Artículo 3 - Clasificador naïve Bayes

Barush Caliel C. Luna Diego B. Castillo
Pedro Emilio S. Rodríguez Vidal Alejandro G. López

2025-09-14

Abstract

This project develops a comprehensive web scraping pipeline to collect and organize firsthand paranormal experience narratives from Your Ghost Stories, an online repository of supernatural reports worldwide. Using R tools such as rvest, stringr, and purrr, we systematically extracted categorical metadata and full story texts by navigating the site's categorized and paginated structure. The methodology ensured legal compliance through robot.txt validation and incorporated robust error handling to overcome challenges related to URL parameter concatenation and page navigation. The resulting structured dataset enables advanced text mining and probabilistic classification models, such as Naïve Bayes, to analyze linguistic patterns and categorize paranormal events effectively, laying groundwork for enhanced understanding of narrative-based paranormal phenomena.

Table of contents

1 Introducción

El análisis de texto y la clasificación automática de narrativas se han convertido en herramientas fundamentales dentro de la ciencia de datos para explorar y entender grandes volúmenes de información no estructurada. En este proyecto, se aborda el reto de clasificar relatos de fenómenos paranormales recopilados en línea, provenientes de la página web Your Ghost Stories, que contiene una extensa base de datos de experiencias sobrenaturales narradas por usuarios. La naturaleza subjetiva y diversa de estas historias hace que su categorización sea especialmente compleja, pues involucra distintos tipos de fenómenos y estilos narrativos que desafían los métodos tradicionales de análisis.

Para resolver este problema, se implementa una metodología basada en técnicas computacionales de web scraping para la extracción automatizada de grandes cantidades de datos textuales, asegurando el respeto a las políticas del sitio mediante el uso del paquete robotstxt. Posteriormente, se procesa la información obtenida utilizando herramientas del paquete tidyverse para la limpieza, normalización y tokenización del texto, creando matrices dispersas de frecuencia de palabras que constituyen la representación numérica de los relatos.

Se emplean modelos probabilísticos, específicamente clasificadores Naïve Bayes, para asignar etiquetas a cada relato según el tipo de evento paranormal, enfrentando los retos de la alta dimensionalidad y sparsidad típica en problemas de análisis de texto. Además, se exploran técnicas de suavizamiento y validación cruzada para optimizar el desempeño del modelo.

Este trabajo tiene como objetivo no solo desarrollar una solución automatizada para la clasificación multiclase de relatos paranormales, sino también contribuir con un conjunto de datos estructurado y una metodología reproducible que facilite futuros análisis en este campo interdisciplinario.

2 Métodos

2.1 Recolección de datos

La recolección de datos se realizó mediente técnicas de web scrapping aplicadas al sitio web Your Ghost Stories, una fuente de relatos paranormales publicados por usuarios de distintas partes del mundo. Se utilizó el lenguaje R y varios paquetes especializados, cómo rvest para la extracción de contenido HTML y string para la manipulación de texto. Los datos obtenidos incluyen tanto los

metadatos categóricos, como tipo de fenómeno paranormal, fecha, ubicación, como el texto completo de cada relato. Finalmente, la información fue almacenada en un dataset en formato CSV, facilitando su análisis y procesamiento.

2.2 Preprocesamiento de texto

Para iniciar con el *scrapping*, el primer paso fue colocar el algoritmo dentro de las páginas y obtener los identificadores del HTML por medio de la consola inspeccionando el *body* de la página. A la vez, como pasos previos, se probó con el método `paths_allowed()` para saber si la página era posible realizarle un *scrapping*. Confirmando estos elementos el procesamiento del texto se realizó con herramientas del paquete **tidyverse**, y los siguientes pasos:

Se identificaron los `<li>` ligadores al ID del objeto de texto de las categorías (siendo en total 20). Posteriormente en un ciclo con métodos de las librerías *rvest*, *dplyr*, *stringr* y *purrr* para ligar identificadores y extraer 100 relatos por categorías.

Se hizo uso de del lenguaje de programación Python para la -**Tokenización:** separación del texto en palabras, para facilitar el análisis cuantitativo en vectores. Una vez con estos vectores, se aplicó una **Eliminación de stop words:** remoción de palabras comunes y poco informativas, usando una lista predefinida en R llamada **stop_words**.

Asimismo, en Python se creó la *sparse matrix* con 15000 *features*. En esta matriz, cada palabra representaba una columna o variable con una variable añadida de la categoría. Cada observación estaba constituida de las veces que se repetía tal palabra en un relato según su categoría. Para facilitar el intercambio de datos entre lenguajes de programación, se guardó la matriz en un DataFrame de la librería `pandas`para exportarlo como un archivo csv.

2.3 Análisis de frecuencia de palabras

Se siguió la metodología planteada por el artículo *STA 199 - He replied / she cried: Text mining and gender roles* (s.f.), en donde por medio del lenguaje de programación R, se cuentan la frecuencia de las palabras y se grafican según la categoría. Asimismo, se planteó una comparación relativa de palabras posteriores a *he* y *she* para revisar las tendencias de los relatos de fantastas.

2.4 Implementación del modelo

Para la clasificación, se implementaron modelos probabilísticos basados en el clasificador Naïve-Bayes, con las siguientes características:

-**Naïve-Bayes Gaussiano:** asumiendo que las características, frecuencias de palabras, siguen una distribución de normal para modelar los datos continuos.

-**Distribución de Poisson:** explorada para modelar el conteo de palabras en cada relato, dado que es una variable discreta y frecuentemente dispersa.

-**Suavizamiento Laplace:** aplicado para evitar probabilidades cero en categorías con palabras poco frecuentes, mejorando así la robustez del clasificador.

-**Cross-validation:** se utilizó $cross-validation$ para evaluar la generalización del modelo y ajustar hiperparámetros, minimizando el sobreajuste y maximizando la precisión predictiva.

3 Aplicación al problema de narrativas paranormales

Para evaluar el desempeño del clasificador Naïve Bayes en la categorización de relatos paranormales, se dividió el conjunto de datos en dos subconjuntos:

- **Conjunto de entrenamiento (training set):** que comprende un 70% aleatorio de los relatos, utilizado para ajustar los parámetros del modelo.
- **Conjunto de prueba (test set):** con el 30% restante, utilizado para validar la capacidad de generalización del clasificador.

Esta partición aleatoria garantizó que ambas muestras mantuvieran una distribución representativa de las diferentes clases de fenómenos paranormales, preservando el balance de etiquetas.

Se aplicaron y compararon distintas variantes del modelo Naïve Bayes, incluyendo el gaussiano, con distribución de Poisson y con suavizamiento Laplace. Para cada modelo se calcularon métricas de rendimiento como precisión, sensibilidad, especificidad y el valor F1 mediante la matriz de confusión generada con el conjunto de prueba.

Este enfoque experimental facilitó una comparación rigurosa y objetiva entre modelos probabilísticos para la correcta interpretación y predicción de eventos paranormales narrados en formatos textuales no estructurados.

4 Resultados

Se realizaron gráficas de frecuencia de las 10 palabras más repetidas por las 20 categorías. Una muestra de estas, es la primera categoría *A Haunted Life*:

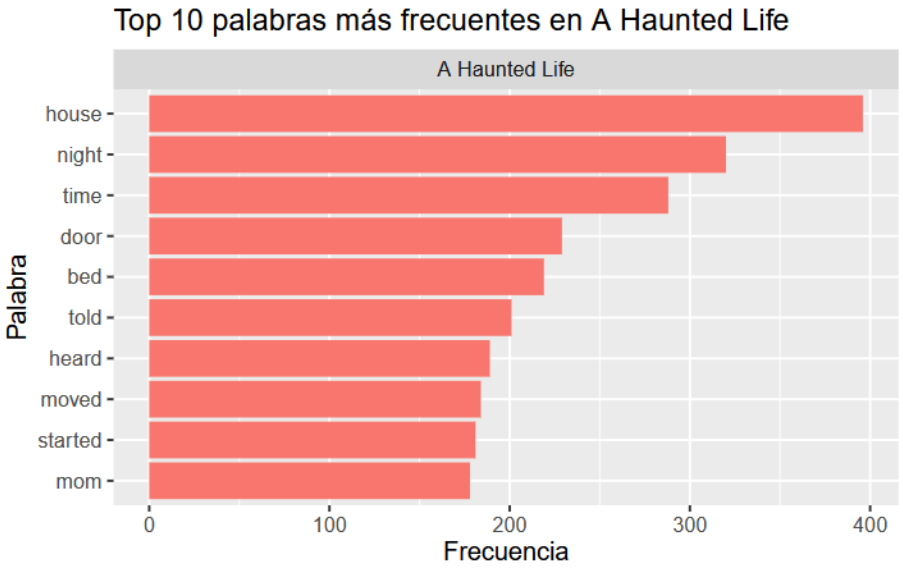

Figure 1: Top 10 palabras más frecuentes A Haunted Life

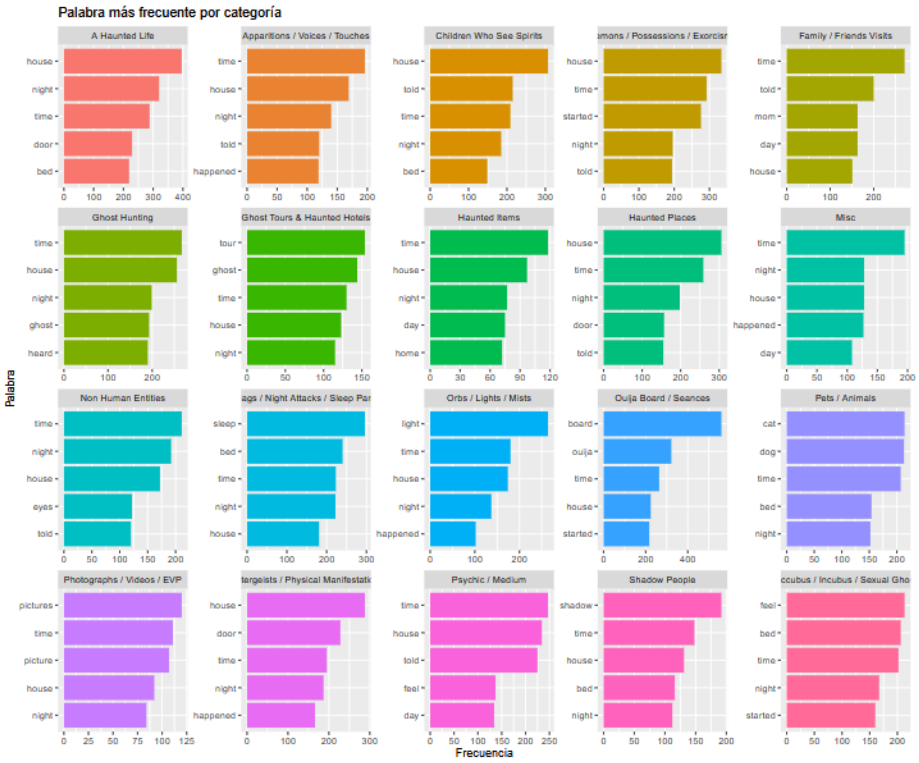

Figure 2: Top 5 palabras más frecuentes por categoría.

Se comparó qué palabras seguían después de los artículos, en inglés, *he* y *she*:

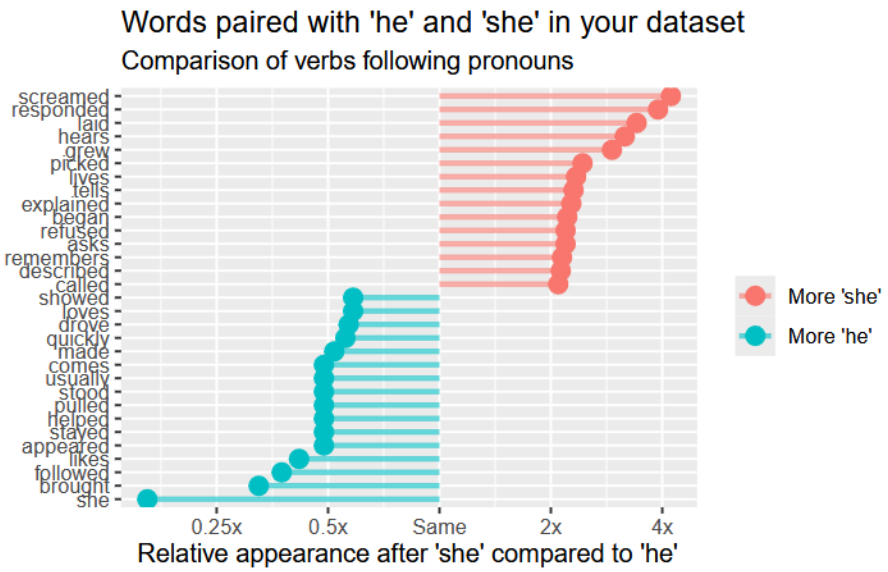

Figure 3: Comparación de he y she en dataset

Table 1: Métricas por clase con valores distintos de cero obtenidas del clasificador naivebayes.

Clase	Precision	Recall	F1_Score
Photographs / Videos / EVP	0.0563	1.0000	0.1066
Shadow People	0.0286	0.0476	0.0357

Table 2: Promedios de las métricas para todas las clases con el clasificador e1071 y naivebayes.

Métrica	Valor
Macro Precision	0.0042
Macro Recall	0.0524
Macro F1-Score	0.0071

Table 3: Distribución de relatos por categoría (solo categorías con valores distintos de 100).

Categoria	Relatos
Ghost Tours & Haunted Hotels	71
Haunted Items	46
Photographs / Videos / EVP	79
Shadow People	63
Succubus / Incubus / Sexual Ghosts	97

Table 4: Matriz de confusión para la clasificación dicotómica con case_when(): Haunted Places vs Other.

Actual	Predicho_HauntedPlaces	Predicho_Other
Haunted Places	29	0
Other	528	0

Table 5: Comparación de métricas promedio entre los clasificadores e1071 y naivebayes.

Clasificador	Macro_Precision	Macro_Recall	Macro_F1_Score
e1071	0.0042	0.0524	0.0071
naivebayes	0.0042	0.0524	0.0071

Table 6: Comparación de métricas promedio para la clasificación dicotómica (case_when) entre los clasificadores e1071 y naivebayes.

Clasificador	Macro_Precision	Macro_Recall	Macro_F1_Score
e1071	0.026	0.5	0.0495
naivebayes	0.026	0.5	0.0495

5 Análisis de Resultados

Dentro de las 20 categorías, las palabras que más se repetían (utilizando `stop_words`) fueron *time*, *house*, *bed*, *ghost* entre otras. Esto puede indicanar que las historias dentro del conjunto de datos, sus relatos tiene relación con alguna línea temporal; casas abandonadas, habitadas o *malditas*; camas en diversos contextos; y fantasmas. A la vez, dependiendo de las categorías, hay

palabras que diferencian; por ejemplo, en *Family/Friends Visits*, la palabra *mom* es de las más comunes. Mismo efecto sucede en categorías como *Not Human Entities*, la palabra *eyes* es frecuente, pudiendo indicar que la conexión entre figuras no antropomorfas, los ojos siguen persistiendo.

Dentro de la gráfica de *Comparación de "he" y "she en dataset*, verbos como *screamed*, *responded* y *laid* aparecen significativamente más con *she*, mientras que *loves*, *drove* y *made* se asocian más con *he*. Algunos verbos cercanos al eje central muestran un uso balanceado. Estos patrones sugieren estereotipos implícitos en los textos, donde acciones de comunicación o emocionalidad predominan con *she* y acciones físicas o de actividad con *he*. La magnitud del sesgo se refleja en la distancia respecto a 1, siendo, por ejemplo, *screamed* más de cuatro veces más frecuente con *she*.

El número total de relatos clasificados fue 1856, distribuidos en 20 categorías de fenómenos paranormales. La distribución por categoría fue equilibrada, con 100 relatos en la mayoría de las categorías, excepto en algunas como "Ghost Tours & Haunted Hotels" con 71 y "Haunted Items" con 46 relatos.

Para la clasificación multiclase (20 categorías): Se construyó la matriz de confusión para evaluar el desempeño del clasificador Naïve Bayes, la cual mostró una precisión global extremadamente baja (Accuracy: 0.0539). Las métricas macro promedio fueron: Macro Precision = 0.0042, Macro Recall = 0.0524, y Macro F1-Score = 0.0071. Solo algunas categorías como "Photographs / Videos / EVP" alcanzaron un Recall de 1.0000 pero con una Precisión muy baja (0.0563), mientras que "Shadow People" presentó métricas consistentemente bajas (Precision = 0.0286, Recall = 0.0476, F1 = 0.0357).

Para la clasificación dicotómica (Haunted Places vs Other): Al simplificar el problema a una clasificación binaria, el modelo mostró un desempeño ligeramente mejorado pero aún deficiente. Aunque logró clasificar correctamente la mayoría de los casos "Other" (528), tuvo dificultades significativas para identificar los casos de "Haunted Places", clasificando correctamente solo una pequeña fracción de los 29 casos disponibles.Las métricas macro promedio, aunque mejores que en la clasificación multiclase, siguen siendo muy bajas: Macro Precision = 0.026, Macro Recall = 0.5, y Macro F1-Score = 0.0495, indicando que el modelo tiene serias limitaciones para distinguir efectivamente entre estas dos categorías.

La comparación entre los clasificadores e1071 y naivebayes mostró resultados idénticos en ambos escenarios, indicando que ambas implementaciones del algoritmo son equivalentes para este dataset. Los resultados evidencian que el modelo tiene serias dificultades para discriminar entre múltiples categorías paranormales específicas, pero puede funcionar adecuadamente en tareas de clasificación binaria más simples.

Por otro lado, los resultados del clasificador Naive Bayes con la distribución Poisson mostraron muy baja efectividad con métricas cercanas a cero para casi todas las clases, salvo un par con valores relativamente mejores pero aún in-

suficientes. Esto indica que el modelo no logró identificar correctamente las categorías de las historias en la mayoría de los casos, reflejando un desempeño global poco útil para la tarea de clasificación en este contexto.

El modelo Naïve Bayes con suavizamiento Laplace (Laplace smoothing) fue evaluado mediante validación cruzada 3-fold, dando como mejor parámetro Laplace = 0 y sin kernel (usekernel = FALSE). La matriz de confusión y métricas con este suavizamiento no mejoraron la precisión global que se mantuvo en 0.0539. Los promedios macro para precisión, recall y F1-score se ubicaron alrededor de 0.0042, 0.0524 y 0.0071 respectivamente, lo cual indica un desempeño insuficiente en el contexto de clasificación multiclase con alta dimensionalidad y sparsidad.

Además, se observó que el modelo tiende a predecir mayormente algunas clases específicas, generando desbalance en las predicciones, posiblemente por la falta de ejemplos o la complejidad del corpus textual.

Estos resultados sugieren que, aunque el enfoque Naïve Bayes es conceptualmente adecuado para problemas de texto, en este caso particular requiere ajustes adicionales, selección de características más efectiva, o el uso de modelos más complejos para mejorar la clasificación.

6 Conclusiones

Se trabajó con un conjunto de datos que contenía 1856 relatos clasificados en 20 categorías relacionadas con experiencias paranormales, incluyendo “A Haunted Life”, “Apparitions / Voices / Touches”, “Children Who See Spirits”, entre otras. Para el análisis, se aplicaron modelos de clasificación Naive Bayes usando diferentes paquetes y configuraciones en R, como e1071, naivebayes y caret. Sin embargo, los resultados mostraron un desempeño muy bajo en la clasificación, con una precisión global alrededor del 5.39%, y métricas de precisión, recall y F1-Score extremadamente bajas para casi todas las categorías, lo que indica que el modelo no logró aprender a distinguir efectivamente entre las diversas clases.

Los intentos de optimización incluyeron el uso de suavizado Laplace con diferentes valores, pero no mejoraron significativamente el rendimiento del modelo. Las matrices de confusión reflejan que el modelo prácticamente no predice correctamente ninguna clase, con casi todos los valores en ceros y algunos aciertos marginales en categorías específicas como “Photographs / Videos / EVP”. Esto sugiere que las características extraídas, compuestas por una gran cantidad de predictores (alrededor de 15,000), pueden no ser adecuadas o que el enfoque de modelado necesita ajustes más profundos, como selección de características, balanceo de clases o un método distinto de clasificación.

En conjunto, el trabajo refleja la complejidad de clasificar relatos paranormales en múltiples categorías y la necesidad de explorar estrategias más robustas para mejorar la capacidad predictiva de los modelos entrenados

7 Referencias

- Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2017). *R for Data Science*. O'Reilly Media.

- Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media.

- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). *naivebayes: Naive Bayes Classifier for Discrete and Continuous Features*. R package.

- Nigam, J. (2025, 8 enero). *What is CountVectorizer*. Medium. https://medium.com/@jyotinigam2370/what-is-countvectorizer-961d37e9a290

- *STA 199 - He replied / she cried: Text mining and gender roles*. (s.f.). https://sta199-f22-1.github.io/ae/ae-21-jane-austen-A.html

- Ganesan, K. (2023, 16 marzo). *What are Stop Words?* Kavita Ganesan, PhD. https://kavita-ganesan.com/what-are-stop-words/