UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

PROYECTO DE TESIS :

PERCEPCIÓN DEL CONSUMIDOR LIMEÑO SOBRE LAS TIENDAS POR DEPARTAMENTOS EN LA RED SOCIAL TWITTER USANDO ANÁLISIS DE SENTIMIENTOS

Presentado por
Jaime Gómez Marín

Docente
Ph.D. Augusto Bernuy Alva

La Molina, 2019

Índice general

Índice de figuras

Índice de cuadros

Capítulo 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.

La situación en el Perú no es ajena, de acuerdo al estudio de mercado realizado por la consultora CPI (Compañia Peruana de Estudios de Mercados y Opinión Pública S.A.C.) sobre el uso de redes sociales en Lima de

agosto del 2018, como se aprecia en la tabla [3] , se evidencia el uso de Facebook como la red social mas popular con el 72.7 %, seguido de WhatsApp al 68.5 %, Instragram al 25.0 % y como cuarto puesto a Twitter con el 12.7 %.

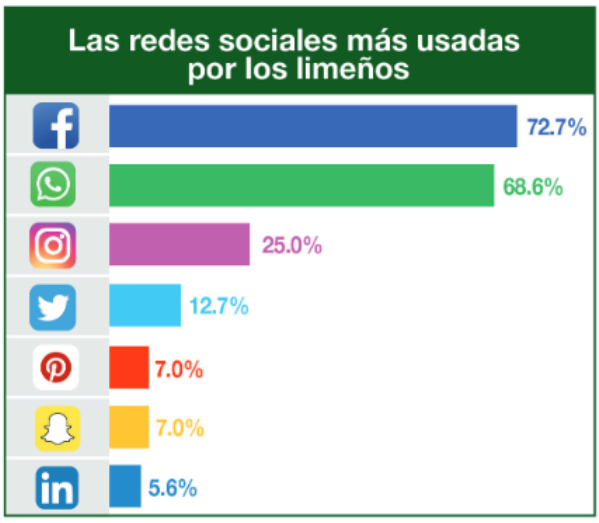

Figura 1.1: Redes Sociales mas usada por los limeños - Agosto 2018 - Fuente : CPI

El uso de la redes sociales por generación es mostrado a continuación:

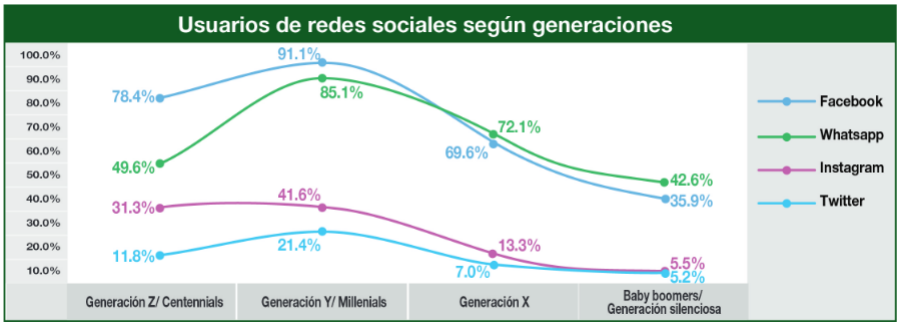

Figura 1.2: Uso de redes sociales según generación - Agosto 2018 - Fuente : CPI

Capítulo 2

Justificación de la investigación

Actualmente no existen estudios sobre las preferencias de los limeños sobre las tiendas por departamentos empleando técnica de Análisis de Sentimientos aplicados a la red social Twitter, esto se ha verificado al realizar una búsqueda en el banco de datos de trabajos de investigacón del SUNEDU [4].

El estudio permitirá entender las opiniones que tienen los consumidores sobre los centros comerciales 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 [5].

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 consumidores.

Capítulo 3

Objetivos de la investigación

3.1. Objetivos Generales

Clasificar las opiniones de los consumidores aplicando análisis de sentimiento sobre la compra realizado por los limeños en las tiendas por departamento a través de Twitter con la finalidad de que ser usado en estrategias de mercado.

3.2. Objetivos Especificos

- Evaluar plataforma de software de código libre que sirvan para el procesamiento de datos .
- Recopilar las opiniones más comunes del comprador Limeño y los periodos cuando realiza compras en tiendas por departamentos.
- Determinar formas óptimas de obtener datos de la red social Twitter.
- Evaluar herramientas de Machine Learning para al Análisis de Sentimiento.
- Presentar los resultados por zonas geográficas de Lima Metropolitana y compararlo con estudios de mercados realizados

Capítulo 4

Formulación de 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 tiendas por departamento de la ciudad de Lima para poder ser usado en estrategias de mercado.

Capítulo 5

Marco Teórico

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

5.1. Twitter

Es una plataforma de servicio de microblogging [2] 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.

A continuación se muestra la estructura de un 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 envia el mensaje de texto.
Coordinates	Coordinates	Coordenadas de la localización geográfica.
Retweeted	Boolean	Verifica si el mensaje ha sido re-enviado.
Lang	String	Idioma del mensaje.

Cuadro 5.1: Estructura de un tweet obtenido de la Empresa Twitter

5.2. Data Mining

El Data Mining [6] es un campo de la estadística y las ciencias de la computación relacionado con el proceso de descubrir patrones en grandes volúmenes de conjuntos de datos . Utiliza métodos de inteligencia artificial [9], aprendizaje automático, estadística y sistemas de 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

5.3. Text Mining

Es el análisis de información no estructurada [7] , 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.

5.4. Análisis de Sentimiento

Es el uso del procesamiento del lenguaje natural [8], análisis de texto y lingäística computacional para identificar y extraer información subjetiva de los recursos. Desde el punto de vista de la minería 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".

5.4.1. Técnicas de Análisis de Sentimiento

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:

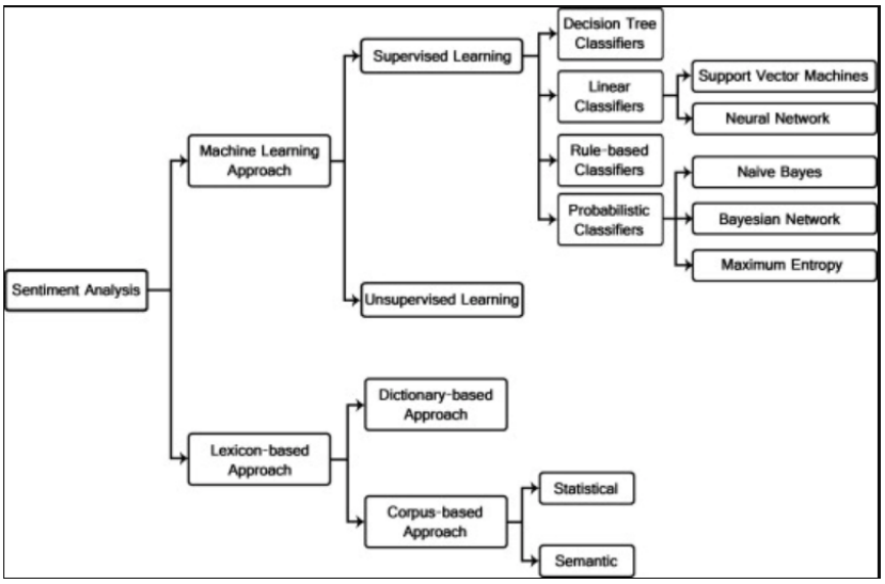

Figura 5.1: Técnicas de Análisis de Sentimiento según Medat

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.

Capítulo 6

Metodología

En la presente Proyecto de Tesis se ha decidido usar la metodología CRISP-DM [10] (Cross-industry standard process for data mining). Esta metodología es ampliamente usada en minería de datos y es empleado por expertos en esta materia.

La metododologá consiste en dividir el proceso de data minining en 6 fases las cuales se mencionan:

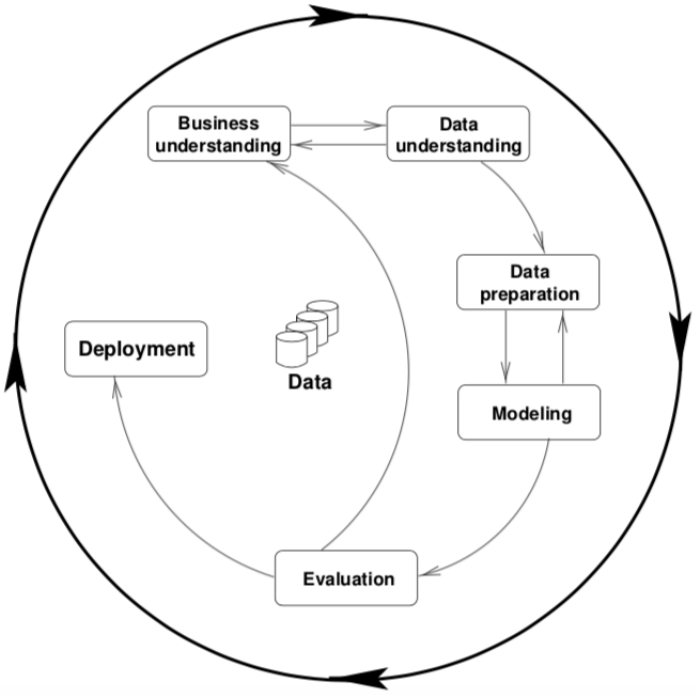

Figura 6.1: Fase del proceso de modelo CRISP-DM

- Entender el negocio
- Entender los datos
- Preparar los datos
- Modelar
- Evaluar
- Desplegar

6.1. Entender el negocio

En este apartado se define que debemos conocer claramente el negocio sobre el cuás estamos analizando los datos

6.2. Entender los datos

Comprender la naturaleza del datos, identificando nuestra variables independientes, en este apartado es importante identificar la fuentes de los datos

6.3. Preparar los datos

Es un mecanismo donde aplicamos técnicas para mitigar los datos faltantes , son las conocidas como técnicas de clean data

6.4. Modelar

Con los datos listos para ser procesados, se realiza una análisis descriptivo y elegimos un modelo de acuerdo a la naturaleza de los datos.

6.5. Evaluar

Evaluamos los datos obtenidos por el modelo con los datos reales y verificamos si nuestro modelo es el correcto

6.6. Desplegar

Una vez obtenido el modelo más parsimonioso procedemos al despliegue del modelo en un entorno productivo para su uso por los usuarios finales.

Capítulo 7

Cronograma

A continuación se presenta el diagrama de Gantt del Proyecto de Tesis y de la Tesis, se ha considerado el uso de la metodología CRISP-DM.

	Nombre	Duracion	Inicio	Terminado	Predecesores
1	**Proyecvto de Tesis**	75 days?	**08/04/19 08:00 AM**	**19/07/19 05:00 PM**	
2	Definciòn del Tema	14 days	08/04/19 08:00 AM	25/04/19 05:00 PM	
3	Definiciòn de los objetivos	14 days	26/04/19 08:00 AM	15/05/19 05:00 PM	2
4	Formulación de la hipotesis	10 days	16/05/19 08:00 AM	29/05/19 05:00 PM	3
5	Marco Teòrico	18 days	30/05/19 08:00 AM	24/06/19 05:00 PM	4
6	Definir Metodologia de trabajo	10 days	25/06/19 08:00 AM	08/07/19 05:00 PM	5
7	Realizar el Preaupuesto	8 days	09/07/19 08:00 AM	18/07/19 05:00 PM	6
8	Sustentaciòn	1 day?	19/07/19 08:00 AM	19/07/19 05:00 PM	7
9	**Tesis**	108 days	**22/07/19 08:00 AM**	**18/12/19 05:00 PM**	**1**
10	**Entender el negocio**	28 days	**22/07/19 08:00 AM**	**28/08/19 05:00 PM**	
11	Revisar estudios de preferencia de consumidores	21 days	22/07/19 08:00 AM	19/08/19 05:00 PM	
12	Capacitaciòn en Anállisis de Sentimiento	28 days	22/07/19 08:00 AM	28/08/19 05:00 PM	
13	**Entender los datos**	35 days	**29/08/19 08:00 AM**	**16/10/19 05:00 PM**	**10**
14	Evaluar la plataforma de desarrollo de la soluciòn	7 days	29/08/19 08:00 AM	06/09/19 05:00 PM	
15	Seleccionar el lenguaje de programaciòn	7 days	09/09/19 08:00 AM	17/09/19 05:00 PM	14
16	Obtener una cuenta de desarrollo de Twitter	7 days	09/09/19 08:00 AM	17/09/19 05:00 PM	14
17	Obtener los datos de Twitter	21 days	18/09/19 08:00 AM	16/10/19 05:00 PM	15;16
18	**Preparar los datos**	14 days	**17/10/19 08:00 AM**	**05/11/19 05:00 PM**	**13**
19	Usar tècnica de data clean	14 days	17/10/19 08:00 AM	05/11/19 05:00 PM	
20	**Modelar**	14 days	**06/11/19 08:00 AM**	**25/11/19 05:00 PM**	**18**
21	Evaluar modelos de Machine Learning	14 days	06/11/19 08:00 AM	25/11/19 05:00 PM	
22	**Evaluar**	7 days	**26/11/19 08:00 AM**	**04/12/19 05:00 PM**	**20**
23	Usar validación cruzada en equipos con GPU	7 days	26/11/19 08:00 AM	04/12/19 05:00 PM	
24	**Desplegar**	10 days	**05/12/19 08:00 AM**	**18/12/19 05:00 PM**	**22**
25	Preparar la aplicaciòn en la web	10 days	05/12/19 08:00 AM	18/12/19 05:00 PM	

Figura 7.1: Cronograma del Proyecto de Tesis y la Tesis

Capítulo 8

Presupuesto

Respecto al presupuesto, el valor estimados es de USD 12,450 conforme se presenta en la tabla 8.1. Los principal costo corresponde a la adquisición de las computadoras, software y pagos a desarrolladores, con un total de 72 %. Le siguen la investigaciones y capacitaciones en técnicas de análisis de sentimiento que son aproximadamente el 17 % , resaltar el rubro correspondiente a capacitaciones online en plataforma de e-learning de universidades extranjeras en temas de análisis de sentimiento. Finalmente el 10 % para los gastos de reuniones, insumos y empaste de tesis.

Recursos	**Costo Unit.($)**	**Cant.**	**Subtotal ($)**	**Observaciones**
Computadora	$1500	2	$3000	Computadoras iCore 7 con tarjeta gráfica NVIDIA GTX 1060
Software	$2000	1	$2000	Se usará Python v 3.7, Servidor en la nube (AWS)
Programadores	$2000	2	$4000	
Cursos online sobre Análisis de Sentimiento	$200	3	$600	
Seminarios	$150	3	$450	
Libros, papers	$100	10	$1000	
Reuniones y gastos diversos	$100	10	$1000	
Impresiones	$400	1	$400	
—	—	**Total**	$12450	-

Cuadro 8.1: Presupuesto del Proyecto de Tesis

Bibliografía

[1] Moisés Naím. *Ni Facebook, ni Twitter: son los fusiles* https://elpais.com/diario/2011/02/27/internacional/1298761206_850215.html , 2011.

[2] Twitter. *web page* https://twitter.com/?lang=es , 2019.

[3] CPI : Compañia Peruana de Estudios de Mercado y Opinión Pública s.a.c.. *Market Report : Lima Digital* http://cpi.pe/images/upload/paginaweb/archivo/26/MR_Limadigital2018.pdf , pp3 , 01/2019.

[4] SUNEDU *Registro de trabajos de Investigación* Bùsqueda : twitter analisis sentimientos. http://renati.sunedu.gob.pe/simple-search , 2019.

[5] INEI *Una mirada a Lima Metropolitana* https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1168/libro.pdf , pp 9 , sept. 2014.

[6] Taeho Jo-Text. *Mining Concepts, Implementation, and Big Data Challenge*. Springer, pp 3, 2018.

[7] Bing Liu. *Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data. Second Edition*. Springer, pp 17, 2017.

[8] Hassan, A. - Korashy, H. *"Sentiment analysis algorithms and applications: A survey",Ain Shams Engineering Journal*. Springer, vol. 5, no 4, pp 1093–1113, 2017.

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

[10] Daniel S. Putler , Robert E. Krider *Customer and Business Analytics*. CRC Press, pp 19, 2012.

[11] Abid Hussain , Ravi K. Vatrapu *Social Data Analytics Tool*. https://www.researchgate.net/publication/295257564_Social_Data_Analytics_Tool_SODATO. Conference Paper , 2014.

[12] Niels Buus Lassen, Ravi Vatrapu, Lisbeth la Cour, Rene Madsen, and Abid Hussain *Towards a Theory of Social Data: Predictive Analytics in the Era of Big Social Data*. https://research-api.cbs.dk/ws/portalfiles/portal/55458783/niels_buss_lassen_towards_a_theory_of_social_data_pu January 2018.

[13] Zahra Iman, Scott Sanner, Mohamed Reda Bouadjenek, Lexing Xie *A Longitudinal Study of Topic Classification on Twitter*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/download/15625/14852. (ICWSM 2017

[14] Niels Buus Lassen , Lisbeth la Cour, and Ravi Vatrapu *Predictive Analytics with Social Media Data*. The SAGE Handbook of Social Media Research Methods , pp 328 , 2016

[15] Carrie Winterer *Thesis : Predicting Twitter TIme Series Using Generalized Linear Models*. Case Western Reserve University School of Graduates Studies - EEUU. Department of Mathematics, Applied Mathematics and Statistics , August, 2018

[16] Manuel Alejandro Rodríguez Santana *Thesis : Predicción del éxito de los mensajes de Twitter*. Máster en Ingeniería Informática, Facultad de Informática, Universidad Complutense de Madrid , Junio, 2018

[17] Xinsong Du , Jiang Bian and Mattia Prosperi *An Operational Deep Learning Pipeline for Classifying*. Dept. of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL 32611, USA 2 Dept. of Epidemiology, University of Florida, Gainesville, FL 32611, USA