project

Estefany Carolina Segura Linares
Brayan David Prieto Aya
Alejandro Jaramillo Vallejo

2025-09-08

Justificación de su importancia.

1. Definición del problema

1.3 Preguntas de investigación o hipótesis a responder.

Si las empresas consideran la presencia de sus servicios, para mejorar su rendimientos recurrirán a reseñas spam para aumentar sus ventas de prestación de servicios.

¿Existe una relación entre la edad y la cantidad de reseñas que publica una persona con respecto a un producto?.

Si la mayoría de usuarios del mundo están conectados a windows, será el sistema operativo que tenga más reseñas de carácter spam.

¿Existe algún tipo de relación de las reseñas de carácter spam con las aplicaciones más importantes?.

¿La dispersión es muy grande tanto para las reseñas verdaderas como de spam?.

2. Descripción del datataset

2.2 Variables disponibles y su significado

- **review_id**: Número identificativo de la reseña.
- **user_id**: Número identificativo del usuario.
- **app_name**: Nombre de la app desde donde se hizo la reseña.
- **app_category**: Categoría de la app desde donde se hace la reseña.
- **review_text**: Contenido de texto de la reseña.
- **review_language**: Lenguaje de la reseña.
- **rating**: Puntuación de la reseña.
- **review_date**: Fecha de publicación de la reseña.
- **verified_purchase**: Verificación de validez de compra.
- **device_type**: Tipo de dispositivo desde donde se hizo la reseña.
- **num_helpful_votes**: Número de votaciones de ayuda de la reseña por otros usuarios.
- **use_age**: Edad de usuario que realizo la reseña.
- **user_country**: País del usuario que hizo la reseña.
- **user_gender**: Género del usuario que hizo la reseña.
- **app_version**: Versión de la app desde donde se hizo la reseña.

3. Análisis Exploratorio de Datos (EDA) Básico

3.1.1 Importación de dataset.

```
data_reviews <- read.csv("multilingual_mobile_app_reviews_2025.csv")
```

3.1.3 Identificar cantidad de filas y columnas.

```
dim(data_reviews)
```

```
## [1] 2514   15
```

El dataset presenta 2514 fila con 15 columnas.

3.3.1 Contar cuántos valores faltantes hay por variable (sin necesidad de imputarlos todavía).

```
null_data <- colSums(is.na(data_reviews))
print(null_data)
```

```
##         review_id           user_id          app_name      app_category
##                 0                 0                 0                 0
##       review_text   review_language            rating       review_date
##                 0                 0                37                 0
## verified_purchase       device_type num_helpful_votes          user_age
##                 0                 0                 0                 0
##      user_country       user_gender       app_version
##                 0                 0                 0
```

Importación de librerías.

```
library(ggplot2)
library(dplyr)
library(tidyverse)
```

4.Revisión bibliográfica.

5.Plan de trabajo.

5.3 Herramientas y lenguajes de programación.

Lenguaje R

R como lenguaje cumple un factor decisivo para la realización del proyecto, sobre todo para el análisis de datos.Según su definición R es una sistema para la computación estadística y gráficas enfocado a un lenguaje de programación para generación de gráficas optimizadas y limpieza de errores[1].

Además de su diversidad de elementos que ofrece para el tratamiento de los datos según las necesidades del proyecto, consolidado por su vasto dominio de paquetes que tienen dominios como la estadística, aprendizaje automatizado, visualización de datos y análisis de series de tiempo definido bajo una red comprensiva de archivos (CRAN) con alrededor 18000 librerías desarrollada por la comunidad[2].

RStudio

Posterior al entendimiento del lenguaje R, destacamos que Rstudio como editor de código abierto, que nos permite escribir citaciones, visualizar la generación de código y el comportamiento de las variables de datos involucradas[3].

Con características destacables como su consola, editor de scripts, un ambiente de panel que permite observar las variables afectadas en tiempo real, panel de historia que muestra los comando ejecutados y un espacio de proyectos para guardar el trabajo en conjunto a la solución de un proyecto[2].

RMarkdown

Por último, para documentación del proceso de tratamiento de datos por la identificación de spam por medio de la ecuación Naive Bayes en reseñas, nos permite formatear las sustentaciones del código por medio de markdown para indicar la estética de los textos implícitos en los proceso escrito del proyecto, diferente a aplicaciones como word[4].

Facilitado por su portabilidad, y cambios de las aplicaciones de edición en código. RMarkdown es una herramienta de valor para el ámbito investigativo para la preservación de conocimiento en el tiempo con adaptación estandarizada que ofrece markdown para libros, tesis de universidades[4], etc.

6. Expectativas y retos.

6.1 Posibles dificultades y estrategias para resolverlas.

Dificultad: Poca claridad de palabras claves que nos permiten identificar un comportamiento de spam para una contraseña.

Estrategia: Investigación de proyectos previos que nos indiquen que palabras clases influyen en las reseñas podrían ser consideradas spam o no.

Dificultad: Presencia de datos nulos en la variable review_text que no faciliten el procesamiento de clasificación de las reseñas, considerado como un dato atípico que puede entorpecer los cálculos de probabilidad o verosimilitud de spam.

Estrategia: Limpieza de datos de datos que sean nulos, para no permitir el entorpecimiento del cálculo de palabras claves para Naive Bayes.

Dificultad: Presencia de otras variables que tengan relación con el comportamiento de spam presentado en el dataset.

Estrategia: Identificación relación entre variables por medio de modelos estadísticos complementarios que nos permitan definir relaciones con el spam.

Dificultad: Definir porcentaje de éxito de clasificación spam por medio de la ecuación Naive Bayes.

Estrategia: Definición de KPI o parámetro evaluables que midan el estado del arte del modelo para identificar spam.

Dificultad:Presencia de valores atípicos diferentes review_text que afecten la detección de spam.

Estrategia: Definición de procesos de análisis de relación(sea de cualitativo o cuantitativo) que definan su efecto en la generación de reseñas de spam.