Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

E-Commerce Text Classification

PLN -> Naive Bayes, SVM, Decision Trees (XGB)

Chaparro Sicardo Tanibeth, Malváez Flores Axel Daniel

Project #1

2023

Machine Learning

Introducción

Contexto

El comercio en México tanto al por mayor como al por menor se encuentra dentro del Top 3 de las actividades que más impacto tienen en el crecimiento del PIB de nuestro país. [Sta23] A lo largo de la historia, el comercio ha experimentado una evolución constante y radical. Durante la época del Porfiriato, el comercio en la iniciativa privada estaba centralizado y era autoritario, favoreciendo a las grandes compañías y la inversión extranjera. Durante la Revolución se buscaba equidad y una mayor justicia social, pero la inestabilidad política y económica limitaba su crecimiento. En la actualidad, en la era de la globalización, las empresas mexicanas han tenido que adaptarse a la creciente competitividad mediante la adopción de procesos como el control de calidad y una gestión eficiente.[Bar+23] La llegada del internet ha revolucionado el comercio, dando origen al "*e-commerce*" o comercio electrónico, que involucra transacciones tanto físicas como virtuales.[Inv23]

Descripcion problema

En los últimos años la popularidad del comercio virtual se ha mantenido a la alza, pues únicamente en 2021 las ventas mundiales por internet sumaron aproximadamente 5.2 billones de dólares, y se prevé que esta cifra se eleve un 56% para 2026.[Ste22] Por lo anterior, es de esperarse que el número de compradores por internet aumente, y con ellos los datos generados por los usuarios. Con el propósito de optimizar la experiencia de compra de los usuarios, es de suma importancia ofrecer recomendaciones de productos personalizadas basadas en patrones encontrados en las búsquedas de los clientes. Es por ello que necesario conocer los tipos de productos con los que el usuario ha interactuado. Lo anterior presenta un nuevo desafío, pues el aumento de productos de venta en línea dificulta la tarea de categorizar los productos y realizar una clasificación manual es sumamente ineficiente y propensa a errores. La implementación de sistemas automatizados de clasificación, respaldados por algoritmos de *machine learning* y PLN, ofrece una solución efectiva a este problema.

Motivacion

Desde un punto de vista empresarial, clasificar productos basados en su descripción genera un impacto en la experiencia del usuario y el rendimiento comercial. Al implementar un sistema de clasificación eficiente, podremos ofrecer a los usuarios recomendaciones personalizadas y precisas, lo que mejorará significativamente su satisfacción. No obstante, a un nivel personal nuestra **motivación** principal es poder trabajar en un proyecto que involucre un tratamiento especial de los datos (en este caso texto) pues es un reto que hasta la fecha no hemos tenido.

Estado del arte

Los modelos de *Machine Learning* y técnicas de PLN han sido utilizados con diferentes enfoques, entre los cuales están:

- Clasificación de opiniones de clientes: Estos proyectos utilizan técnicas de PLN para realizar análisis de sentimientos a las opiniones de clientes con el fin de determinar si una opinión es positiva o negativa. Los modelos de clasificación usados son Naive Bayes y Support Vector Machines. [Emr22]
- Prevención de abandono de clientes: Debido al alto costo de obtener nuevos clientes, la retención de clientes es esencial para las empresas. El objetivo de estos proyectos es realizar *churning analysis*. Los modelos de Support Vector Machines son populares en este tipo de proyectos. [Sho23]
- Predicción de ventas: Las empresas necesitan generar pronósticos de ventas precisos, que les permitan predecir las ventas de sus productos con meses de anticipación y tomar decisiones

acertadas. Entre los modelos usados para estos proyectos están Random Forest, Gradient Boosting Trees y Redes Neuronales. [Kar20]

No obstante, es importante conocer del lado de la academia cómo se encuentra la clasificación de texto en esta área del comercio. Es por ello que nos dimos a la tarea de investigar varios artículos en donde su tarea principal fuera similar, los resultados obtenidos fueron los siguientes:

- *A Real-World Text Classification Application for an E-commerce Platform* : Este artículo presenta un estudio sobre el uso de algoritmos de clasificación de texto para la categorización de productos en una plataforma de comercio electrónico. Los datos utilizados fueron recopilados de una plataforma de e-commerce de moda. Se entrenaron los modelos de *ML* Multinomial Naïve Bayes, Complement Naive Bayes, Linear Support Vector Classifier, Stochastic Gradient Descent, Ridge Classifier. El algoritmo Linear Support Vector Classifier (LinearSVC) logró la tasa de precisión media más alta de 96,08%. Se utilizó Stemming y CountVectorizer como un preprocesamiento destacable. [Yıl+19]
- *A FRAMEWORK FOR PRODUCT DESCRIPTIONCLASSIFICATION IN E-COMMERCE* : En este artículo se presenta un *framework* para la clasificación de descripciones de productos en el *e-commerce*. Se basa en un modelo de machine learning que utiliza una variedad de características, incluyendo el vocabulario, la sintaxis y la estructura de las descripciones. En la parte del preprocesamiento se concluyó que *information gain* (IG) y χ^2 *statistics* (CHI) abarcan hasta el 90% del espacio original de características o términos (sin pérdida de *accuracy*). Se utilizó un modelo de SVM con un accuracy final del 90%. [VFK17]
- *GoldenBullet: Automated Classification of Product Data in E-commerce* : En este artículo el objetivo es aplicar técnicas de PLN y Machine Learning para clasificar productos online. Dentro del preprocesamiento de datos, se utilizó *Stemmer* y *Tagged*. Remoción de *stopwords*, etc. Los SVM fueron los que alcanzaron un accuracy mayor al 90%.
- *Sentiment Analysis and Product Review Classification in E-commerce Platform* : Dentro de este artículo, si bien la clasificación no fue dentro de categorías de *e-commerce*, se realizó un análisis de sentimientos con redes neuronales. De este artículo es importante destacar el procesamiento utilizado, el cuál remueve *stopwords*, puntuaciones y caractéres especiales, finalmente para la creación de vectores se utilizó **fastText** que implementa un algoritmo preentrenado para vectorizar texto.[MRB20]

Justificacion

Nuevamente desde un punto de vista empresarial, esta solución de clasificación no solo mejorará la experiencia del usuario, sino que también proporcionará a nuestra empresa una posición estratégica en el mercado logrando la capacidad de optimizar nuestras operaciones. Esto nos genera un crecimiento sostenible a largo plazo logrando traducirse en mayores ingresos.

Por otro lado, de nuevo desde un punto de vista personal, este proyecto nos permitirá poner en práctica técnicas de Procesamiento de Lenguaje Natural (las cuáles estaremos a la par adquiriendo en otra materia) y Machine Learning en un contexto con datos reales. Enriqueciendo nuestro portafolio de proyectos (el cuál siempre elegimos problemas distintos, para conocer de diversas áreas). Además, como ya se vio la clasificación de texto en *e-commerce* es un desafío actual y relevante en el mundo empresarial, lo que nos permitirá demostrar nuestras habilidades de *problem-solving* y aumentará nuestra experiencia en proyectos al momento de afrontar el mundo laboral.

Definición del problema

Con el crecimiento exponencial de los datos y de las ventas en línea, se espera que el usuario pueda recibir ofertas personalizadas en función de sus compras y búsquedas en el *search engine (Google, Yahoo, Bing, etc)*, no obstante para poder realizar estas ofertas personalizadas es necesario conocer lo que ha buscado nuestro cliente y si corresponde a comprar algún producto, de ser así deseamos saber qué tipo de producto fue el que buscó y por tanto ofrecer dicho producto o productos similares, para esto necesitamos conocer cuáles son dichos productos y su tipo de segmento. Esta última tarea es la que atacaremos proponiendo un clasificador de texto correspondiente a los productos de venta en línea y su descripción, esto será de gran ayuda para no tener que leer y clasificar cada descripción de productos que tengamos (lo cuál sería bastante ineficiente).

Objetivos

`Objetivo general`

El objetivo de este proyecto es desarrollar un modelo que automatice el proceso de categorización de productos de venta en línea a partir del análisis de su descripción. Con lo anterior se busca brindar una experiencia óptima a los usuarios en línea al ofrecerles recomendaciones de compra personalizadas. Además del hecho de tener éxito en este nuevo reto que estamos tomando y el cuál iremos aprendiendo a la par que lo vamos desarrollando.

`Objetivos especificos`

- Procesamiento del Lenguaje Natural: Aplicación de técnicas de PLN para la extracción de información relevante de las descripciones de los productos. Generación de etiquetas. Extracción de características y vectorización.
- Selección del algoritmo: Elegir el mejor algoritmo de clasificación de textos para nuestra tarea. Los algoritmos con los que se trabajará son Naive Bayes, Support Vector Machines y XGBoost.
- Entrenamiento, validación y pruebas: Utilizar el conjunto de datos para entrenar el modelo seleccionado. Realizar múltiples pruebas para evaluar la precisión del modelo y a partir de los resultados realizar los ajustes necesarios al modelo.

Contribuciones

En términos empresariales, la automatización de la clasificación permitirá a las empresas realizar decisiones informadas, pues les ayudará comprender las preferencias de los usuarios y con ellas personalizar las recomendaciones de productos, lo que se espera afecte positivamente a las ventas. Al mismo tiempo, se espera que estas recomendaciones mejoren la experiencia de compra de los usuarios.

Desde un punto de vista académico, realizar un buen y acertado preprocesamiento del texto esperamos que incremente significativamente la precisión de los modelos. Hasta el momento, la mayoría de artículos no realizan un gran preprocesamiento de texto, esta es un área de oportunidad para nosotros debido a que incluyendo lo básico como remoción de *stopwords*, caractéres especiales, *lowercase*. Utilizaremos stemming y lematización (comparar entre sí). Posteriormente la vectorización es un punto importante debido a que no siempre se menciona y

en este caso haremos matriz TF-IDF y algunos modelos embebidos (por definir entre Word2Vec o Glove). Hasta este punto tendremos en el caso de la matriz TF-IDF posiblemente de una alta dimensión, es por esto que aplicaremos las pruebas *Information Gain* y χ^2 *statistic*, las cuales nos ayudarán a seleccionar las características (palabras o términos) más informativas para la tarea de clasificación. Finalmente, del lado de *Machine Learning*, haremos un benchmark para saber cuáles de los algoritmos propuestos tienen un mejor tiempo-accuracy y tratar de encontrar el que más nos convenga en función de los recursos con los que contamos.

Metodología

Recoleccion y Preparacion de Datos

Para este proyecto se utilizará un conjunto de datos obtenido de la plataforma Kaggle [Sha19] que contiene la descripción de productos pertenecientes a las categorías *Electronics*, *Household*, *Books* y *Clothing & Accesories*. Los datos se encuentran en formato ".csv" y contienen dos columnas, una columna de texto con la descripción de los productos y la segunda la etiqueta de su categoría. La exploración inicial indica que en total se cuenta con 50,425 registros.

Extraccion de Caracteristicas usando PLN

Previo al entrenamiento del modelo es necesario extraer la información relevante de las descripciones de los productos. Este proceso se puede dividir en dos secciones.

1. *Preprocesamiento de texto.* Esta sección consiste en segmentar y estandarizar nuestro texto. El primer paso del preprocesamiento consiste en la tokenización; es decir, separar el texto por segmentos de palabras. Posteriormente se realiza la transformación de palabras a minúsculas. Por último, se deben eliminar *stop-words* y símbolos de puntuación, pues estos no proporcionan información relevante para el análisis del texto.

2. *Extracción de características.* Debido a que los algoritmos de aprendizaje automático no pueden trabajar directamente con el texto, se emplean diferentes métodos de extracción que transforman el texto a matrices o vectores de características. Entre estos métodos se encuentran TF-IDF vectorizer y Word2Vec o algún modelo embebido. [Chr09]

3. *Selección de Características.* Utilizaremos las pruebas antes mencionadas para seleccionar aquellas características o términos más relevantes y reducir el costo computacional que requerirán nuestros algoritmos de *Machine Learning* para aprender.

Desarrollo de Modelos de Aprendizaje Automatico

Se explorarán diferentes modelos de aprendizaje automático que pueden ser usados para la clasificación de textos. Para este proyecto nos centraremos en los siguientes modelos.

1. *Naive Bayes Classifier.* En estos clasificadores, cada característica cuenta para determinar la etiqueta asignada a la entrada. Se comienza calculando la probabilidad posterior de cada etiqueta a partir de la frecuencia en el conjunto de entrenamiento. La contribución de cada característica se combina con su probabilidad posterior para llegar a la estimación de verosimilitud de cada una. A la entrada se le asigna la etiqueta con la estimación más alta. [Ste09]

2. *XGBoost.* XGBoost es un algoritmo de boosting que utiliza múltiples árboles de decisión. Cada árbol en XGBoost se construye para corregir los errores del anterior, lo que permite capturar patrones más complejos en los datos. XGBoost tiene una mayor capacidad de aprendizaje en comparación con un único árbol de decisión simple. Un árbol simple consiste en nodos de decisión, los cuales verifican el valor de las características, y nodos hoja, que

asignan etiquetas. Además utiliza optimizaciones como la paralelización y el corte temprano de nodos para acelerar el proceso de entrenamiento y predicción. [CG16]

3. *Support Vector Machines.* Este clasificador define el criterio a considerar para que una superficie de decisión esté lo más alejada posible a cualquier punto de datos. La distancia entre la superficie de decisión y el punto más cercano determina el margen del clasificador. La función de decisión depende usualmente por un pequeño conjunto de datos que define la posición de la frontera, estos puntos son llamados vectores de soporte. [Chr09]

Entrenamiento, Validacion y Evaluacion

Se utilizará un conjunto de entrenamiento para entrenar los modelos mencionados en la sección anterior. Se evaluará la precisión y eficiencia de los clasificadores usando métricas como precisión, recall y F1-score. Los parámetros se ajustarán dependiendo de los resultados de las métricas de evaluación.

Cronograma

TAREAS	FECHA DE INICIO	FECHA DE ENTREGA	% AVANCE
Definir el problema que queremos resolver	21/08/23	25/08/23	100%
Obtener los datos que necesitamos	21/08/23	25/08/23	100%
Propuesta del Proyecto Final	28/08/23	31/08/23	100%
Presentar la Propuesta	05/09/23	05/09/23	0%
Preparar y explorar los datos	11/09/23	17/09/23	0%
Limpieza (stopwords, caractéres especiales)	18/09/23	24/09/23	0%
Stemming y Lemmatizer	25/09/23	30/09/23	0%
Convertir texto en vectores	25/09/23	01/10/23	0%
Comprender los algoritmos (SVM, XGBoost, NN)	02/10/23	22/10/23	0%
Crear algoritmos con datos de entrenamiento y prueba	02/10/23	22/10/23	0%
Fine-tuning de los algoritmos	23/10/23	29/10/23	0%
Comprender los resultados	30/10/23	05/11/23	0%
Escribir reporte final de actividades y resultados y una presentación	06/11/23	17/11/23	0%
Presentación y entrega del proyecto	20/11/24	24/11/23	0%

Bibliography

MexicoHistory [Bar+23] Barbosa, O. A. G. et al. '100 years of Management in Mexico, background and current trends'. In: *South Florida Journal of Development* vol. 4, no. 2 (Apr. 2023), pp. 835–843. URL: https://ojs.southfloridapublishing.com/ojs/index.php/jdev/article/view/2492/1979.

xgboost [CG16] Chen, T. and Guestrin, C. *XGBoost: A Scalable Tree Boosting System*. New York, NY, USA, 2016. URL: https://doi.org/10.1145/2939672.2939785.

inforetrieval [Chr09] Christopher D. Manning Prabhakar Raghavan, H. S. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

customerreview [Emr22] Emre Deniz Hasan Erbay, M. C. 'Multi-Label Classification of E-Commerce Customer Reviews via Machine Learning'. In: *Axioms* (2022).

investopedia [Inv23] Investopedia. *E-commerce Defined: Types, History, and Examples*. Accessed: 2023-08-29. 2023. URL: https://www.investopedia.com/terms/e/ecommerce.asp#:~:text=Electronic%20commerce%20or%20e-commerce%20%28sometimes%20written%20as%20eCommerce%29,all%20four%20of%20the%20following%20major%20market%20segments%3A.

salespred [Kar20] Karandeep Singh Booma P M, U. E. 'E-Commerce System for Sale Prefiction Using Machine Learning Technique'. In: *Journal of Physics: Conference Series* (2020).

9392710 [MRB20] Munna, M. H., Rifat, M. R. I. and Badrudduza, A. S. M. 'Sentiment Analysis and Product Review Classification in E-commerce Platform'. In: *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. 2020, pp. 1–6.

shahane2019 [Sha19] Shahane, S. *E-commerce text dataset*. Version 2. 2019. URL: https://doi.org/10.5281/zenodo.3355823.

churningcus [Sho23] Shobana J Rakesh Kumonar, J. B. 'E-commerce customer churn prevention using machine learning-based business intelligence strategy'. In: *Measurement: Sensors* (2023).

statista_pib [Sta23] Statista Research Department. *Aportación al PIB trimestral de las actividades económicas México [Gráfico]*. Recuperado el 29 de agosto, 2023, de. June 2023. URL: https://es.statista.com/estadisticas/585037/aportacion-al-pib-trimestral-de-las-actividades-economicas-mexico/#:~:text=Publicado%20por%20Statista%20Research%20Department%2C%201%20jun%202023,ambas%20actividades%20con%20una%20contribuci%C3%B3n%20de%20un%2010%2C1%25..

PLNpython [Ste09] Steven Bird Ewan Klein, E. L. *Natural Language Processing with Python*. O'Reilly, 2009.

statistaweb [Ste22] Stephanie Chevalier. *Global retail e-commerce sales 2014-2026*. Recuperado el 29 de agosto, 2023. 2022. URL: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

artcomm [VFK17] VANDIC, D., FRASINCAR, F. and KAYMAK, U. 'A FRAMEWORK FOR PRODUCT DESCRIPTION CLASSIFICATION IN E-COMMERCE'. In: *Journal of Web Engineering* vol. 17, no. 1-2 (Oct. 2017), pp. 001–027. URL: https://journals.riverpublishers.com/index.php/JWE/article/view/3233.

8946337 [Yıl+19] Yıldırım, F. M. et al. 'A Real-World Text Classification Application for an E-commerce Platform'. In: *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*. 2019, pp. 1–5.