**Universidad Nacional Agraria La Molina**

**Maestría de Estadística Aplicada**

**Percepción de los usuarios sobre los centros comerciales usando Análisis de Sentimiento en Lima Metropolitana.**

**Semestre: 2018-2**

**Alumno: Jaime Gómez Marín**

**Asesor: Docente Ana Cecilia Vargas Paredes**

1. **INTRODUCCIÓN**

En los últimos 10 años, la revolución que ha causado el empleo de teléfonos inteligentes en la población ha permitido masificar el uso del internet y podríamos decir sin temor a equivocarnos, la democratización de su uso. La consecuencia es que las personas están más comunicadas y/o conectadas en tiempo real a través de comunidades en internet, las cuales son más conocidas con el nombre de “redes sociales”, el caso más emblemático de su empleo en una revolución políticas sucedió entre los años 2011 y 2012 , con el movimiento denominado “Primavera Árabe” (1), donde las redes sociales tuvieron una vital importancia para la organización de este movimiento, siendo las redes sociales más empleadas Facebook y Twitter.

Twitter como se mencionó anteriormente es una red social donde un usuario, que previamente se ha registrado, puede enviar mensaje de texto sobre un tema que él considere importante o puede interactuar con otros mensajes dando a conocer su punto de vista; el texto enviado es conocido como tweets y tiene un máximo de 280 caracteres. En la página de Twitter, la empresa se define a sí misma con las siguientes palabras “*lo que está pasando en el mundo y los temas sobre los que está hablando la gente*.” (2). Un componente opcional en los tweets es que incluye su geolocalización, lo cual es una pieza importante en el análisis del comportamiento porque nos permite determinar la opinión de las personas por zonas geográficas.

En el presente documento se realizará un estudio de investigación sobre las opiniones que tienen la población limeña sobre las compra navideña en los centros comerciales del año 2018 en el mes de diciembre; se usará como base de datos la información de la red social Twitter, obteniéndola de 4 geolocalizaciones correspondiente a las denominadas 5 Limas: Lima Sur, Lima Norte, Lima Este y Lima Centro. Para tal fin se empleará el análisis de sentimiento para determinar las preferencias de los limeños, los tweets serán obtenidos en geolocalizaciones de las 5 Limas: Lima Sur, Lima Norte, Lima Este, Lima Centro y Lima Moderna; y para analizar los datos se usará la técnica de text mining, el cual es el proceso de obtener conocimiento implícito desde una información basado en texto.

La finalidad de esta investigación es poder contar con una herramienta de estudio de mercado basada en text mining que nos permite reducir en tiempo y costo con un estudio de mercado tradicional en el campo de marketing.

1. **JUSTIFICACIÓN DE LA INVESTIGACIÓN**

Actualmente no existen estudios sobre las preferencias de los limeños acerca de los centros comerciales usando técnica de Análisis de Sentimientos empleando las redes social Twitter como fuente de datos, esto se puede comprobar revisando el registro de trabajos de investigación del SUNEDU,

El estudio permitirá entender las opiniones que tienen los clientes sobre los centros comerciales durante la época navideña del mes de diciembre del 2018, agrupadas por su localización geográfica dentro de Lima, para tal fin se ha considerado las 4 divisiones que tiene Lima: Lima Sur, Lima Norte, Lima Este y Lima Centro

El estudio tiene como finalidad crear una aplicación informática que pueda ser usado por las empresas para desarrollar estrategias de mercado orientado a conocer a sus potenciales clientes.

1. **OBJETIVOS DE LA INVESTIGACIÓN**

Clasificar las opiniones de los consumidores aplicando análisis de sentimiento sobre los centros comerciales en las 5 Limas en el mes de diciembre del 2018 a través de los tweets generados por los usuarios.

Proporcionar elementos para que las empresas tomen decisiones y/o elaboren estrategias de marketing basado en la percepción de los usuarios en la red social Twitter.

1. **FORMULACIÓN DE LA HIPÓTESIS**

La hipótesis consiste en que es posible aplicar el análisis de sentimiento en la red social Twitter sobre las opiniones de la población limeña de los más importantes centros comerciales de la Ciudad para poder ser usado en estrategias de mercado.

1. **MARCO TEÓRICO**

.

Se definen las tecnologías y conceptos teóricos que se utilizarán en el desarrollo del presente trabajo.

* 1. **Twitter**:

Es una plataforma de servicio de microblogging que permite el envió de mensajes de texto en un tamaño máximo de 280 caracteres denominados tweets, los usuarios pueden subscribirse a los tweets de otros usuarios y hacerles seguimientos, esta acción es conocida como seguidores o followers. Por defecto los mensajes son públicos y pueden contener la zona geográfica de la persona que ha enviado el tweet.

|  |  |  |
| --- | --- | --- |
| **Atributo** | **Tipo** | **Descripción** |
| created\_at | String | Hora de creación |
| Id | Int64 | Identificador del tweet |
| text | String | El mensaje de texto. |
| source | String | Usado para postear un texto en HTML |
| user | User | Usuario que envie el mensaje de texto. |
| Coordinates | Coordinates | Coordenadas de la localización geográfica. |
| Retweeted | Boolean | Verifica si el mensaje ha sido reenviado. |
| Lang | String | Idioma del mensaje. |

Tabla 01 : Estructura de un tweet obtenido del la Empresa Twitter

* 1. **Data Mining**

El Data Mining es un campo de la estadística y las [ciencias de la computación](https://es.wikipedia.org/wiki/Ciencias_de_la_computaci%C3%B3n) relacionado con el proceso de descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza métodos de [inteligencia artificial](https://es.wikipedia.org/wiki/Inteligencia_artificial), [aprendizaje automático](https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico), [estadística](https://es.wikipedia.org/wiki/Estad%C3%ADstica) y sistemas de [bases de datos](https://es.wikipedia.org/wiki/Bases_de_datos)

La forma para aplicar Data Mining se basa en las siguientes etapas:

* Comprensión del Negocio
* Comprensión de los Datos
* Preparación de los datos
* Modelado
* Evaluación
* Desarrollo
  1. **Text Mining**

Es el análisis de información no estructurada, la cual se puede encontrar en redes sociales, para tal fin se emplea técnicas de lingüística, modelamientos estadísticos y técnicas de aprendizaje para descubrir conocimientos que no existen explícitamente en ningún texto de la colección, pero que surgen al relacionar el contenido de muchos de ellos.

Se suelen aplicar a encuestas de opinión, encuestas de satisfacción, libros de reclamación, etc.

La forma para aplicar Text Mining se basa en las siguientes etapas:

* Preparar texto para el análisis
* Extraer conceptos
* Aplicar el análisis de enlace de texto
* Construir categorías
* Desplegar modelos predictivos

Los beneficios del uso del text mining es que nos permite identificar hechos o datos puntuales a partir del texto de los documentos, agrupándolo en clustering e igualmente determinar el tema o temas tratados en los documentos mediante la categorización automática de los textos y crear redes de conceptos.

El text mining se puede aplicar en:

* Resumen automático de textos
* Detección de fraudes
* Tendencial electorales
* Análisis de Sentimiento
* Clasificación de textos.
  1. **Análisis de Sentimiento**

Es el uso del procesamiento del lenguaje natural, [análisis de texto](https://es.wikipedia.org/wiki/Miner%C3%ADa_de_textos) y [lingüística computacional](https://es.wikipedia.org/wiki/Ling%C3%BC%C3%ADstica_computacional) para identificar y extraer información [subjetiva](https://es.wikipedia.org/wiki/Subjetividad) de los recursos. Desde el punto de vista de la [minería de textos](https://es.wikipedia.org/wiki/Miner%C3%ADa_de_textos), el análisis de sentimientos es una tarea de clasificación masiva de documentos de manera automática, en función de la connotación positiva o negativa del lenguaje ocupado en el documento. Es importante mencionar que estos tratamientos generalmente "se basan en relaciones estadísticas y de asociación, no en análisis lingüístico".

1. **METODOLOGÍA**

Según Medhat, las principales técnicas de análisis de sentimiento se dividen en dos grandes grupos: las que se basan en aprendizaje automático (machine learning approach) y las que se basan en diccionarios (lexicon-based approach).

La siguiente figura muestra las 2 técnicas:

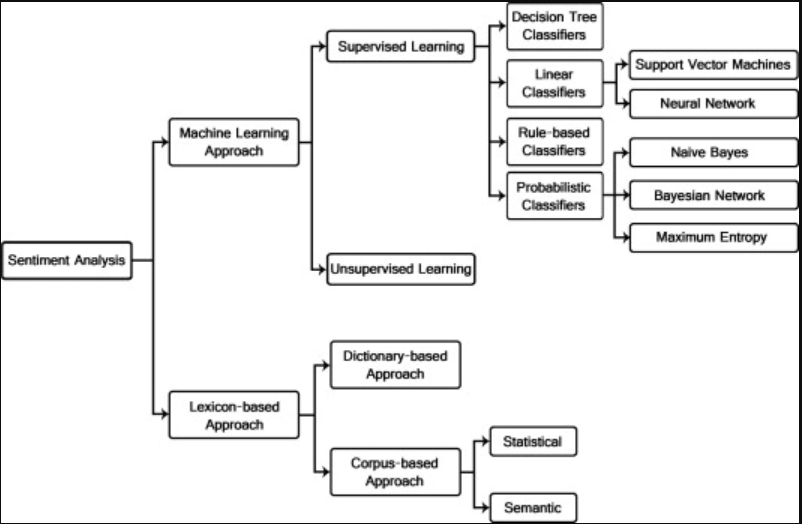


Figura 01 : Técnicas de Analisis de Sentimiento según Medat

Una vez mostrado los métodos de análisis, en la presente investigación se usará la estrategia de los algoritmos de aprendizaje automático que tiene relación con el campo de la informática y más en concreto, de la inteligencia artificial.

El procesamiento de datos es a través del aprendizaje para tal fin se extrae patrones de comportamiento a partir de las entradas recibidas, y en base a dicha información aprendida o asimilada, realice la evaluación de nuevas entradas. Los algoritmos internos que constituyen la base de este aprendizaje tienen un fuerte componente estadístico y algebraico, con la consiguiente capacidad de cálculo.

1. **CRONOGRAMA**

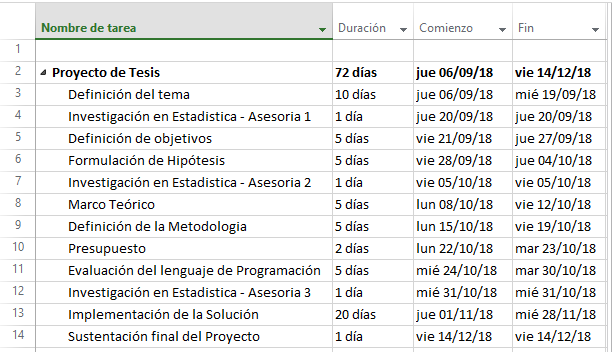
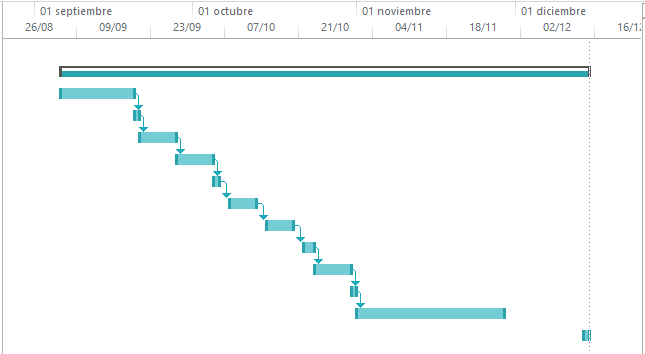


Figura 02 : Cronograma de Trabajo

Figura 03 : Gráfico del Cronograma de Trabajo

1. **PRESUPUESTO**

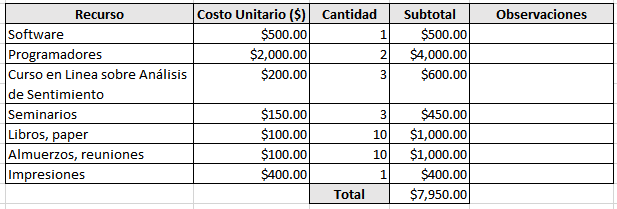


Tabla 02 : Presupuesto del Proyecto de Tesis

1. **ANEXO**

Script para obtener los datos de Twitter basado en un patrón de búsqueda

|  |
| --- |
| ##########################################################  # SCRIPT PARA OBTENER DATOS DE TWITTER  # BASADOS EN UN PATRON DE BUSQUEDA  ##########################################################  # install rtweet from CRAN  # install.packages("rtweet")  # install.packages("tidyverse")  ## load rtweet package  library(rtweet)  library(httr)  library(tidyverse)  #  twitter\_token <- create\_token(  app = "InstituteOfLima3",  consumer\_key = "ejXoPRrKQOXWlIr7BlumWCCxD",  consumer\_secret = "y4nSiBnWfm2y0I7Od2hYuIiGvNMvfhz28ZqwCpPMQxGyZIphmX",  access\_token = "200779440-zUuSrp11W6JWtb2jygPU9fPMvJ9Rrx5E2e8fqK7t",  access\_secret = "YdOIG8etShZXQ5HzkHtZzTyme3nb3tHv41Seyo6W8zjxO")  identical(twitter\_token, get\_token())  #########################################################3  # LIMA MODERNA  #########################################################3  SEARCH\_PATTERN <- "navidad OR sagafalabella OR ripley OR oechsle OR gamarra OR regalos OR compras"  #############################################3  # LIMA MODERNA  # LUGAR : COLISEO DIBOS  # DISTRITO : SURQUILLO  # COORDENADAS : -12.111982,-77.000294  ##############################################  lm <- search\_tweets(  SEARCH\_PATTERN, geocode = "-12.111982,-77.000294,5mi", n = 3000 , retryonratelimit = TRUE  )  lm[,"LIMA"] <- "MODERNA"  #############################################3  # LIMA SUR  # LUGAR : ESTACION DEL METRO FINAL  # DISTRITO : VILLA SALVADOR  # COORDENADAS : -12.206978,-76.933511  ##############################################  ls <- search\_tweets(  SEARCH\_PATTERN, geocode = "-12.206978,-76.933511,10mi", n = 3000 , retryonratelimit = TRUE  )  ls[,"LIMA"] <- "SUR"  #############################################3  # LIMA ESTE  # LUGAR : OVALO SANTA ANITA  # DISTRITO : SANTA ANITA  # COORDENADAS : -12.054857,-76.964667  ##############################################  le <- search\_tweets(  SEARCH\_PATTERN, geocode = "-12.054857,-76.964667,5mi", n = 3000 , retryonratelimit = TRUE  )  le[,"LIMA"] <- "ESTE"  #############################################3  # LIMA NORTE  # LUGAR : MEGA PLAZA NORTE  # DISTRITO : INDEPENDENCIA  # COORDENADAS : -11.994278,-77.061239  ##############################################  ln <- search\_tweets(  SEARCH\_PATTERN, geocode = "-11.994278,-77.061239,5mi", n = 3000 , retryonratelimit = TRUE  )  ln[,"LIMA"] <- "NORTE"  #############################################3  # LIMA CENTRO  # LUGAR : GAMARRA  # DISTRITO : LA VICTORIA  # COORDENADAS : -12.065656,-77.012336  ##############################################  lc <- search\_tweets(  SEARCH\_PATTERN, geocode = "-12.065656,-77.012336,5mi", n = 3000 , retryonratelimit = TRUE  )  lc[,"LIMA"] <- "CENTRO"  lima <- rbind(lm,ls,le,ln,lc)  save\_as\_csv(lima, "resultados-lima-navidad.csv", prepend\_ids = TRUE, na = "",  fileEncoding = "ISO-8859-1") |

1. **REFERENCIAS**

Taeho Jo-Text, [2018], Mining Concepts, Implementation, and Big Data Challenge-Springer

Bing Liu, [2011], Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data. Second Edition

Hassan, A. - Korashy, H., [2014], Medhat, W.“Sentiment analysis algorithms and applications: A survey”, Ain Shams Engineering Journal, vol. 5, no 4, pp. 1093–1113.

Francois Chollet , [2017], Deep Learning with Python-Manning Publications, pp. 178-223