada observación es un vector real de d dimensiones, k-means construye una partición de las observaciones en k conjuntos (k ≤ n) a fin de minimizar la suma de los cuadrados dentro de cada grupo (WCSS):

S = {S1, S2, …, Sk}

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Donde μi es la media de puntos en Sí. Mahajan et al (2009).

Regresión lineal

La regresión lineal simple hace referencia a la cantidad de cambio que experimenta una variable dependiente (Y), en relación al cambio de una unidad de una variable independiente (X), la regresión es un concepto estadístico vinculado con el concepto de correlación, mientras la regresión estudia la relación de dos variables dependientes la correlación estudia la estreches de la relación de estas dos variables, una depende de la otra. (Little ,1978).

El modelo de regresión lineal viene dado de la siguiente manera:

$$\mathbf{Y_i = B_0 + B_1 X_i + e_i}.$$

Dónde:

Y_1: es la variable dependiente .

B_0: es la ordenada en el origen o bien el intercepto.

B_1: es la pendiente de la recta de regresión

e_i: es el termino del error , es decir la diferencia entre los valores predichos por la regresión y los valores reales.

El coeficiente de correlación es el que mide el grado de asociación entre las variables X e Y, mientras más alto sea este valor nos indicara que mayor es la dependencia entre una variable y otra. (Pedroza, 2006).

3.5 Estado del Arte

Según los cambios y avances en las tecnologías de la información y la participación de las personas en los medios sociales a nivel mundial, en un estudio realizado por la asociación española, que representa al sector de la publicidad y la comunicación digital en España denominada Interactive Adverstising Bureau(IAB) ,en lo que va del año 2017, tiene como dato que el 86% de los internautas de 16 a 65 años de edad, que representan a 19 millones de usuarios en España utilizan redes sociales, dado esto se puede mentalizar la cantidad de información que es ingresada en las redes sociales, la cual con un correcto procesamiento y análisis sirve para resolver problemas y crear nuevas oportunidades en el ámbito tecnológico, gubernamental, médico o cualquier área en la que se pueda aplicar las técnicas necesarias de minería de datos , dependiendo de los objetivos planteados. (IAB,2017).

En empresas importantes como IBM, se realizó un análisis de sentimiento enfocado a identificar tendencias en Twitter y en América Latina relacionadas con la educación, futuro y salud. (IBM,2014). Como resultado se obtuvo que el 25% de los “twitteros” proponen ideas constructivas, mientras un 40% reacciona de manera inmediata a las noticias de los medios. En lo cotidiano, el 41% de los latinos son optimistas, datos como

estos, nos permiten saber qué está en la mente de las personas detrás de un computador, celular o interfaz, que le permite plasmar su preferencia, gusto o pensamiento.

Diez países fueron parte de este experimento: Argentina, Bolivia, Chile, Colombia, Ecuador, México, Paraguay, Perú, Uruguay y Venezuela destacando que los latinos son altamente reactivos ante las noticias de los medios de comunicación y que se preocupan por su futuro, la educación y la salud pública. De los 10 países, 7 de ellos tienen un porcentaje superior en torno a la discusión de la educación en sus países, 6 de ellos discuten su futuro, y 5 el de salud pública. Los argentinos mencionan los conceptos de "gratitud" y "oportunidad" en las categorías de educación y salud.

En ecuador 2,8 millones de personas, declara usar redes sociales a través de su teléfono inteligente, según datos del Instituto Nacional Estadísticas y Censos (Inec). (Metro ,2016). Este tipo de información nos da una idea del universo que podría ser parte de la investigación planteada, puesto que en estos millones de personas están incluidos propietarios de grandes empresas que requieren obtener predicciones y análisis de opinión en cuanto a su producto.

Por la gran cantidad de datos que se almacenan, con gran variedad de tipos y su crecimiento exponencial, se da la necesidad de tener mecanismos que nos permitan manipular y obtener conocimiento en base a esta data, la minería de datos nos brinda la oportunidad con las metodologías diseñadas.

Rodríguez et al (2016) ha escogido la metodología CRISP-DM para realizar la adecuación a sus problemas no supervisados tipo atributo-valor, se logra un nivel de especificación más profundo en cada una de las 6 fases propuestas y así se ahorra tiempo. Los métodos de minería de datos llevan asociados una serie de mecanismos (estimación de errores, matrices de confusión, matrices de pérdida, curvas de esfuerzo y aprendizaje, análisis sensitivo de entradas...) que permiten realizar una mejor validación empírica

de los modelos y un análisis de resultados más completo y fiable que el que ofrece el enfoque clásico.

Desde otro punto de vista para otros investigadores como Arcila et. al (2016) en su artículo acerca de las técnicas para análisis de textos a gran escala dentro del periodismo, nos dice que la capacidad de procesamiento aumenta y el volumen de información es infinita, las nuevas lógicas implican la necesidad de construir grandes centros de análisis de big data para facilitar el desarrollo de proyectos. Dentro de este artículo observamos que las aplicaciones de aprendizaje supervisado requieren algoritmos especializados que detecten patrones en los datos. Estos algoritmos pueden implementarse en lenguajes de programación como Python, pero al igual que en el análisis de contenido automatizado, si se aplican sobre grandes cantidades de datos requieren plataformas distribuidas para el procesamiento en paralelo. Para superar las dificultades que implica el desarrollo de código y el despliegue de centros de cómputo en la nube, ha prosperado una serie de servicios comerciales que permiten el aprendizaje automático de manera mucho más sencilla.

Existe una plataforma AWS que incluye un módulo llamado Amazon Machine Learning (AML) que incorpora tanto asistentes como software de visualización, o los servicios de la empresa Databrick que basan sus servicios de computación en la nube exclusivamente en Spark. Para el aprendizaje supervisado, modelos como el de predicción basado en máquinas de vectores soporte (SVM) que pasan nuestros datos a un espacio multidimensional, permiten crear algoritmos potentes a partir de datos existentes (un ejemplo: conjuntos de noticias ya separadas por tema) para crear patrones que permitan categorizar, automáticamente, nuevos conjuntos de textos. Arcila et. al (2016).

En otras investigaciones se ha utilizado diferentes motores de análisis para la minería de texto; no se centra sólo en la estadística descriptiva. Por ejemplo, se realizó un análisis de los sentimientos, de KOL(Key Opinión Leaders), la creación de comunidades entre los temas y personas, así como

también se ha desarrollado la capacidad de monitorear el comportamiento social en torno a tópicos.

Diferentes autores han aplicado diversas técnicas de agrupamiento para detectar comunidades en la red social aplicando agrupación jerárquica, esta técnica es una combinación de muchas técnicas utilizadas para agrupar nodos en la red que revelan la fuerza de los grupos individuales que luego se utiliza para distribuir.

El análisis de sentimientos se puede utilizar en la estimación del impacto de una publicación como lo ha utilizado el autor Hernández et. al (2014). En el Análisis de sentimiento aplicado a referencias Bibliográficas usando el contexto de las citas para detectar la intensificación y sentimientos de un autor al hacer una referencia o cita bibliográfica dentro de un artículo científico. Con esta información se puede mejorar el análisis bibliométrico de las referencias.

Según el artículo, Evaluación de algoritmos de clasificación supervisada para minado de opinión en twitter (Estevez, 2015). Nos dice que entre las diferentes técnicas para reducción de dimensiones, la técnica de clustering con su algoritmo K-meas entre otras, que realizan este mismo trabajo, se han destacado como las mejores; para la clasificación los algoritmos con un mayor porcentaje de acierto fueron árboles de decisión y Linear SVM, los mismos que determinaron si un twitt es positivo, neutro o negativo.

Para determinar un sentimiento de manera automatizada en combinación de algoritmos de clasificación, se requiere evaluar una herramienta con un alto grado de precisión, puesto que se tiene que tomar en cuenta características como: el idioma, la extensión de la frase que va a ser analizada, ortografía, que se asume tiene que ser correcta, y el criterio de clasificación. la herramienta WordNet-Affect Hierach se ha usado para determinar si una frase contiene sentimientos de: tristeza, enojo o alegría , de esta manera lo ha explicado un grupo de Mexicanos en su artículo, Arquitectura Web, para análisis de sentimientos en Facebook con enfoque

semántico, donde llega a un 63% de precisión realizando este análisis, aplicando el algoritmo de clasificación Naive Bayes. Acevedo et. Al (2014).

Es importante contar con una herramienta que nos permita realizar de manera amigable y fácil el procesamiento de información, en este caso la aplicación de algoritmos, brindándonos los resultados de manera concreta y entendible, WEKA ha sido usada como herramienta base para experimentos según el artículo Estudio de las categorías LIWC para el análisis de sentimientos en español , donde el algoritmo SMO y J48 son los de mayor precisión en el momento de clasificar los sentimientos positivos, negativos y neutros. Salas et. Al (2011).

4 CAPITULO III

MARCO METODOLÓGICO

Aplicación de la Metodología CRISP-DM

En la investigación que se va a realizar para el análisis las técnicas de minería de datos aplicadas a la información de la marca Fideos Cayambe en las redes sociales, será de tipo experimental. En este punto del proyecto se aplicará cada una de las fases de la metodología y se detallará su desarrollo.

4.1 Comprensión del negocio

El nombre de fideos Cayambe, evidencian la trayectoria del producto y el orgullo de su origen, como una marca de alto reconocimiento en su sector, ya que alimenta desde varios años a las familias ecuatorianas de la zona y de todo el país.

Hace 6 años, se decidió darle un impulso adicional a la producción de pasta con la adición de un molino semolero, que produce sémola de Trigo Durum, que también se lo conoce como trigo para pasta por su dureza, alto contenido proteínico, por su sabor y cualidades de cocción. Así se aseguró una materia prima de excelente calidad para el producto final:

Fideo Cayambe con sémola de la mejor calidad.

El objetivo de tener presencia en las Redes Sociales, como Facebook, es tener una cercanía con los consumidores de la marca y lograr una afinidad en dónde los clientes y consumidores puedan compartir sus vivencias y recetas con la marca y a la vez que se enteren de lo que la misma aporta a este segmento de comunidad virtual generando un nivel de engagemet (relación) entre marca y consumidor.

Objetivos del negocio

Es importante tener una visión a largo plazo, por ende el objetivo principal de Fideos Cayambe es incrementar el posicionamiento de la marca en el marketshare (cuota de mercado) de hogares ecuatorianos y conseguir un 15% de cuota de mercado adicional cada año.

Evaluación del negocio

Fideos Cayambe es una de las marcas de pastas de más trayectoria en el Ecuador, sus productos son bastante diversos, entre ellos tenemos tallarín, broca, lazo, macarrón, concha, peques dinosaurios, entre otros. Sin embargo el producto que el consumidor ecuatoriano está más acostumbrado a adquirir es el tradicional tallarín, en presentaciones de 200 y 400 gramos el paquete.

Según informes de la empresa Analitics, por medio de impactos y participaciones en incrementos ha llegado al 16% luego de haber empezado con un 2%.

Objetivos de la minería de datos

Verificar la existencia de la relación entre el consumidor y el producto de la marca.

Determinar la participación del consumidor en la red social Facebook por periodos de tiempo.

Determinar las opiniones existentes de los consumidores en relación a la marca.

Realizar el plan de proyecto

Etapa 1: Análisis de la estructura de los datos y la información de la base de datos

Etapa 2: Exploración y calidad de datos.

Etapa 3: Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la minería de datos sobre ellos.

Etapa 4: Elección de las técnicas de modelado y ejecución de las mismas sobre los datos.

Etapa 5: Análisis de los resultados obtenidos en la etapa anterior, si fuera necesario repetir la etapa 4.

Etapa 6: Evaluación de los resultados obtenidos en función de los objetivos de negocio.

Etapa 7: Conclusiones.

Evaluación inicial de las herramientas y técnicas

Para el desarrollo de esta etapa se ha seleccionado Qlick View, ya que cuenta con los filtros necesarios para el análisis de datos, exploración y validación de los mismos.

Por otro lado para la aplicación de los algoritmos dentro de las técnicas de análisis que serán evaluadas, será utilizada la herramienta WEKA de fácil accesibilidad, que cuenta con los algoritmos que vamos aplicar en la investigación y muestra los resultados de una manera comprensible y gráfica.

4.2 Comprensión de los datos

En esta etapa de la metodología CRISP–DM se realiza la recolección de datos, que será la fuente en la que se aplicará nuestro análisis, este punto es donde de manera inicial se tiene que crear una relación con los datos, explorarlos y medir su calidad, para así tener el conocimiento concreto de la información a utilizar.

Comprensión de los datos

La data que será procesada para esta investigación es la referente a la página Fideos Cayambe en la red social Facebook. Utilizando la herramienta Facepager se ha logrado descargar todos los post(publicaciones), comentarios, likes y datos de personas que han participado de cada una de las publicaciones realizadas en esta página en los últimos dos años.

Con la url de la página de Facebook que se desea extraer los datos, en este caso de la marca Fideos Cayambe como se puede ver en la Fig. 6:

Fig. 6 Url de Fideos Cayambe en Facebook.

Ingresamos en la página https://findmyfbid.com/ donde nos brindará un identificador único para la extracción de datos.

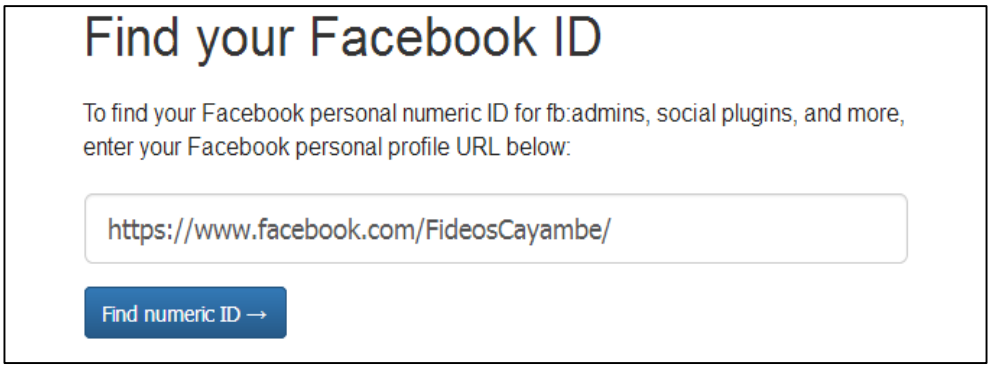


Fig. 7 Buscador de Facebook ID.

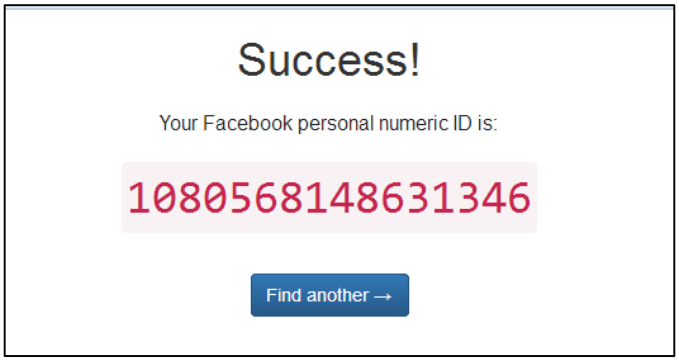


Fig. 8 Facebook ID.

Con este identificador numérico procedemos a utilizarlo en nuestra herramienta Facepager.

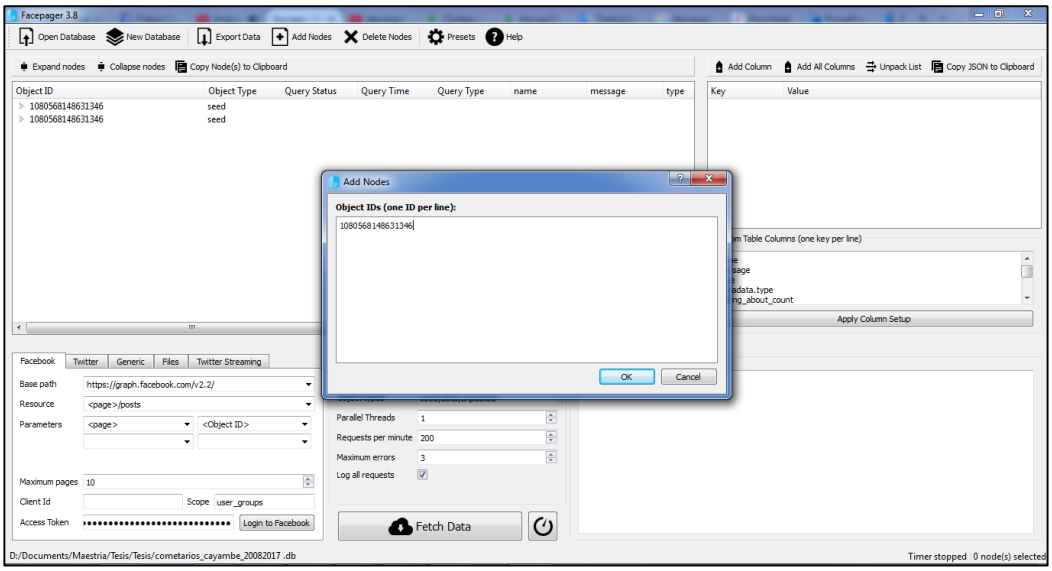


Fig. 9 Facepager Extracción de publicaciones.

Una vez ingresado el id numérico obtenemos los post realizados hasta agosto del 2017, seguido de esto se tiene que extraer también los comentarios de cada uno de estas publicaciones para que nuestra data sea más completa.

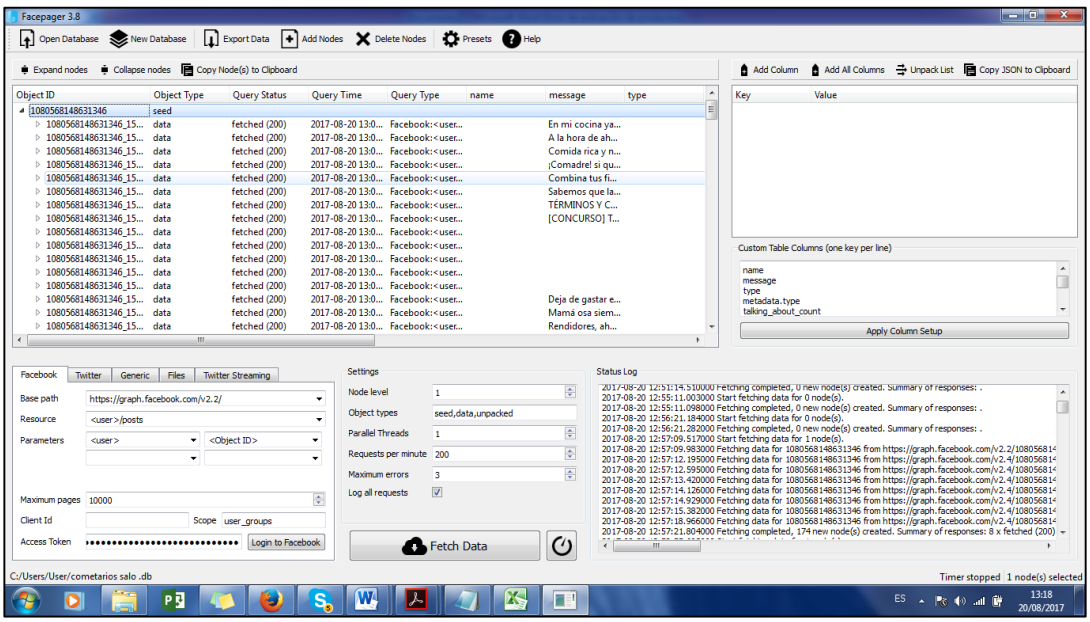


Fig. 10 Facepager Extracción de comentarios.

Los datos acerca de las publicaciones extraídas, se muestran en una tabla de la siguiente manera:

Tabla 1:

Descripción de campos obtenidos por Facepager

COLUMNAS	DESCRIPCION
ObjectID	Es la llave única que identifica el objeto se post o comentario
ObjectType	Tipo de objeto
QueryStatus	Estado del query ejecutado
QueryTime	Fecha de descarga de la data
Query Type	Tipo de query en este caso puede ser post o comentario,potos,viedeos, etc
Message	Mensaje
CreatedTime	Fecha de creación del evento

Los datos relevantes para nuestro análisis serán:

- QueryStatus
- QueryType
- Message
- CreatedTime

Descripción de los datos

Una vez que tenemos los datos vamos a describir su volumen, es decir, el número de registros y campos por registro, identificación y descripción de cada uno.

Nuestra base consta de 12822 registros y los campos consolidados correspondientes son los siguientes:

Tabla 2:

Descripción de la tabla base

CAMPO	TIPO	DESCRIPCION
Id	Numérico	Identificación única del campo
Id_hijo	Numérico	Identificación segundo nivel
nivel	Numérico	Nivel del registro
Tipo_objeto	Cadena	Describe si es comentario, post, comentario del pos, comentario del video
Persona	Cadena	Persona que realiza la acción
Sexo	Char	F = femenino M = masculino
ubicacion	Char	Ubicación de la persona que realiza la acción
mensaje	Varchar2	Mensaje del post o del comentario
Fecha_descarga	Date	Fecha de descarga de la data
Fecha_creacion	Date	Fecha de creación de la acción

Explorar los datos

Una vez que se han descrito los datos, se procede a explorarlos, esto es aplicar pruebas estadísticas que revelarán propiedades de los datos. Este informe sirve principalmente para determinar la consistencia y completitud de los datos. Mediante la herramienta Qlick View aplicamos este proceso.

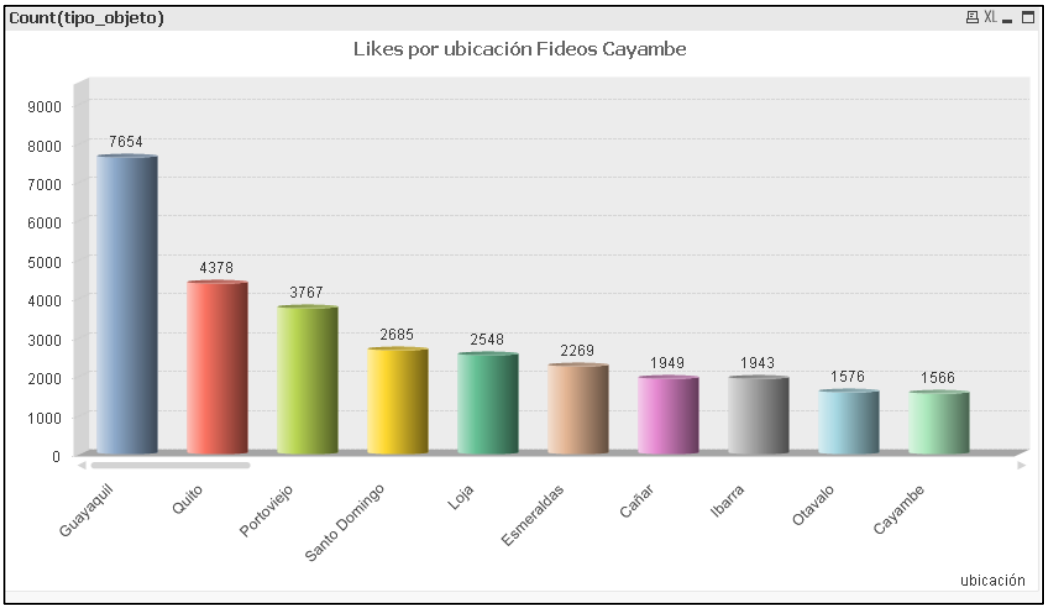


Fig. 11 Likes por ubicación Fideos Cayambe.

Como podemos ver en la Fig. 11, tenemos los likes representados en cada una de las ciudades del Ecuador en cuanto a los post que publica Fideos Cayambe en su red social.

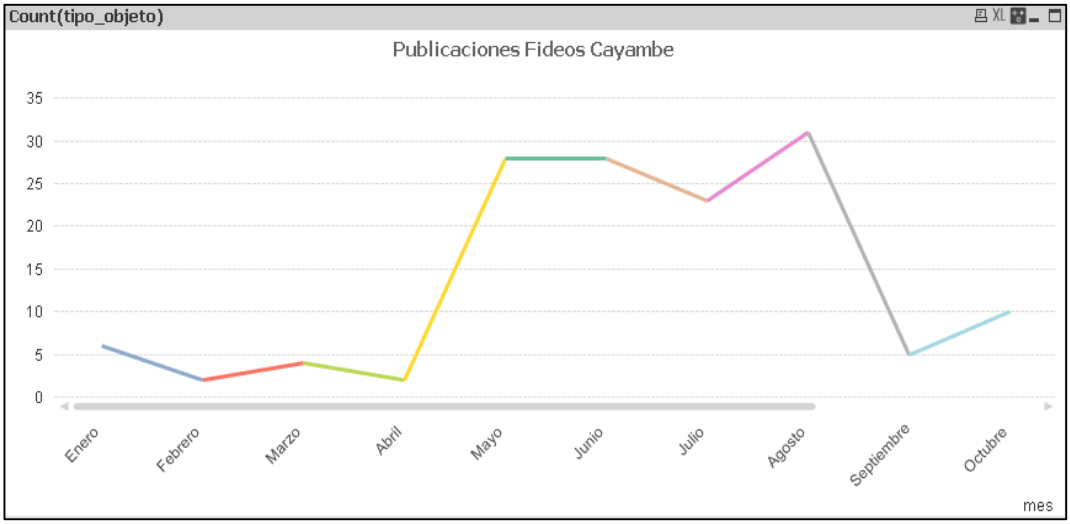


Fig. 12 Post publicados mensualmente 2017.

En la fig. 12 se muestra los post que son publicados mensualmente y se puede observar como existe un incremento de publicaciones en los

meses: mayo, junio y agosto, que como estrategia de marketing existen promociones y concursos, son meses en los que la mayor parte de personas interactúan de manera frecuente en redes sociales.

Verificar la calidad de los datos

Después de hacer la exploración de los datos iniciales nos encontramos con post que no tenían ninguna fecha de publicación o no tenían mensaje, por tanto son datos que no aportan a la investigación realizada, entonces se tiene que excluir de nuestra base.

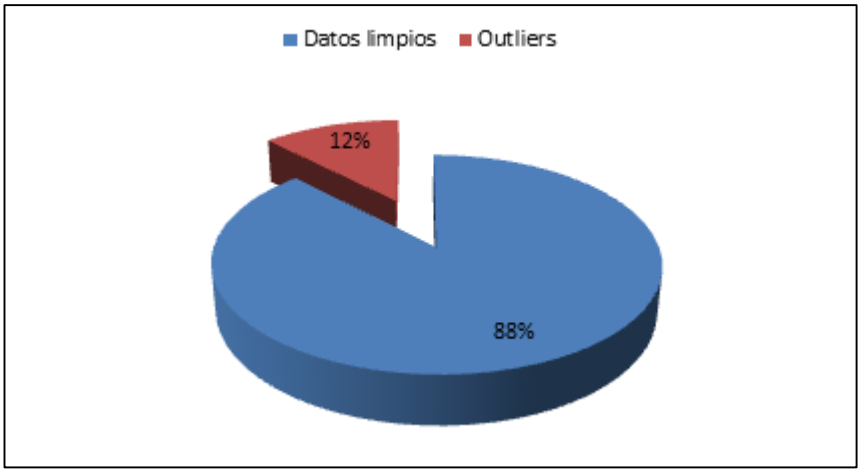


Fig. 13 Datos Fideos Cayambe Comentarios.

En los comentarios, que será nuestra data para realizar el análisis de sentimiento, se encontró un 12% de Outliers que serán discriminados y por tanto se trabajará con el 88% de los comentarios existentes.

4.3 Preparación de los datos

La fase de preparación de datos consiste en adecuarlos a las técnicas de minería de datos que se van aplicar en ellos; para esto se debe seleccionar el subconjunto que se va a ser procesado, realizar la adecuada limpieza y añadir campos relacionados con los existentes, de manera que ayuden al desarrollo de la técnica utilizada, así como también hay que darles el formato necesario para la utilización de la herramienta seleccionada.

Selección de los datos

Los registros a ser usados para el análisis de sentimientos serán los de la base de comentarios de Fideos Cayambe que se encuentra en un archivo plano .csv, específicamente creada para este fin, se ha depurado este archivo estableciendo los campos necesarios para nuestros objetivos y se ha descartado los campos que no serán de utilidad en este proceso.

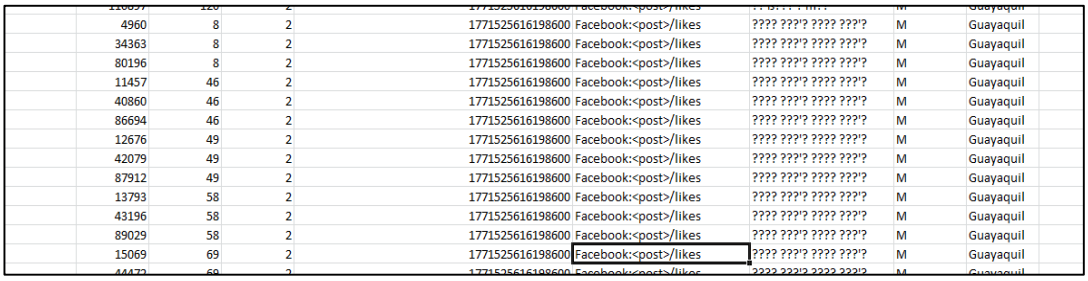
Tabla 3:

Descripción campos para análisis

CAMPO	TIPO	DESCRIPCION
Id	Numérico	Identificación única del campo
Id_hijo	Numérico	Identificación segundo nivel
nivel	Numérico	Nivel del registro
Tipo_objeto	Cadena	Describe si es comentario, post, comentario del pos, comentario del video
Persona	Cadena	Persona que realiza la acción
Sexo	Char	F = femenino M = masculine
ubicacion	Char	Ubicación de la persona que realiza la acción
mensaje	Varchar2	Mensaje del post o del comentario
Fecha_creacion	Date	Fecha de creación de la acción

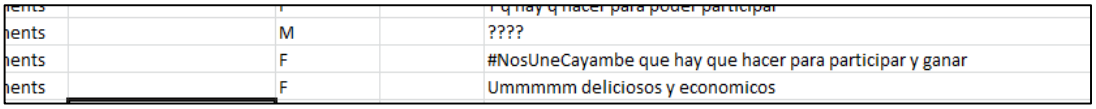
Limpieza de datos

Para realizar un mejor análisis de datos se requiere tener los datos de manera estructurada y limpia, por tanto se procede a eliminar los registros que contengan caracteres especiales, campos vacíos o comentarios sin sentido alguno.



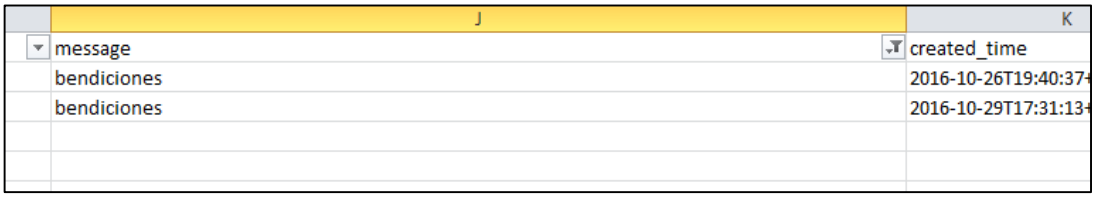
	4960	8	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	34363	8	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	80196	8	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	11457	46	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	40860	46	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	86694	46	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	12676	49	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	42079	49	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	87912	49	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	13793	58	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	43196	58	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	89029	58	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil
	15069	69	2	1771525616198600	Facebook:<post>/likes	???? ???'? ???? ???'?	M	Guayaquil

Fig. 14 Data basura.



ients		M		????
ients		F		#NosUneCayambe que hay que hacer para participar y ganar
ients		F		Ummmmm deliciosos y economicos

Fig. 15 Data inconsistente caracteres especiales.



	J	K
	message	created_time
	bendiciones	2016-10-26T19:40:37+
	bendiciones	2016-10-29T17:31:13+

Fig. 16 Data inconsistente sin mensaje.

Integrar los datos

Atributos derivados

Los datos a ser evaluados, en este caso, serán los comentarios emitidos por las personas, para esto se ha procedido a la agregación de campos, donde se evalúan los criterios para pasar de manera cualitativa a cuantitativa la opinión, determinando si existen comentarios con las siguientes palabras o sinónimos: económico, delicioso, aceptado, rendidor, nutritivo, feo, caro; analizando cada uno de los comentarios para poder establecer 1 cuando cumple y 0 cuando no lo cumple.

Formato de los datos

En esta fase de formato de datos, se ha realizado la descomposición de ciertos campos, como: fecha de creación, la cual se divide en mes, año y hora de publicación, para mejor uso de estos criterios.

Tabla 4:

Descripción de tipos de datos

CAMPO	TIPO	DESCRIPCION
Id	Numérico	Identificación única del campo
Id_hijo	Numérico	Identificación segundo nivel
Tipo_objeto	Cadena	Describe si es comentario, post, comentario del pos, comentario del video
Sexo	Char	F = femenino M = masculino
Ubicacion	Varchar2	Ubicación de la persona que realiza la acción
Mensaje	Varchar2	Mensaje del post o del comentario
Fecha_creacion	Date	Fecha de creación de la acción
Mes	Numérico	Mes representado en números
Hora_creacion	Numérico	Hora de creación
Económico	Numérico	1 si cumple 0 si no cumple
Delicioso	Numérico	1 si cumple 0 si no cumple
Aceptado	Numérico	1 si cumple 0 si no cumple
Rendidor	Numérico	1 si cumple 0 si no cumple
Nutritivo	Numérico	1 si cumple 0 si no cumple
Feo	Numérico	1 si cumple 0 si no cumple
Caro	Numérico	1 si cumple 0 si no cumple

4.4 Modelado

En esta fase se evaluará las técnicas para los criterios antes descritos, y realizar la aplicación de algoritmos para determinar la técnica más apropiada para el procesamiento de la data de la marca Fideos Cayambe, y cumplir de esta manera los objetivos propuestos.

Selección de la técnica del modelado

Para la selección de las técnicas a utilizar se ha tomado de las experiencias descritas en el estado del arte, donde queremos realizar la comparación de las técnicas: árboles de decisión, Clustering y regresión lineal mediante la herramienta Weka.

Construcción del modelo

Árboles de clasificación

Ahora se procede a ejecutar el algoritmo J48 sobre los datos de Fideos Cayambe, para evaluar los resultados que nos permitirán visualizar la respuesta hacia nuestros objetivos, cabe recalcar, que la variable dependiente a usarse, en este caso, es categórica.

Se procesa el algoritmo J48 para el resultado el criterio “Positivo”.

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

aceptado_binarized = 0
|   delicioso_binarized = 0
|   |   economico_binarized = 0: NO (198.0/1.0)
|   |   economico_binarized = 1: SI  (4.0)
|   delicioso_binarized = 1: SI  (28.0)
aceptado_binarized = 1: SI  (968.0)

Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1197               99.9165 %
Incorrectly Classified Instances         1                0.0835 %
Kappa statistic                          0.997
Mean absolute error                      0.0017
Root mean squared error                  0.029
Relative absolute error                  0.6043 %
```

Fig. 17 Resultado para el criterio Positivo I.

En esta gráfica, podemos apreciar algunos parámetros calculados, como: la clasificación correcta, que nos da un 99.91% correctas, y un 0.083 % incorrectas, lo que nos dice que se aplicó a casi todos los datos que han

sido cargados.

```
Root relative squared error              7.8159 %
Total Number of Instances             1198
Ignored Class Unknown Instances                48

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0,999    0,000    1,000      0,999    1,000      0,997    0,922     0,963     SI
                1,000    0,001    0,995      1,000    0,997      0,997    0,994     0,941     NO
Weighted Avg.   0,999    0,000    0,999      0,999    0,999      0,997    0,934     0,960

=== Confusion Matrix ===

   a   b   <-- classified as
1000   1 |   a = SI
   0 197 |   b = NO
```

Fig. 18 Resultado para el criterio Positivo II.

Se han ignorado 48 instancias, y como se muestra en la matriz de confusión, tenemos en la variable a = SI que son 1000 registros y 197 para la variable b =NO, por tanto, la mayoría de los clientes de Fideos Cayambe opinan, en Facebook, de manera positiva.

Generamos la visualización del árbol, donde se puede observar los factores, que son determinantes, para el criterio “Positivo”.

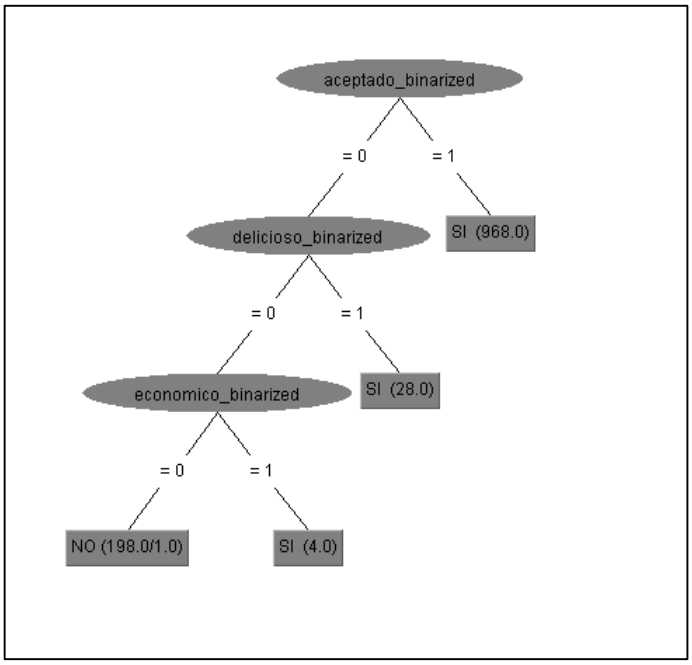


Fig. 19 Resultado árboles para el criterio Positivo II.

En este tipo de visualización, podemos ver, que el factor más determinante es aceptado. En menor proporción, las personas opinan que es delicioso y económico, en general: tienen una opinión “Positivo” de los clientes, en la red social Facebook, en la página comercial de Fideos Cayambe. Estas personas han realizado comentarios en los que expresan que su sabor les parece delicioso. Otro factor importante, para ellas, es, el precio de venta, aunque el factor de mayor importancia, para que este criterio, sea “Positivo”, es que tiene que ser aceptada la marca Fideos Cayambe. Dentro de estos comentarios, se ha encontrado la aceptación con la marca para el grupo de 968 personas.

Se procesa el algoritmo J48 para el resultado el criterio “Negativo”.

```
J48 pruned tree
------------------

feo_binarized = 0: NO (198.0/1.0)
feo_binarized = 1: SI (9.0)

Number of Leaves  :     2

Size of the tree :      3


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         206               99.5169 %
Incorrectly Classified Instances         1                0.4831 %
Kappa statistic                          0.9448
Mean absolute error                      0.0096
Root mean squared error                  0.0697
Relative absolute error                  9.9972 %
Root relative squared error             32.4973 %
Total Number of Instances              207
Ignored Class Unknown Instances               1039
```

Fig. 20 Resultado para el criterio Negativo I.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,100    0,995      1,000    0,997      0,946    0,500     0,158     NO
               0,900    0,000    1,000      0,900    0,947      0,946    0,905     0,901     SI
Weighted Avg.  0,995    0,095    0,995      0,995    0,995      0,946    0,519     0,194

=== Confusion Matrix ===

   a   b   <-- classified as
 197   0 |   a = NO
   1   9 |   b = SI
```

Fig. 21 Resultado para el criterio Negativo II.

A simple vista, los resultados nos dicen que: 197 personas, opinan que no es “Negativo”, es decir, este número de personas opinan: que Fideos Cayambe no es “feo”, y tan solo 9 personas opinan que, sí es “Negativo”, teniendo una precisión de 99,51 % de clasificación correcta, tomando en cuenta 199 registros de la base total y tan solo un 0,48% no clasificados correctamente, opinión obtenida con mayor precisión.

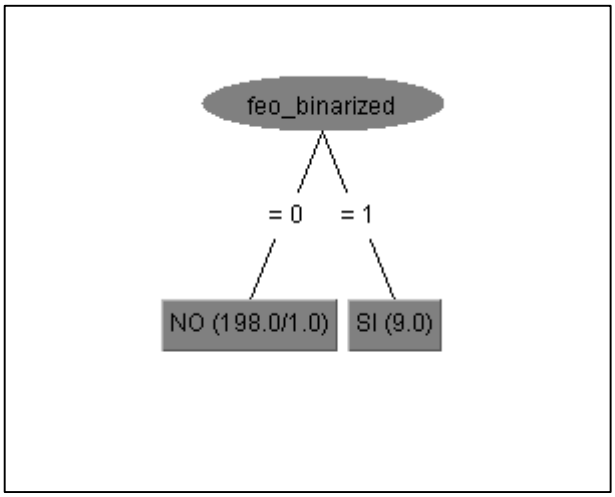


Fig. 22 Resultado árboles para el criterio Negativo.

Claramente se observa que el factor determinante para la opinión "Negativo" está dado por las personas que comentaron que Fideos Cayambe no es "feo".

Se procesa el algoritmo J48 para el resultado el criterio "Neutro".

```
Correctly Classified Instances        1194               99.6661 %
Incorrectly Classified Instances         4                0.3339 %
Kappa statistic                          0.9878
Mean absolute error                      0.0067
Root mean squared error                  0.0578
Relative absolute error                  2.4166 %
Root relative squared error             15.5853 %
Total Number of Instances             1198
Ignored Class Unknown Instances                48

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,998    0,010    0,998      0,998    0,998      0,988    0,894     0,952     NO
                 0,990    0,002    0,990      0,990    0,990      0,988    0,989     0,970     SI
Weighted Avg.    0,997    0,009    0,997      0,997    0,997      0,988    0,910     0,955

=== Confusion Matrix ===

   a   b   <-- classified as
 999   2 |   a = NO
   2 195 |   b = SI
```

Fig. 23 Resultado árboles para el criterio neutro .

Para el criterio "Neutro" contamos con el 99,66% de la clasificación correcta es decir 1194 registros han sido clasificados de manera correcta y un 0.33% de manera incorrecta lo cual no indica un gran porcentaje de

precisión en la clasificación, de tal manera que tenemos en la matriz de confusión que b= SI lo que nos indica que 195 personas se mantienen con un comentario “Neutro”, no les agrada Fideos Cayambe pero tampoco les disgusta

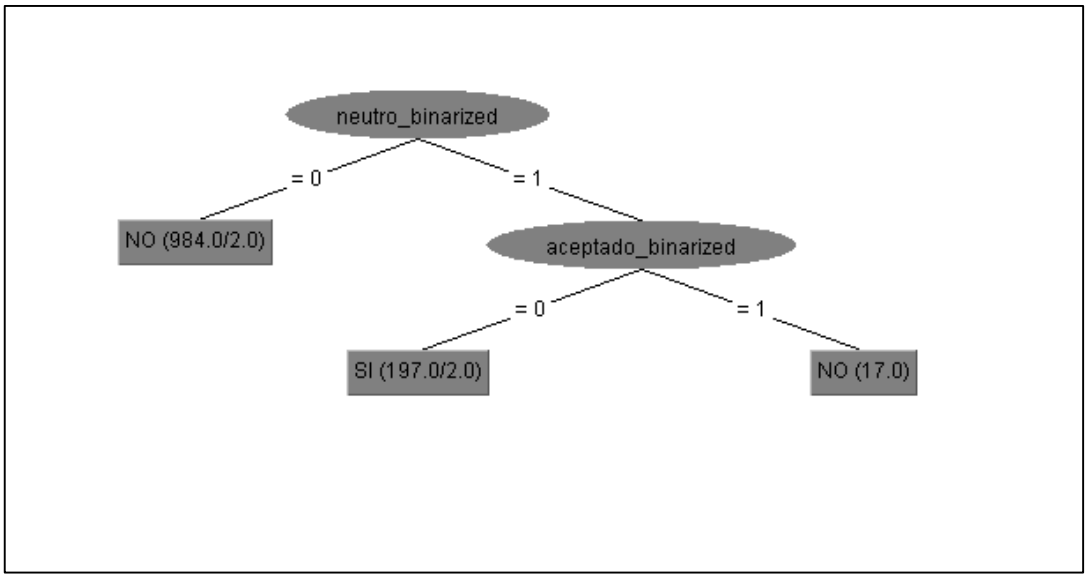


Fig. 24 Resultado árboles para el criterio neutro.

Los factores determinantes para este criterio han sido los comentarios que tienen una opinión neutra y que de una manera subjetiva han aceptado la marca Fideos Cayambe en los comentarios analizados, y de estos tenemos personas que aceptaron la marca.

Aplicación de Relación y participación con la marca tomando en cuenta la ubicación aplicando el algoritmo J48.

```
tipo_objeto = Facebook:<post>/comments
|   sexo = F: Guayaquil (792.0/639.0)
|   sexo = M: Santa Elena (413.0/337.0)
tipo_objeto = Facebook:<video>/comments
|   sexo = F: Quito (28.0/20.0)
|   sexo = M: Guayaquil (13.0/7.0)

Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         243               19.5024 %
Incorrectly Classified Instances      1003               80.4976 %
Kappa statistic                          0.0289
Mean absolute error                      0.105
Root mean squared error                  0.2294
Relative absolute error                 99.5562 %
Root relative squared error             99.929  %
Total Number of Instances             1246
```

Fig. 25 Resultado de relación con la marca por ubicación I.

```
Total Number of Instances            1246

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,713    0,631    0,198      0,713    0,309      0,065    0,534     0,196     Guayaquil
               0,000    0,000    0,000      0,000    0,000      0,000    0,451     0,067     Portoviejo
               0,367    0,324    0,184      0,367    0,245      0,034    0,499     0,163     Santa Elena
               0,000    0,000    0,000      0,000    0,000      0,000    0,498     0,116     Machala
               0,000    0,000    0,000      0,000    0,000      0,000    0,468     0,044     Esmeraldas
               0,000    0,000    0,000      0,000    0,000      0,000    0,482     0,067     Loja
               0,000    0,000    0,000      0,000    0,000      0,000    0,508     0,030     Otavalo
               0,000    0,000    0,000      0,000    0,000      0,000    0,439     0,072     Cayambe
               0,000    0,000    0,000      0,000    0,000      0,000    0,371     0,044     Ibarra
               0,000    0,000    0,000      0,000    0,000      0,000    0,430     0,025     Cañar
               0,000    0,000    0,000      0,000    0,000      0,000    0,390     0,026     Cuenca
               0,000    0,000    0,000      0,000    0,000      0,000    0,454     0,039     Ambato
               0,000    0,000    0,000      0,000    0,000      0,000    0,507     0,038     Latacunga
               0,000    0,000    0,000      0,000    0,000      0,000    0,245     0,006     Riobamba
               0,000    0,000    0,000      0,000    0,000      0,000    0,215     0,001     Quevedo
               0,000    0,000    0,000      0,000    0,000      0,000    0,216     0,001     Jipijapa
               0,421    0,016    0,286      0,421    0,340      0,335    0,695     0,137     Quito
Weighted Avg.  0,195    0,167    0,070      0,195    0,101      0,022    0,480     0,104
```

Fig. 26 Resultado de relación con la marca por ubicación II.

Se puede observar, que en general, no se cuenta con un alto porcentaje de precisión sobre, la participación y relación con la marca, y los que cuentan y se destacan son ubicaciones de: Guayaquil, Santa Elena y Quito, que constan con una precisión, más alta, en comparación a las demás ubicaciones.

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   <-- classified as
 159   0  62   0   0   0   0   0   0   0   0   0   0   0   0   0   2 |   a = Guayaquil
  58   0  34   0   0   0   0   0   0   0   0   0   0   0   0   0   2 |   b = Portoviejo
 128   0  76   0   0   0   0   0   0   0   0   0   0   0   0   0   3 |   c = Santa Elena
  91   0  56   0   0   0   0   0   0   0   0   0   0   0   0   0   1 |   d = Machala
  34   0  24   0   0   0   0   0   0   0   0   0   0   0   0   0   2 |   e = Esmeraldas
  52   0  36   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   f = Loja
  25   0   8   0   0   0   0   0   0   0   0   0   0   0   0   0   2 |   g = Otavalo
  69   0  34   0   0   0   0   0   0   0   0   0   0   0   0   0   3 |   h = Cayambe
  48   0  24   0   0   0   0   0   0   0   0   0   0   0   0   0   1 |   i = Ibarra
  27   0   9   0   0   0   0   0   0   0   0   0   0   0   0   0   1 |   j = Cañar
  24   0  13   0   0   0   0   0   0   0   0   0   0   0   0   0   2 |   k = Cuenca
  37   0  18   0   0   0   0   0   0   0   0   0   0   0   0   0   1 |   l = Ambato
  37   0  11   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   m = Latacunga
   7   0   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   n = Riobamba
   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   o = Quevedo
   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 |   p = Jipijapa
   7   0   4   0   0   0   0   0   0   0   0   0   0   0   0   0   8 |   q = Quito
```

Fig. 27 Matriz de confusión por ubicación II.

Como se muestra en la matriz de confusión determinamos que las ubicaciones con mayor determinación son la a,c,q.

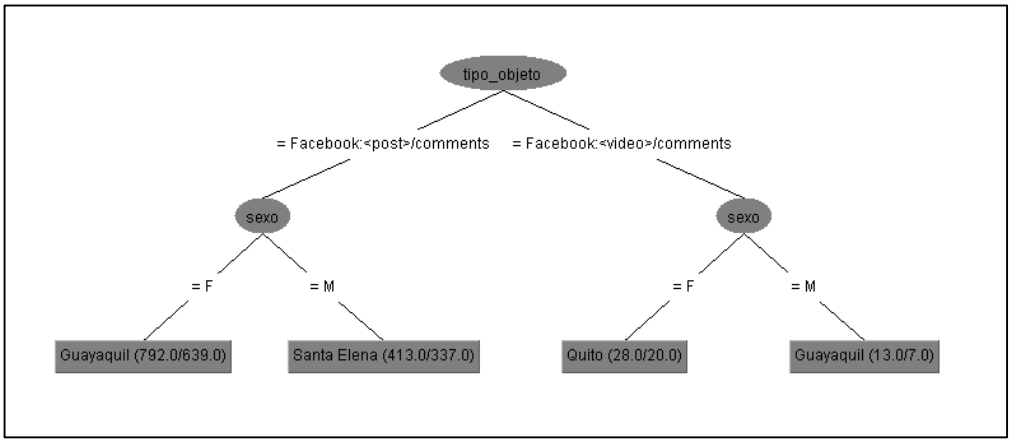


Fig. 28 Resultado de árboles para relación con la marca por ubicación.

Según la Fig. 28 se observa que, la participación de las personas con la marca Fideos Cayambe está determinada por: tipo-objeto, que son: los

comentarios de los post, y los comentarios de los videos publicados. El factor determinante para esta participación es: el sexo, donde se observa que: en Guayaquil, se muestra 639 personas pertenecientes al sexo Femenino, y en Santa Elena, 337 del sexo masculino, que participan en los comentarios de los post publicados. En la ciudad de Quito: 20 personas del sexo femenino y 7 del sexo masculino, participan de los comentarios de los videos publicados, dejando en menor número los comentarios de los videos, entonces, las mujeres en Guayaquil son las que más participan de los post de las publicaciones de Fideos Cayambe

Participación en periodos de tiempo M5P.

```
año <= 2016.5 : LM1 (893/86.969%)
año >  2016.5 : LM2 (353/83.99%)

LM num: 1
mes =
        0.024 * tipo_objeto=Facebook:<video>/comments
        - 0.0332 * año
        + 0.7062 * ubicación=Esmeraldas,Ibarra,Cayambe,Guayaquil,Portoviejo,Loja,Latacunga,Ambato,Machala,Quito,Riobamba,Quevedo,Jipijapa
        + 0.5275 * ubicación=Loja,Latacunga,Ambato,Machala,Quito,Riobamba,Quevedo,Jipijapa
        + 1.4137 * Positivo=NO
        + 75.246

LM num: 2
mes =
        0.0592 * tipo_objeto=Facebook:<video>/comments
        - 0.0818 * año
        + 0.0192 * ubicación=Esmeraldas,Ibarra,Cayambe,Guayaquil,Portoviejo,Loja,Latacunga,Ambato,Machala,Quito,Riobamba,Quevedo,Jipijapa
        + 0.376 * ubicación=Portoviejo,Loja,Latacunga,Ambato,Machala,Quito,Riobamba,Quevedo,Jipijapa
        + 0.0182 * ubicación=Loja,Latacunga,Ambato,Machala,Quito,Riobamba,Quevedo,Jipijapa
        - 0.7007 * ubicación=Machala,Quito,Riobamba,Quevedo,Jipijapa
        + 0.0291 * Positivo=NO
        + 172.2291

Number of Rules : 2
```

Fig. 29 Resultado de participación por periodos de tiempo.

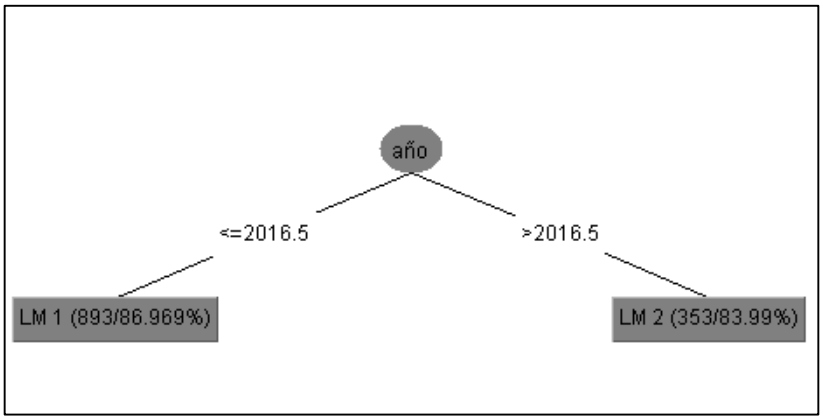


Fig. 30 Resultado participación por año.

Se observa que, para definir la participación de las personas, se toma como el factor determinante, el año, considerando el anterior y posterior al 2016, pues es el período de mayor participación de comentarios, con respecto, a los post publicados por Fideos Cayambe.

```
LM num: 1
año =
        -0.1404 * mes
        + 0.1016 * sexo=F
        + 0.0105 * ubicación=Ambato,Latacunga,Guayaquil,Esmeraldas,Portoviejo,Santa Elena,Loja,Quito,Ibarra,Cuenca,Cañar,Otavalo
        - 0.006 * ubicación=Santa Elena,Loja,Quito,Ibarra,Cuenca,Cañar,Otavalo
        + 0.0163 * ubicación=Loja,Quito,Ibarra,Cuenca,Cañar,Otavalo
        + 0.5079 * Positivo=NO
        + 2016.8768

LM num: 2
año =
        -0.0008 * mes
        + 0.1825 * ubicación=Ambato,Latacunga,Guayaquil,Esmeraldas,Portoviejo,Santa Elena,Loja,Quito,Ibarra,Cuenca,Cañar,Otavalo
        - 0.0812 * ubicación=Santa Elena,Loja,Quito,Ibarra,Cuenca,Cañar,Otavalo
        + 0.1695 * ubicación=Loja,Quito,Ibarra,Cuenca,Cañar,Otavalo
        + 0.3424 * Positivo=NO
        + 2016.4385

LM num: 3
año =
        -0.0022 * mes
        + 0.002 * ubicación=Ambato,Latacunga,Guayaquil,Esmeraldas,Portoviejo,Santa Elena,Loja,Quito,Ibarra,Cuenca,Cañar,Otavalo
        + 0.003 * ubicación=Loja,Quito,Ibarra,Cuenca,Cañar,Otavalo
        + 0.0064 * Positivo=NO
        + 2016.0227

Number of Rules : 3
```

Fig. 31 Resultado participación por mes I.

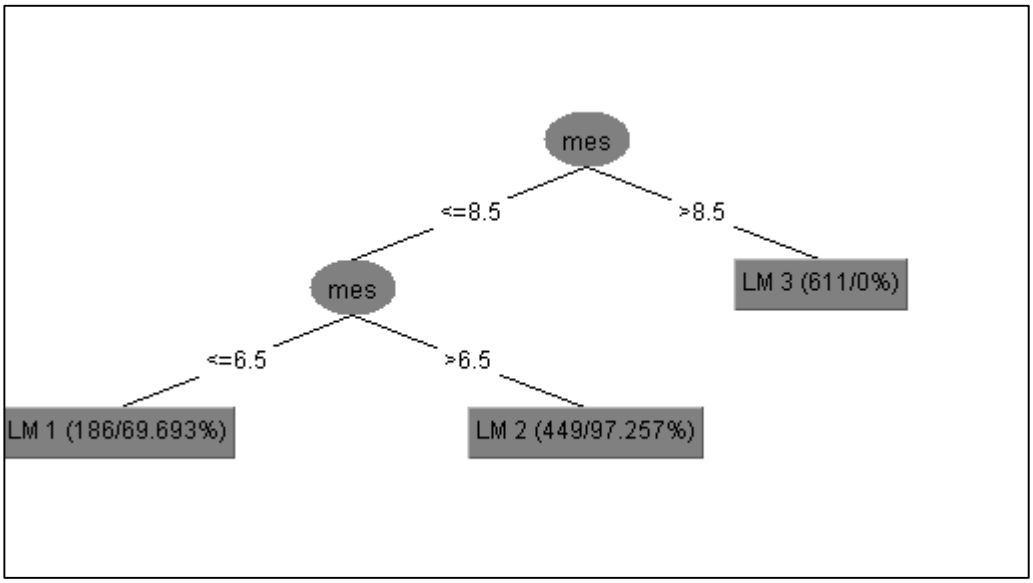


Fig. 32 Resultado participación por mes II.

En los meses: Junio y Agosto existen el mayor número de participaciones de comentarios, fenómeno que obedece a la campaña de regreso a clases por Fideos Cayambe en este período. Por consiguiente se confirma que ha tenido una mayor recepción por parte de los seguidores.

Clustering

Se aplica la técnica de “Clustering” para la clasificación de los datos otorgados por Fideos Cayambe, que permite ejecutar el algoritmo KMeans y con este procesamiento se obtiene los siguientes resultados.

Se aplica el algoritmo KMeans que nos da como resultado dos clusters 0 – 1 con todos los campos obtenidos de la base Fideos Cayambe.

Attribute	Full Data (1246.0)	Cluster# 0 (426.0)	1 (820.0)
tipo_objeto	Facebook:<post>/comments	Facebook:<post>/comments	Facebook:<post>/comments
mes	8.7055	8.6901	8.7134
año	2016.2833	2016.2793	2016.2854
sexo	0.6581	0	1
ubicación	Guayaquil	Santa Elena	Guayaquil
Positivo	0.8034	0.8028	0.8037
masculino	0.3419	1	0
femenino	0.6581	0	1
guayaquil	0.179	0.1596	0.189
Negativo	0.008	0.007	0.0085
economico	0.1164	0.1268	0.111
delicioso	0.3917	0.3967	0.389
aceptado	0.8042	0.7958	0.8085
rendidor	0.0225	0.0188	0.0244
nutritivo	0.0144	0.0141	0.0146
feo	0.0072	0.007	0.0073
caro	0.0008	0	0.0012
neutro	0.1717	0.169	0.1732

Fig. 33 Resultado Cluster.

Tenemos dos clusters claramente diferenciados: el Cluster 0, nos da un 34% (426) , y en el Cluster 1, nos da 66% (820) de la información, lo que, nos permite ver que en el cluster 1 se encuentra la mayor población con pensamiento positivo y perteneciente al género femenino en Guayaquil, que a su vez, han opinado de manera positiva en cuanto a los post publicados por Fideos Cayambe, de igual manera en Santa Elena, tenemos opiniones positivas, como son económico y delicioso, que no llegan a ser un número mayor, pero se destaca el pensamiento positivo.

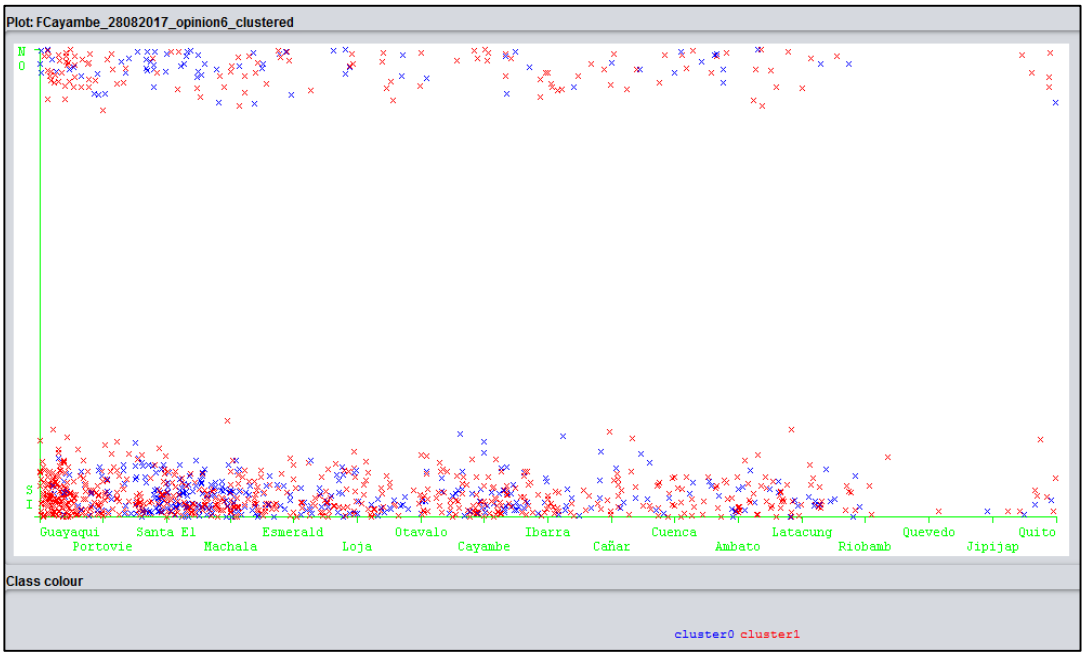


Fig. 34 Resultado Cluster gráfico I.

Se puede observar, que la gran mayoría de personas se han clasificado en el cluster 1, de acuerdo al análisis de los comentarios publicados por los seguidores de la marca Fideos Cayambe. Identificamos que la gran mayoría emite pensamientos positivos con relación a la marca.

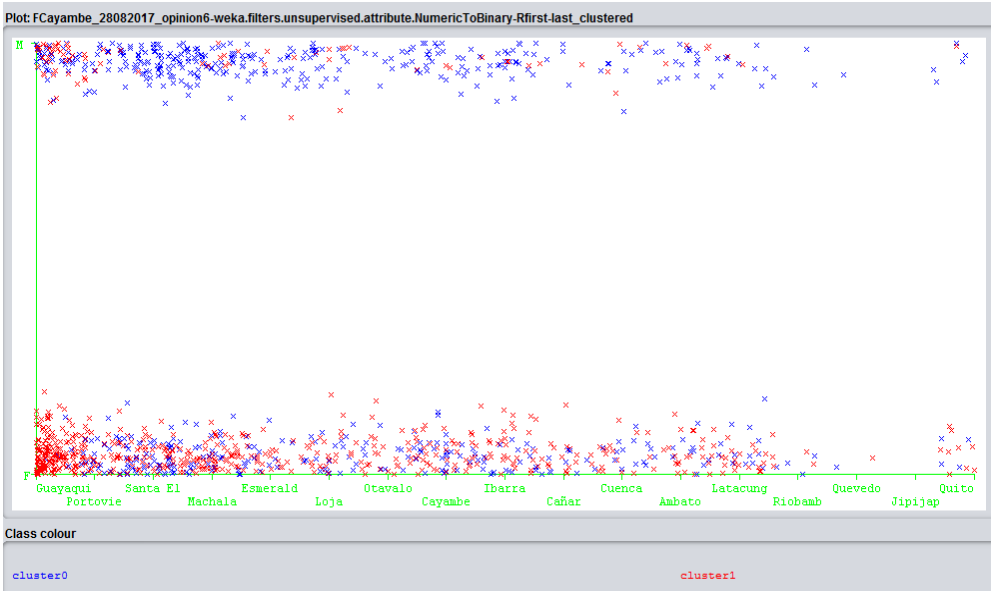


Fig. 35 Resultado Cluster gráfico II.

En la Fig.32, se observa por la ubicación, como está la clasificación por cada cluster, dado el sexo Masculino y femenino, donde se aprecia claramente con color rojo las personas de Guayaquil y el sexo masculino en su mayoría en Santa Helena.

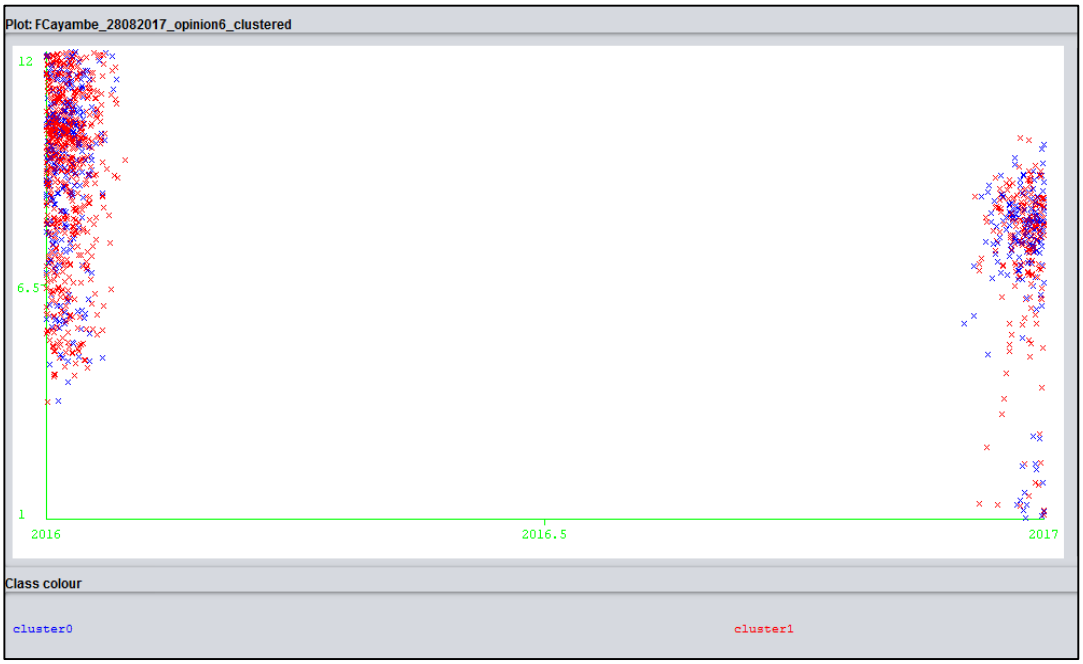


Fig. 36 Resultado Cluster gráfico III.

En la figura 36, se aprecia la participación de las personas en el rango de años 2016 y 2017, existe más participación identificada por el cluster 1, que nos indica que hay más participación en el mes de junio para adelante en el año 2016, y que en ese mismo mes en el 2017 se acumula la mayor parte de participación, y esto lo explica claramente por la campaña lanzada por período de salida a vacaciones y regreso a clases, que es donde se promociona más la marca Fideos Cayambe, como estrategia de marketing.

Regresión Lineal

Se ha aplicado el modelo de regresión lineal simple y se ha obtenido los siguientes resultados con los datos de Fideos Cayambe.

Positivo

```
Linear Regression Model

Positivo =

      0.0718 * economico +
      0.0728 * delicioso +
      0.8102 * aceptado +
     -0.7243 * rendidor +
     -0.7424 * nutritivo +
      0.142

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.8995
Mean absolute error                      0.0902
Root mean squared error                  0.1736
Relative absolute error                 28.5536 %
Root relative squared error             43.6856 %
Total Number of Instances             1246
```

Fig. 37 Resultado para el criterio "Positivo".

En este resultado, se puede apreciar que existe una correlación entre las variables de un 89%, lo cual indica que la variable con mayor relación para que un criterio sea Positivo, es la variable aceptado, con menor porcentaje delicioso y económico con un 28 % de error, entonces esto indica que las personas opinan de manera Positiva en cuanto a la marca Fideos Cayambe; de igual manera la opinión: que la marca es "delicioso", "económico" y que la mayor parte de personas aceptan el producto

Negativo

```
Linear Regression Model

Negativo =

      1      * feo +
      1      * caro +
      0

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correlation coefficient                  1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             1246
```

Fig. 38 Resultado para el criterio Negativo.

Para el criterio Negativo, las variables con relación directa son “feo” y “caro”, que son las de mayor influencia para que las personas opinen de esta manera en cuanto a la marca Fideos Cayambe; seguido de esto podemos observar que tenemos un coeficiente de correlación de 1 lo cual demuestra que es una relación directa y que tiene un error relativo del 0%.

Neutro

```
     0.0031 * mes +
     0.0212 * año +
     0.0136 * ubicación=Santa Elena,Esmeraldas,Ambato,Portoviejo,Guayaquil,Loja,Otavalo,Ibarra,Latacunga,Cuenca,Cañar,Quito +
     0.036  * ubicación=Loja,Otavalo,Ibarra,Latacunga,Cuenca,Cañar,Quito +
    -0.0186 * ubicación=Otavalo,Ibarra,Latacunga,Cuenca,Cañar,Quito +
     0.0555 * ubicación=Cañar,Quito +
    -0.0799 * ubicación=Quito +
     0.418  * Positivo=NO +
     0.2644 * Muy positivo=NO +
     0.2442 * Negativo=NO +
     0.418  * Neutro=SI +
    -0.0439 * economico +
    -0.0292 * delicioso +
    -0.055  * aceptado +
     0.0876 * rendidor +
     0.064  * nutritivo +
   -43.1308

Time taken to build model: 0.09 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9394
Mean absolute error                      0.0403
Root mean squared error                  0.1293
Relative absolute error                 14.1651 %
Root relative squared error             34.2633 %
```

Fig. 39 Resultado para el criterio Neutro.

Para el análisis de criterio Neutro, tenemos claramente que un 41% de correlación con la variable neutro, que hace referencia al análisis realizado de los comentarios, en los cuales se ha identificado que las personas no tiene el criterio de delicioso, económico, rendidor, nutritivo sino más bien aceptan la marca pero no están atados a ella, dicho así también existieron comentarios en los que no se refieren a la marca si no que no tienen sentido de contexto.

5 CAPITULO IV

RESULTADOS Y DISCUSIÓN

Dentro de la fase, se toma en cuenta los resultados obtenidos de cada uno de los modelos, verificando el cumplimiento de los objetivos del negocio que van alineados en este caso a los objetivos de minería de datos.

Evaluación de resultados

Objetivo 1

Verificar la existencia de la relación entre el consumidor y el producto de la marca Fideos Cayambe. Dado que 29602 son los seguidores de la página Fideos Cayambe en Facebook y tomando en cuenta que esta estrategia de mercado que tiene como objetivo: incrementar la participación de su target con la marca, se ha logrado identificar la participación de las personas a nivel nacional como se muestra en la Fig. 43

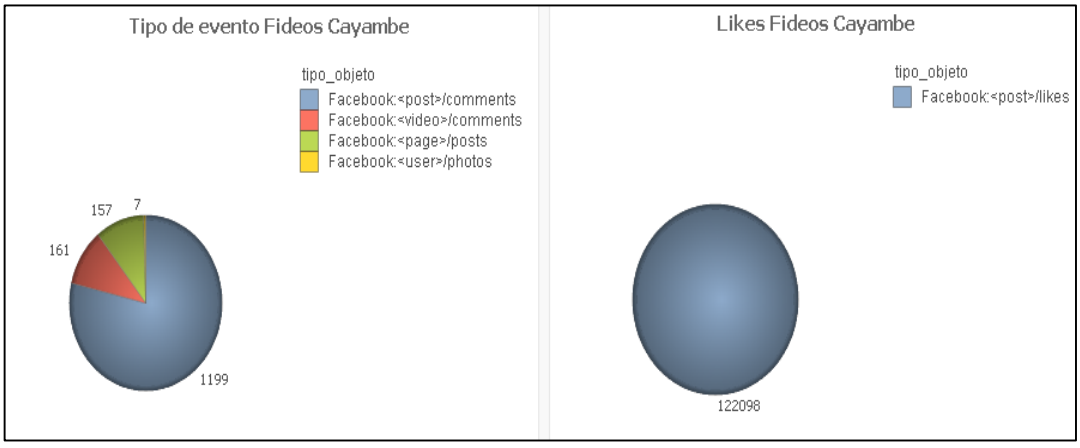


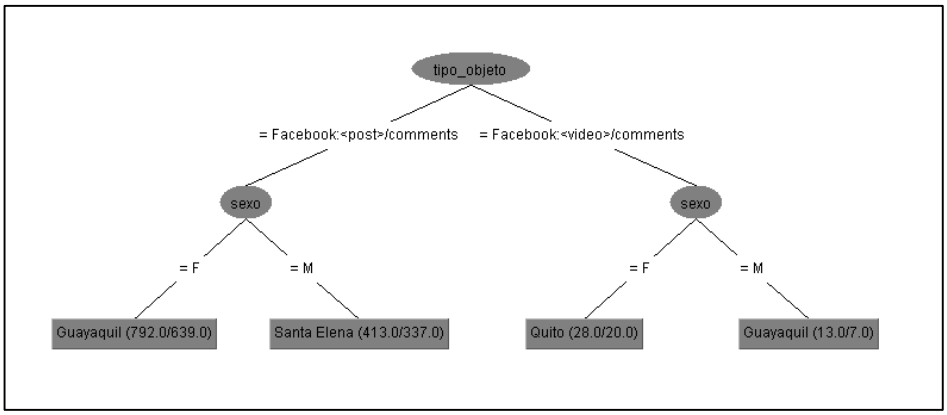
Fig. 40 Participación en Fideos Cayambe.

Si sumamos la participación en la página de Fideos Cayambe en estos dos últimos dos años, como se observa en la imagen, podemos ver que existe 1360 personas, que interactúan realizando comentarios en cuanto a los post

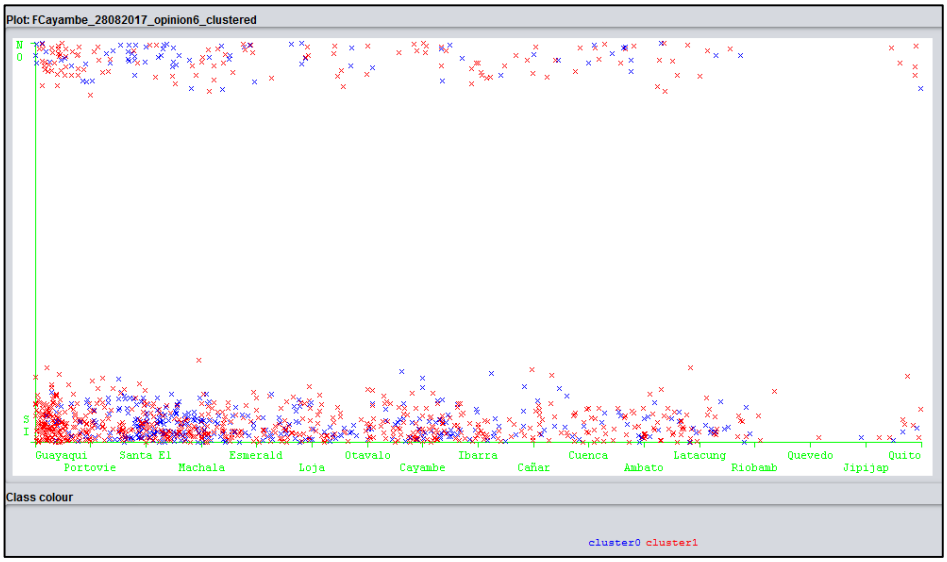
que realiza la marca, y como likes por post, se observa 122098 personas que siguen a esta marca y les gusta sus publicaciones.

Para la clasificación realizada con las tres técnicas, observamos que el factor más determinante para árboles, es el tipo de objeto, en este caso los comentarios realizados divididos por el sexo de las personas que han comentado dependiendo de la ubicación, lo cual nos da una medida cuantitativa de los casos clasificados.

Árboles



Clustering



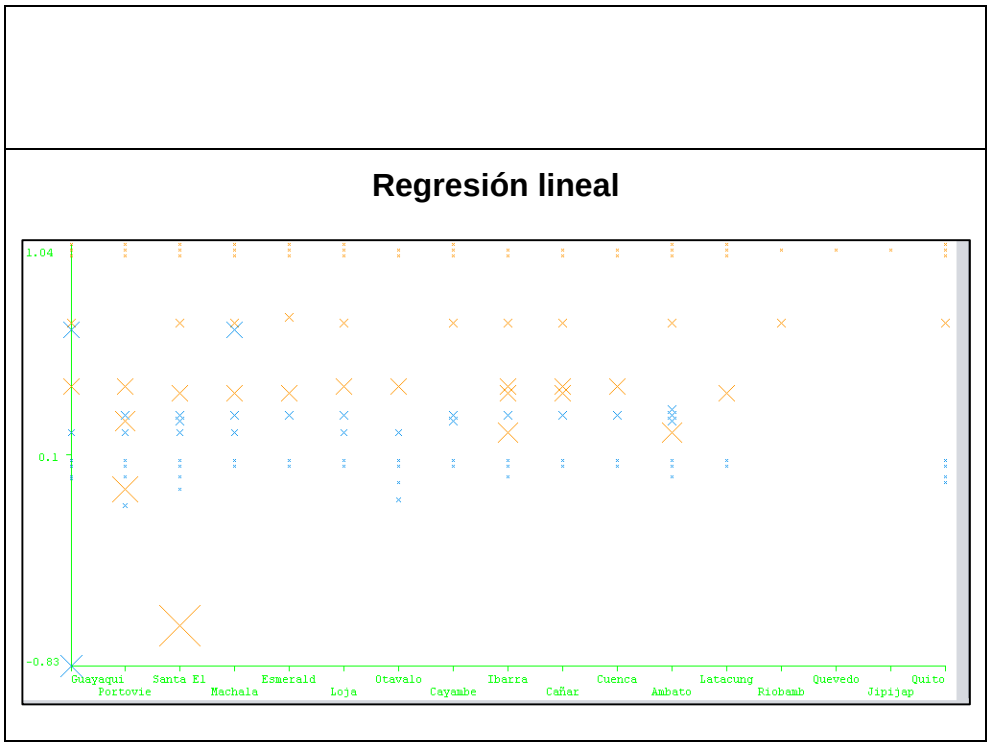


Fig. 41 Grafica comparativa de participación Fideos Cayambe.

La técnica de cluster, por otro lado, nos segmenta en dos grupos, en los cuales también toma en cuenta: la ubicación y la semejanza en las variables, pero no nos da un número exacto de las personas que están en cada una de estas categorías si no un total sumariado.

En la técnica de regresión lineal, nos indica la correlación de las variables, en este caso positivas. en Santa Elena y Guayaquil, como muestra la gráfica de la Fig. 38

Objetivo 2

Determinar la participación del consumidor en la red social Facebook por periodos de tiempo.

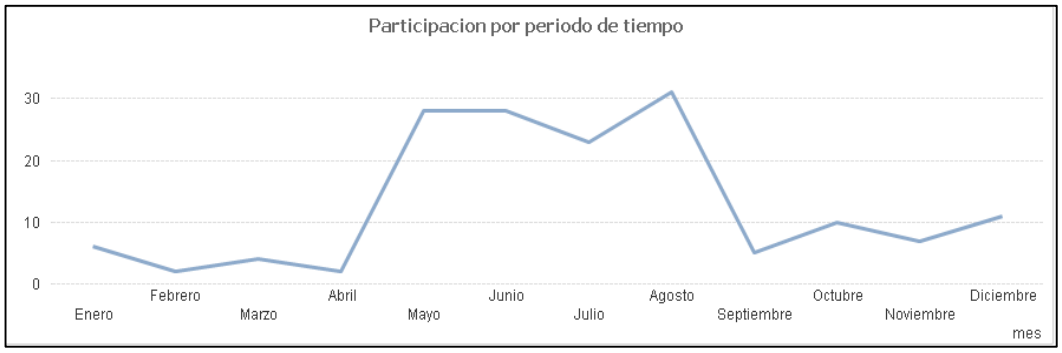
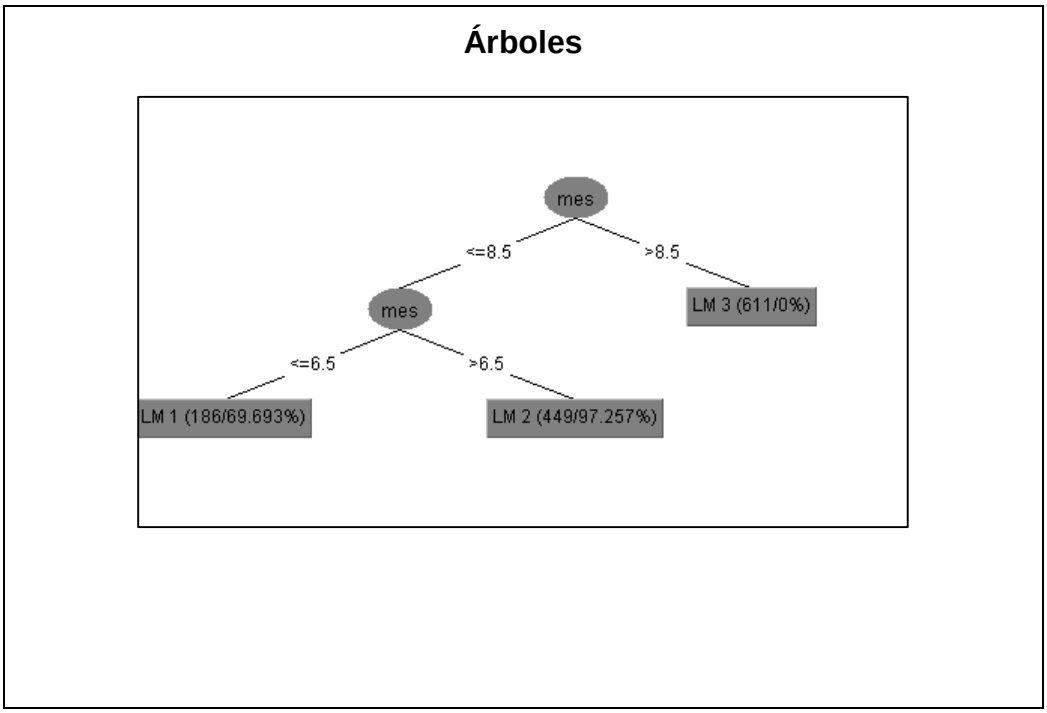


Fig. 42 Participación por periodo de tiempo Fideos Cayambe.

Como se observa en la gráfica, la participación de los seguidores se da más desde el mes de abril, donde concuerda muy bien dado que en estos meses existe una campaña por temporalidad por parte de la marca, es decir, días festivos, como: día de la madre, salida de vacaciones y regreso a clases, son oportunidades de promoción, en los que Fideos Cayambe le saca provecho para aumentar su engagement para su participación, de esta manera, se puede observar su participación y fidelidad a la marca por medio de esta red social.



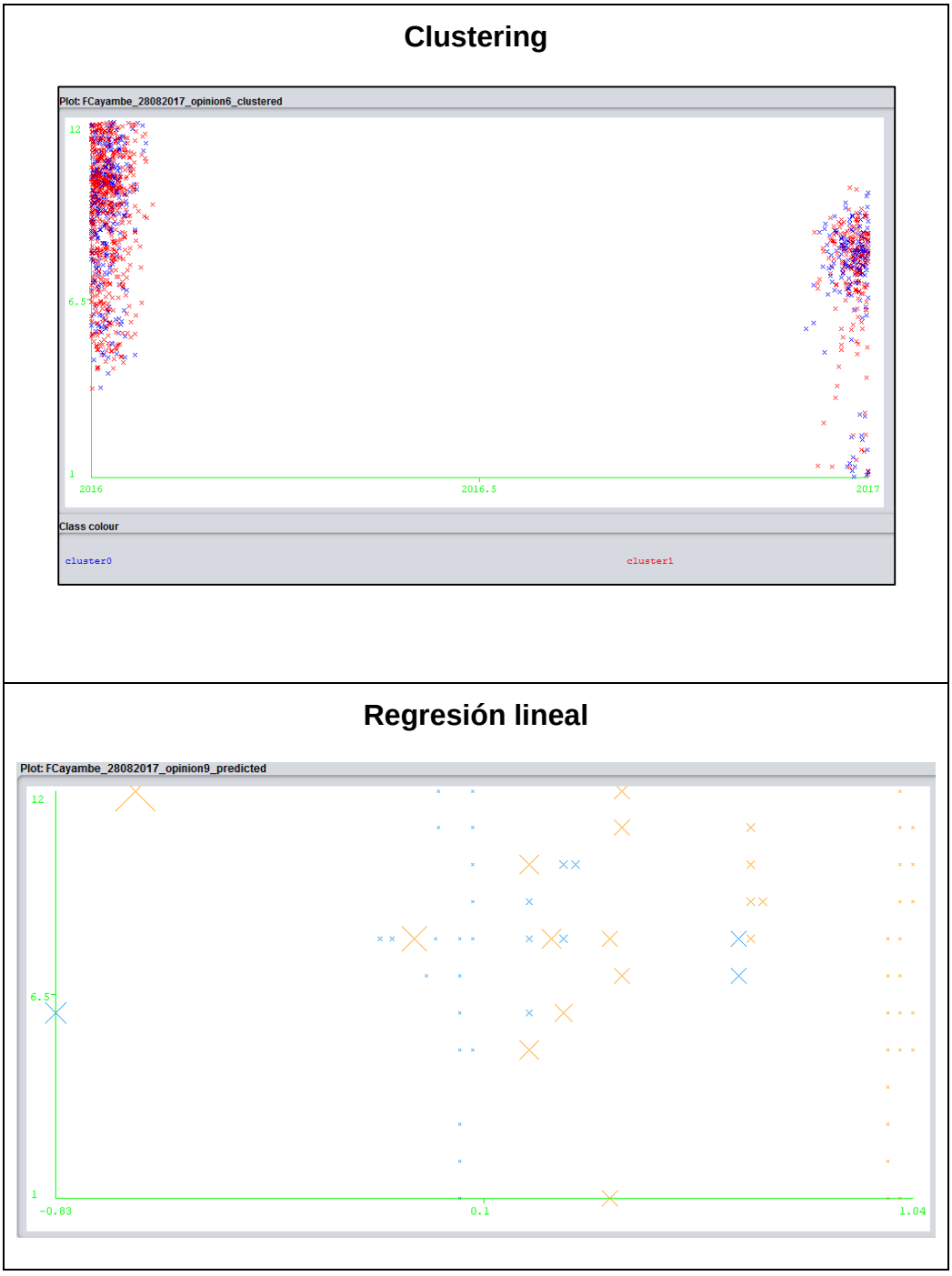


Fig. 43 Comparativo de participación por periodo de tiempo.

Como podemos observar en árboles, tenemos los datos, más precisos ,en cuanto a los meses que se ha destacado por la participación de las personas, meses posteriores a julio, de igual manera en clustering nos indica que los comentarios han sido más el año 2016 , debido a que el año 2017 aún no termina y no se puede estimar este criterio, pero se especifica

que la mayor afluencia esta del mes de julio en adelante, en regresión lineal no se toma en cuenta como un factor determinante el mes ni el año de publicación de un comentario para definir que sea positivo o negativo.

Objetivo 3

Determinar las opiniones existentes de los consumidores en relación a la marca.

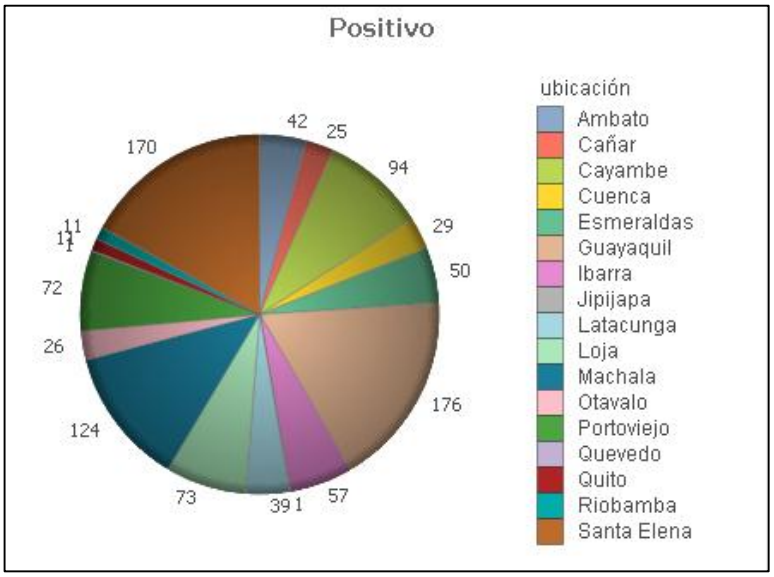


Fig. 44 Criterio Positivo Fideos Cayambe.

Según el análisis realizado en cuanto a los comentarios de los post realizados por las personas seguidoras de la página de Fideos Cayambe en Facebook, notamos que la mayor parte de estos comentarios se segmentan en el criterio positivo, tomando en cuenta los sub criterios delimitados, como son: económico, aceptado, delicioso, rendidor y nutritivo que son las palabras con mayor afluencia dentro de los comentarios.

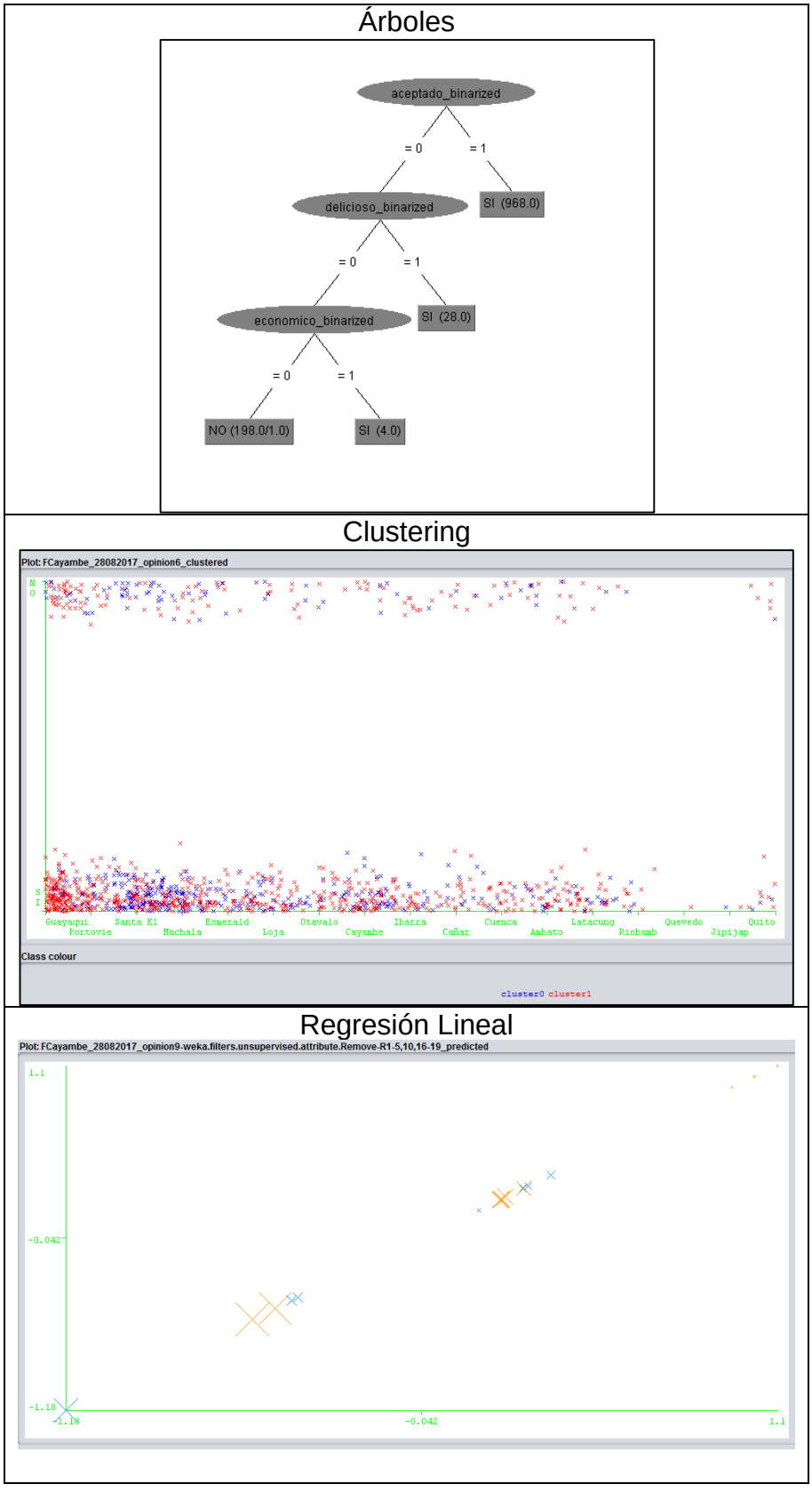


Fig. 45 Comparativo criterio Positivo.

Para el análisis del criterio Positivo, aplicando la técnica de árboles, se ha identificado los factores más determinantes para que sea positivo, el criterio final, se da debido a que existen más personas que opinan que la marca Fideos Cayambe son: aceptado, delicioso y económico; esto nos indica numéricamente ; mientras que, en clustering no agrupa por la cantidad de semejanzas, donde la mayor cantidad de personas está en el cluster 1, identificado por las personas que piensan de manera positiva, con respecto a la marca y son de sexo femenino, para regresión lineal las variables con más relación son: aceptado, económico y delicioso, de manera que, se concluye que, el pensamiento en cuanto a la marca Fideos Cayambe es positiva.

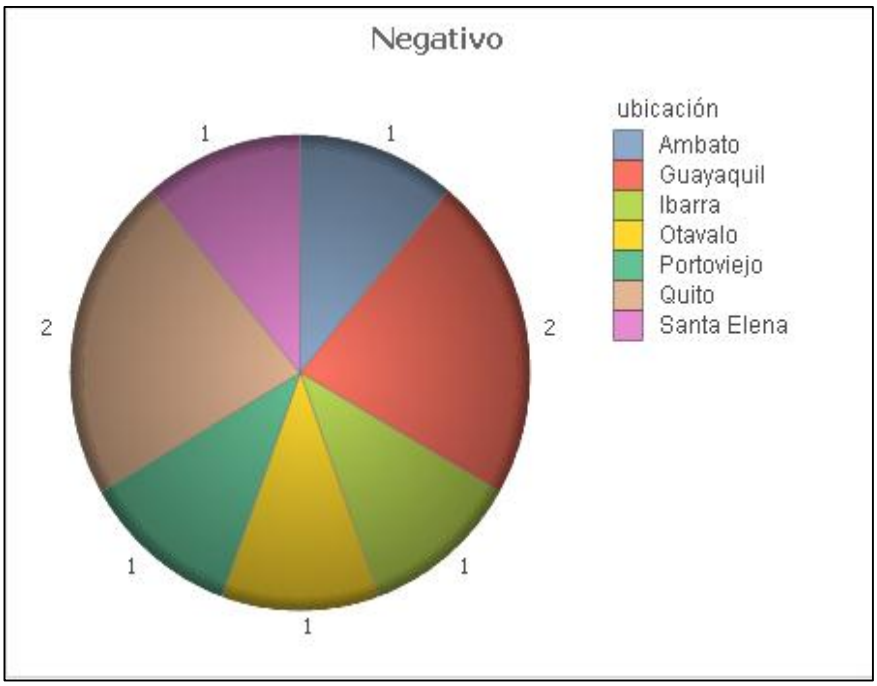


Fig. 46 Pensamiento Negativo Fideos Cayambe.

Para el criterio Negativo, se observa claramente, que no existe mucha complicación para su determinación, debido a que, 10 personas han comentado de manera negativa, sobre la marca Fideos Cayambe a nivel nacional.

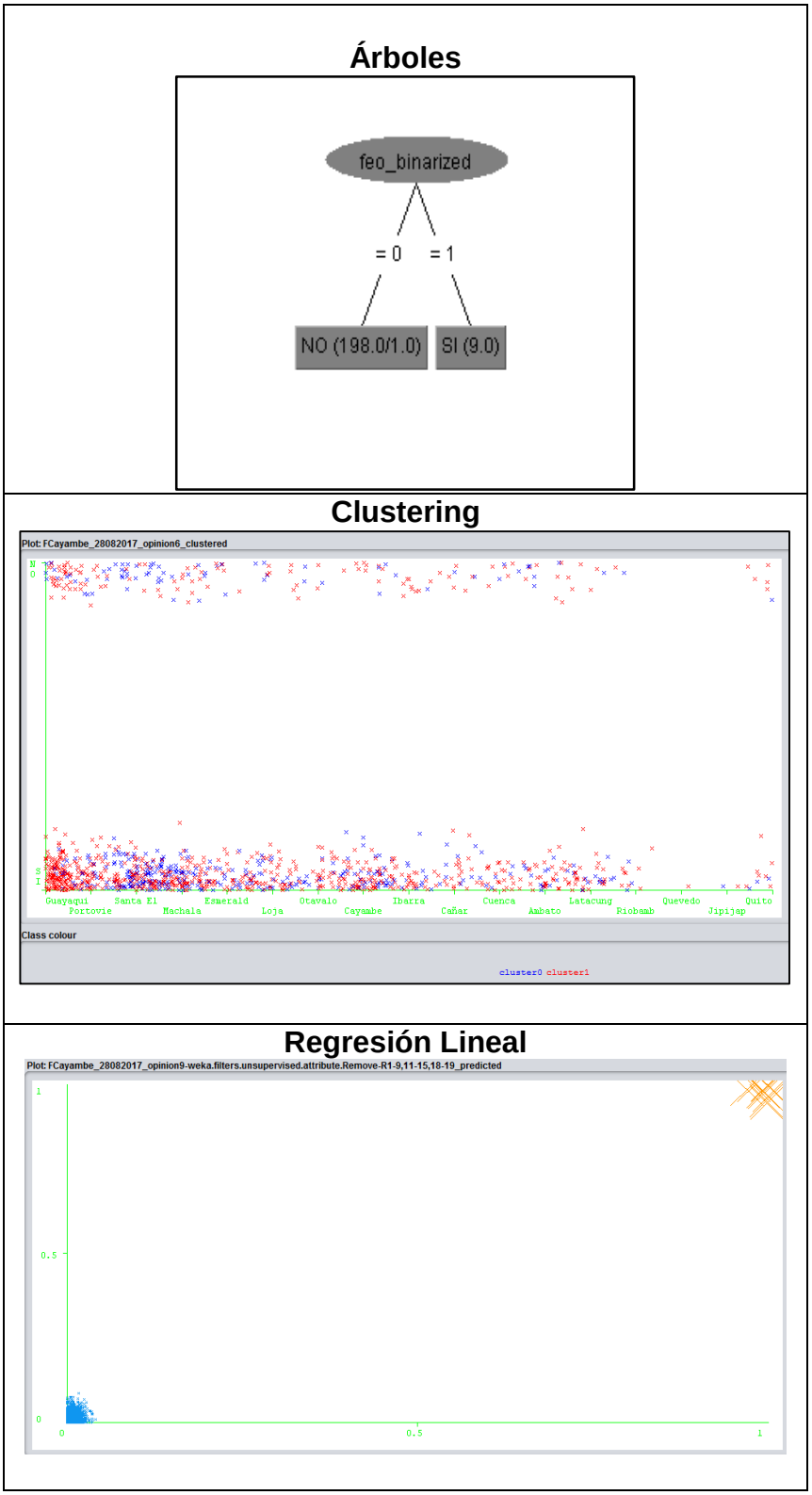


Fig. 47 Comparativo de criterio Negativo Fideos Cayambe.

Para el análisis de sentimiento negativo, las gráficas, fueron muy claras en determinar, la menor parte de personas que piensan de esta manera, con respecto a la marca Fideos Cayambe; en la técnica de árboles

de clasificación, los factores determinantes , son el criterio: feo y caro; para clustering, agrupa a la mayor parte de personas en el pensamiento positivo, lo cual no indica que la parte negativa, es un porcentaje muy insignificante, de igual manera, en regresión lineal, la variable con más correlación es: feo y caro, lo cual no indica también una precisión de un 100%, dado que es una variable directa con respecto a nuestra interrogante, Negativo.

6 CAPITULO V

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- La técnica de árboles de decisión, ha sido la más idónea para el análisis y procesamiento de la información obtenida de Facebook. Sin embargo la técnica de clustering, sirvió de complemento indispensable para la clasificación y uso de las variables en su totalidad.
- La metodología CRISP-DM nos permitió desarrollar la investigación de una manera estructurada y sistemática lo cual permitió destacar las características y propiedades de cada una de las técnicas aplicadas.
- Las técnicas de árboles y clustering, son las que tienen menor porcentaje de error con respecto a la participación de las personas con respecto a la marca.
- Se determina que los meses de mayor participación de las personas en la página de Fideos Cayambe en su red social Facebook, tiene un incremento desde el mes de mayo en adelante, destacando los meses de festividades.
- Al analizar la opinión de las personas con respecto a la marca fideos Cayambe, se determina que, la técnica con mayor precisión, es la de árboles de clasificación, debido a que cuenta con el mayor porcentaje de correlación entre las variables.

- Para la marca Fideos Cayambe, las personas definen en sus comentarios como una opinión positiva, en los post publicados en la red social Facebook, es decir, que la estrategia de marketing para el incremento de engagement está dando resultados positivos.
- La mayor aceptación de los productos de la marca, fideos Cayambe, se determina en las ciudades de: Guayaquil y Santa Elena. Por lo tanto, es la región Costa, la que consume más el Producto.
- Las tres técnicas analizadas han sido de gran utilidad para la marca Fideos Cayambe, pues, la concordancia entre los resultados obtenidos y los objetivos propuestos son parámetros importantes.
- La información obtenida de Facebook, señala que, las personas que participaron en sus comentarios en cuanto al valor nutritivo de los productos de Fideos Cayambe, lo desconocen.

Recomendaciones

- Se recomienda realizar la aplicación de las técnicas de clasificación árboles y clustering ya que presentan los mejores resultados para el tipo de estudio realizado.
- Evaluar periódicamente la información almacenada en la red social Facebook, con el fin de denotar las variables de positividad o negatividad del producto a nivel nacional, para tomar oportunamente los correctivos necesarios.
- Utilizar la metodología CRISP-DM en futuras investigaciones relacionadas con comparativos de técnicas, debido a su alto nivel de aplicabilidad.
- Fideos Cayambe debe implementar una estrategia de marketing dirigida para la región Sierra, con el propósito de una mayor difusión de sus productos a nivel nacional.
- Incrementar su campaña por temporalidad en los meses no festivos, para que no exista tanta variabilidad en la participación de los seguidores mensualmente y así incrementar su enagement todo el año.

- Implementar publicaciones, donde la información que se imparte facilite conocer el valor nutritivo del producto.
- Se recomienda incrementar más eventos como post videos para la campaña de promociones ya que existen muy pocos que se han detectado a los largo de la extracción de la información.
- Los Resultados de las técnicas aplicadas en la presente investigación, permitirán a la Empresa, Fideos Cayambe, consolidar sus estrategias de Marketing.

7 CAPITULO VI

REFERENCIAS BIBLIOGRÁFICAS

Bibliografía

Acevedo Miranda Carlos; Clorio Rodriguez Ricardo; Zagal Flores Roberto; García Mendoza Consuelo V. (2014). Arquitectura Web para análisis de sentimientos en Facebook con enfoque semántico. https://pdfs.semanticscholar.org/ca39/a4018f0cfacacc14922f335d925c1221c678.pdf :[Recuperado el 28/05/2017].

Araujo M. (2016). Analisis de Sentimientos. 2017, de Escuela de Computacion´ Licenciatura en Computacion´ Universidad Central de Venezuela Sitio web: https://www.overleaf.com/articles/analisis-de-sentimientos/knhvsjtbnmyv/viewer.pdf :.[Recuperado el 23/05/2017].

Arcila-Calderón Carlos, Barbosa-Caro Eduar, Cabezuelo-Lorenzo Francisco. (2016). Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística. 2017, de El Profesional de la Información Sitio web: http://recyt.fecyt.es/index.php/EPI/article/view/epi.2016.jul.12

Benkler Yochai.(2015). La riqueza de las redes. Barcelona. http://www.icariaeditorial.com/libros.php?id=1519:[Recuperado el 24/06/2017].

Britos Paola.(2008). Procesos De Explotacion De Informacion Basados En Sistemas Inteligentes. (Tesis de Doctorado).Universidad de la Plata.

Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (1999).CRISP-DM 1.0 Step by step BIguide. http://www.crispdm org/ CRISPWP-0800.pdf.. : [Recuperado el 27/07/2017].

Cortez Vásquez Mg. Augusto; Vega Huerta Mg. Hugo; Pariona Quispe Lic. Jaime. (2009). Procesamiento de lenguaje natural. Facultad de Ingeniería de Sistemas e Informática Universidad Nacional Mayor de San Marcos 2.

http://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923 .[Recuperado el 27/07/2017].

Dufort Guillermo y Álvarez Fabián Kremer; Mordecki Gabriel. (2016). Determinación de la orientación semántica de las opiniones transmitidas en textos de prensa. 2017, de Instituto de Computación, Facultad de Ingeniería Universidad de la República Montevideo, Uruguay Sitio web: https://www.fing.edu.uy/inco/grupos/pln/prygrado/informe_analisis_sentimientos_2015.pdf: [Recuperado el 23/05/2017].

Ecured.cu.(2017).Clustering.https://www.ecured.cu/Clustering:[Recuperado el 27/07/2017].

Estevez Velarde Suilan. Almeida Cruz Yudivian.(2015). EVALUACION DE ALGORITMOS DE CLASIFICACION SUPERVISADA PARA EL MINADO DE OPINION EN TWITTER. Universidad de La Habana, Cuba. http://ojs.uh.cu/InvestigacionOperacional/index.php/InvOp/article/viewFile/480/444 : [Recuperado el 25/08/2017].

Fundeu.(2015).Medios Sociales no Social Media.http://www.fundeu.es/recomendacion/medios-sociales-social-media/: [Recuperado el 27/07/2017]

Hernández M.; Gómez J. (2014). Análisis de Sentimientos Aplicado a Referencias Bibliográficas . 2017, de Escuela Politécnica Nacional, Facultad de Ingeniería de Sistemas Sitio web: http://revistapolitecnica.epn.edu.ec/ojs2/index.php/revista_politecnica2/article/view/126: [Recuperado el 23/05/2017].

IBM DevelopersWorks.(2014).IBM analiza el sentimiento de “twitteros” en Latinoamerica.(2017).Unates Estates. Sitio web: https://www.ibm.com/developerworks/community/blogs/insider/entry/analisis_twitteros_tuiteros_ibm?lang=en

IGELSI Software company.(2015).Casos de éxito.(2017). Ecuador.Sitio web: http://www.ingelsi.com.ec/#contact : [Recuperado el 15/05/2017].

http://iabspain.es/categoria-de-estudio/topic/redes-sociales*: [*Recuperado el 25/05/2017].

Kushal Dave,Steve Lawrence,David M. Pennock. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. 2017, de This work conducted at NEC Laboratories America, Princeton, New Jersey Sitio web: http://www.kushaldave.com/p451-dave.pdf

Mahajan, M.; Nimbhorkar, P.; Varadarajan, K. (2009). «The Planar k-Means Problem is NP-Hard». Lecture Notes in Computer Science 5431: 274–285. doi:10.1007/978-3-642-00202-1_24.

Moreno García María N; Miguel Quintales Luis A; García Peñalvo Francisco J y Polo Martín M. José . (2013). APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN LA CONSTRUCCIÓN Y VALIDACIÓN DE MODELOS PREDICTIVOS Y ASOCIATIVOS A PARTIR DE ESPECIFICACIONES DE REQUISITOS DE SOFTWARE . 2017, de Universidad de Salamanca. Departamento de Informática y Automática Sitio web: http://ceur-ws.org/Vol-84/paper4.pdf: [Recuperado el 27/06/2017].

Mpc.(2015). 14 millones de españoles usan redes sociales.http://www.multiplatformcontent.com/14-millones-de-espanoles-usan-redes-sociales//: [Recuperado el 27/07/2017].

Pedroza Dr. H; Dicovsky Ing. L. (2006). Sistema de Análisis estadístico con SPSS.https://books.google.es/books?id=sE0qAAAAYAAJ&pg=PP90&dq=regresion+lineal&hl=es&sa=X&ved=0ahUKEwj9jL70zoTWAhUISyYKHQ78Bf4Q6AEIPTAF#v=onepage&q=regresion%20lineal&f=false:[Recuperado el 27/07/2017].

Ponce V; Maldonado A.(2016). Redes Sociales:Definición.http://www.citeulike.org/group/20088/article/14034935: [Recuperado el 27/07/2017].

Rodríguez Lic. Ciro; García Lorenzo Dra. C. M. (2016). Adecuación A Metodología De Minería De Datos Para Aplicar A Problemas No Supervisados Tipo Atributo-Valor. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2218362020160004000005&lng=es&nrm=iso :[Recuperado el 27/07/2017].

Rojas Yenory; Ferrández Antonio; Peral Jesús. (2015). Aplicación del Procesamiento de Lenguaje Natural en la Recuperación de Información . 2017, de Departamento de Lenguajes y Sistemas Informáticos Universidad de Alicante Sitio web: https://rua.ua.es/dspace/bitstream/10045/1434/1/PLN_34_02.pdf: [Recuperado el 27/06/2017].

Salas Zárate María del Pilar, Rodríguez García Miguel Ángel, Almela Ángela, Valencia García Rafael. (2011). Estudio de las categorías LIWC para el análisis de sentimientos en español. http://timm.ujaen.es/wp-content/uploads/2014/06/TIMM2014v3.pdf : [Recuperado el 17/06/2017].

Turney P.D.. (2002). Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. 2017, de National Research Council of Canada Sitio web: http://cogprints.org/2322/5/ERB-1094.pdf

Vallez Mari (Universitat Pompeu Fabra) y Pedraza-Jimenez Rafael (Universitat Pompeu Fabra). (2007). El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines. 2017, de Universitat Pompeu Fabra Sitio web: https://www.upf.edu/hipertextnet/numero-5/pln.html:[Recuperado el 27/06/2017].

Vilares David, Miguel A. Alonso y G´omez-Rodr´ıguez Carlos. (2013). Clasificaci´on de polaridad en textos con opiniones en espa˜nol mediante an´alisis sint´actico de dependencias. 2017, de Departamento de Computaci´on, Universidade da Coru˜na Campus de Elvi˜na, 15011 A Coru˜na Sitio web: http://www.grupolys.org/biblioteca/VilAloGom2013a.pdf: .[Recuperado el 23/05/2017].

Estevez Velarde Suilan. Almeida Cruz Yudivian.(2015). EVALUACION DE ALGORITMOS DE CLASIFICACION SUPERVISADA PARA EL MINADO DE OPINION EN TWITTER. Universidad de La Habana, Cuba. http://ojs.uh.cu/InvestigacionOperacional/index.php/InvOp/article/viewFile/480/444 : [Recuperado el 25/08/2017].