Situación Problema: Ghostbusters!

Marissa E. Luna[1], Ximena A. Cantón[1], Nubia S. Garcidueñas[1] and Mariana L. Maldonado[1]

[1] *Instituto Tecnológico y de Estudios Superiores de Monterrey, Escuela de Ingenieria y Ciencias, Guadalajara, Jalisco*

Abstract— Este estudio tiene como objetivo implementar un clasificador naïve Bayes para categorizar relatos de fenómenos paranormales recopilados en línea mediante técnicas de web scraping. La investigación parte de la extracción de historias de la plataforma Your Ghost Stories, donde se recolectaron narrativas etiquetadas según el tipo de evento. Posteriormente, se aplicaron técnicas de procesamiento de lenguaje natural (NLP), incluyendo limpieza de texto, eliminación de palabras vacías y construcción de una matriz dispersa. Con estos datos, se entrenó un clasificador naïve Bayes en distintos escenarios: clasificación multiclase y clasificación dicotómica. Se evaluó el desempeño del modelo a través de métricas como accuracy, precision, recall, F1-score y la matriz de confusión. Los resultados mostraron que la clasificación binaria produjo un desempeño superior frente al esquema multiclase, y que el uso de técnicas como Laplace smoothing y la distribución de Poisson mejoraron la capacidad predictiva del modelo. En conclusión, la combinación de web scraping y métodos probabilísticos de NLP son una técnica eficaz para abordar problemas de categorización en textos narrativos.

Keywords—Naïve Bayes, procesamiento de lenguaje natural (NLP), web scraping, clasificación de texto, fenómenos paranormales, Sparse matrix

I. Introducción

En las últimas décadas, el análisis automático de texto se ha consolidado como una herramienta fundamental dentro del procesamiento de lenguaje natural (NLP), con aplicaciones que van desde la minería de opiniones hasta la clasificación automática de documentos. Uno de los enfoques más simples pero efectivos en este ámbito es el clasificador naïve Bayes, ampliamente utilizado en problemas de categorización de texto gracias a su simplicidad computacional y a su capacidad de ofrecer buenos resultados incluso con conjuntos de datos de tamaño moderado (Manning, Raghavan and Schütze, 2008). [1]

El presente trabajo busca aplicar este modelo al dominio de los relatos paranormales, un campo poco explorado desde la perspectiva computacional. Para ello, se recopilaron narrativas de experiencias sobrenaturales mediante técnicas de web scraping, con el fin de construir un conjunto de datos que pudiera ser analizado bajo un enfoque probabilístico. El interés de este estudio radica en demostrar cómo técnicas clásicas de clasificación de texto pueden adaptarse a dominios no convencionales, contribuyendo a la investigación interdisciplinaria entre ciencias de datos y estudios culturales.

En particular, se plantea la hipótesis de que, mediante un preprocesamiento adecuado de texto y la implementación de variantes del clasificador naïve Bayes (incluyendo suavizamiento y distribución de Poisson), es posible categorizar relatos paranormales con un nivel de precisión significativo. Este ejercicio no solo aporta un ejemplo práctico de las capacidades del NLP aplicado a narrativas, sino que también abre la posibilidad de explorar fenómenos sociales a través del análisis automatizado de relatos.

II. Metodología

La metodología del presente reto, se basa en tres etapas principales: (1) recopilación y preparación de datos, (2) análisis de texto y representación vectorial, y (3) modelado probabilístico mediante un clasificador naïve Bayes. A continuación, se detallan los procedimientos, técnicas y herramientas empleadas de manera conceptual:

a. Recopilación y preparación de datos

La primera fase consistió en obtener relatos de fenómenos paranormales mediante técnicas de web scraping. Recolectando de manera sistemática historias publicadas por usuarios en la plataforma "Your Ghost Stories". Antes de realizar la extracción, se verificó el cumplimiento de las normas de acceso del sitio, garantizando que el scraping fuera legal.

El conjunto de datos construido incluyó cuatro atributos fundamentales: título del relato, lugar, tipo de evento paranormal y descripción completa. Una vez recolectada la información, se aplicaron procesos de limpieza y normalización de texto, como eliminación de caracteres especiales, conversión a minúsculas y eliminación de palabras vacías.

b. Análisis de texto y representación vectorial

Para transformar los relatos en una estructura interpretable por algoritmos de clasificación, se emplearon técnicas de minería de texto. Específicamente, se construyó una matriz dispersa de frecuencias de términos, donde cada fila representa un relato y cada columna una palabra distinta.

Este tipo de representación permite cuantificar la presencia de vocabulario específico en cada relato, reduciendo la subjetividad de la interpretación textual. Además, se aplicaron

criterios de filtrado para eliminar palabras con alta frecuencia general (que aportan poca información) y términos demasiado raros (que dificultan la generalización del modelo).

c. Clasificación mediante Naïve Bayes

La etapa de modelado se basó en la implementación de un clasificador naïve Bayes, este, asume independencia condicional entre las características del texto (palabras) dadas las clases (tipos de fenómenos paranormales). Aunque parece simple, es efectiva en tareas de categorización de texto debido a la naturaleza dispersa y de alta dimensionalidad de los datos.

Se exploraron distintas configuraciones del modelo:

- **Clasificación multiclase**: se mantuvieron las categorías originales de fenómenos paranormales
- **Clasificación dicotómica**: las categorías fueron recodificadas en dos grupos (por ejemplo, Haunted Places vs. Other).

Además, se incorporaron variantes del algoritmo para abordar limitaciones comunes:

- **Suavizamiento de Laplace**:utilizado para manejar el problema de frecuencias nulas en palabras poco comunes.
- **Distribución de Poisson**: alternativa a la distribución normal, dado que los datos representan conteos discretos de palabras.
- **Validación cruzada**: usada para seleccionar parámetros óptimos y reducir el riesgo de sobreajuste.

d. Evaluación del modelo

Finalmente, el desempeño del clasificador se evaluó mediante un conjunto de métricas estándar en problemas de clasificación: accuracy, precision, recall, F1-score y la matriz de confusión. Estas medidas permitieron analizar no solo la capacidad general de predicción, sino también el balance del modelo en la detección de las diferentes categorías de fenómenos.

III. Aplicación

a. Datos y preprocesamiento

Primeramente, y después de verificar que el web scraping estaba permitido por la página web con la que trabajaríamos, por medio de la función *pathsallowed*() del paquete robotstxt, comenzamos con el proceso de web scraping:

El proceso de web scraping consistió en la automatización de la extracción de relatos desde la plataforma *Your Ghost Stories*. Para ello se desarrolló un script en Python que empleó librerías como *requests* y *BeautifulSoup* para acceder al código HTML de cada página, identificar las etiquetas que contenían el título, el contenido narrativo y las categorías del relato, y posteriormente almacenarlos en un archivo estructurado.

Este procedimiento permitió recolectar de manera eficiente un gran volumen de datos que hubiera sido impráctico recopilar manualmente. Además, se establecieron rutinas de control para manejar posibles errores de conexión, evitar la descarga de contenido duplicado y respetar las restricciones de la página mediante la consulta previa del archivo *robots.txt*.

Una vez extraídos, los relatos fueron organizados en un *DataFrame* con formato tabular, lo que facilitó su análisis posterior. El almacenamiento en este formato permitió llevar un control ordenado de cada historia junto con sus metadatos (categoría, título y texto narrativo), lo cual fue esencial para las etapas siguientes de limpieza, representación vectorial y modelado probabilístico.

Ahora bien, antes de realizar el análisis de palabras, se realizó la limpieza del archivo csv generado por el web scraping, en donde se compilaron 3,120 relatos y, tras depurar párrafos ajenos al relato y limpiar duplicados y ruido, se conservaron 2,509 historias con descripciones no vacías.

b. Análisis de Palabras

El texto se tokenizó a nivel de palabra, se eliminaron stop words (nexos, preposiciones, pronombres, etc.) o que contuvieran números, y tokens con dígitos; posteriormente se filtraron términos con frecuencia global en las historias para reducir ruido léxico y la dimensionalidad.

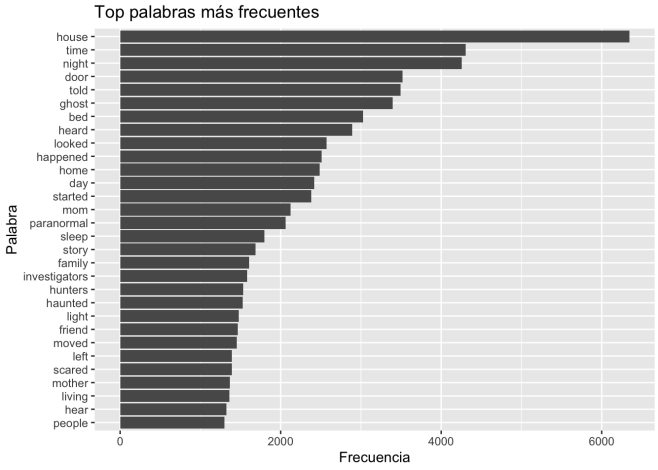

Fig. 1: Top palabras más frecuentes de manera global

Y posteriormente se obtuvo la frencuencia de palabras por cada uno de los relatos, se filtraron aquellas palabras que se repitieran al menos 5 veces de manera global, así como la cantidad de veces que se repetía esta palabra en el relato:

title	word	n
115 Years Old Slave House Haunted	house	4
115 Years Old Slave House Haunted	sarah's	4
115 Years Old Slave House Haunted	night	3
115 Years Old Slave House Haunted	shelly	3
115 Years Old Slave House Haunted	slave	3
11:11 PM Poltergeist Problems	house	3
11:11 PM Poltergeist Problems	evil	2
11:11 PM Poltergeist Problems	fear	2
11:11 PM Poltergeist Problems	flicked	2
11:11 PM Poltergeist Problems	ghost	2

Fig. 2: Tabla de frecuencia por relato (2 primeros relatos)

Y para analizarlo de manera visual, en unas gráficas con algunos ejemplos de relatos:

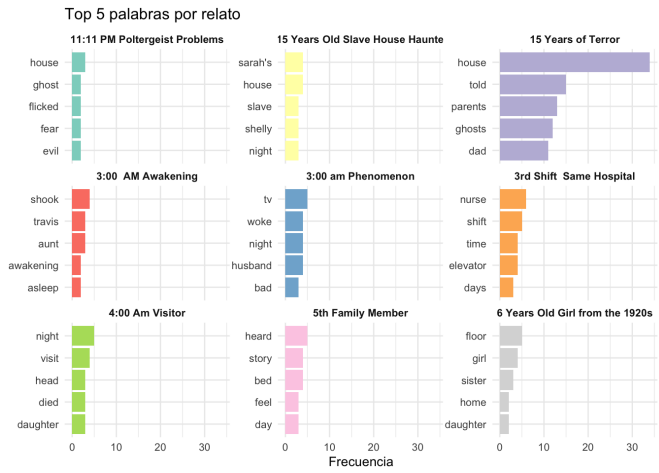

Fig. 3: Gráficas de frecuencia por relato (9 primeros relatos)

Lo anterior se realizó con el objetivo de eliminar ruido (tipo outliers) de nuestro clasificador.

c. Sparse Matrix

Se contruyó un matriz qque contiene la frecuencia de cada palabra en los relatos scrapeados, donde cada renglón representa un relato diferente y donde las columnas son las diferentes palabras, por medio del proceso conocido como "count vectorizer":

doc_id	ago	argue	asleep	bags	bathroom	bedroom	birthday	boarded	breathing	candy	closet
115 Years Old Slave House Haunted	1	1	2	1	2	2	1	1	1	1	2
11:11 PM Poltergeist Problems	1	0	0	0	0	0	0	0	1	0	0
15 Years of Terror	0	0	1	0	1	2	1	0	1	0	0
3:00 AM Awakening	0	0	2	0	0	0	0	0	0	0	0
3:00 am Phenomenon	1	0	0	0	0	0	0	0	0	0	0
3rd Shift Same Hospital	0	0	0	0	0	0	0	0	0	0	0
4:00 Am Visitor	0	0	2	0	0	1	0	0	0	0	0
5th Family Member	2	0	0	0	0	0	0	0	2	0	0
6 Years Old Girl from the 1920s	0	0	1	0	0	1	0	0	0	0	0

Fig. 4: Sparse Matrix de los relatos (primeros renglones)

En este caso, se utilizó un formato de sparse matrix que no necesita tanta memoria. Para visualizar que la conversión haya sido la desada, se convirtió a formato de data frame y se verificó que todo estuviera correcto.

d. Modelos "Y Multiclase"

Primeramente, se dividió el conjunto en training/test con estratificación (80/20) para preservar la distribución de clases.

Se evaluó el desempeño del clasificador Naïve Bayes mediante métricas estándar de clasificación:

accuracy, *precision*, *recall* y *F1-score*. Además, se incluyó la matriz de confusión para analizar el comportamiento del modelo frente a cada clase.

- **Accuracy**: proporción de predicciones correctas respecto al total de observaciones.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Precision**: proporción de observaciones clasificadas como positivas que realmente son positivas.

$$\text{Precision} = \frac{TP}{TP+FP}$$

- **Recall** (sensibilidad): proporción de observaciones positivas que fueron correctamente identificadas.

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **F1-score**: media armónica entre precisión y recall, útil en presencia de clases desbalanceadas.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Los resultados obtenidos muestran que el modelo e1071 alcanza una **Accuracy de 0.01825558**, y en cuanto al modelo naive bayes alcanzó una **Accuracy de 0.01825558**, notando que realmente no existió ninguna diferencia entre ellos.

Sin embargo, en la evaluación del clasificador multiclase se observaron valores NaN en algunas métricas de precisión y recall. Esto ocurre porque en ciertas clases el modelo nunca generó predicciones (lo que hace que el denominador en la fórmula de la precisión sea cero) o bien porque en el conjunto de prueba no había observaciones reales de esa clase (lo que afecta el cálculo del recall). En ambos casos, la métrica queda indefinida y se representa como NaN.

Este resultado indica que existen clases con muy bajo soporte o que son difíciles de identificar para el modelo, lo cual es común en escenarios con datos desbalanceados. Por esta razón, además de los valores por clase, es recomendable reportar métricas globales (como accuracy) y considerar la dicotomización de categorías poco frecuentes para obtener una evaluación más estable.

e. Modelos de 2 clases

Ahora bien, con la ayuda de la función case when() se recodificó el tipo de evento paranormal, en este caso, si era Haunted Places u Other, utilizando la misma partición pero dicotomizamos sobre ytrain y ytest.

Se volvió a entrenar un clasificador naïve Bayes con ambas funciones, para ver si había alguna mejora y los resultados obtenidos fueron los siguientes:

- **Para e1071**: se obtuvo un *accuracy* de **0.31643**.
- **Para Naïve Bayes (paquete naivebayes)**: se obtuvo un *accuracy* de **0.31643**.

Con lo anterior confirmamos que si hubo una mejora con respecto a los modelos multiclase, sin embargo entre ellos muestran un mismo desempeño,

Ahora bien, un problema que enfrentan estas primeras implementaciones del clasificador naïve Bayes es el supuesto

	Reference	
Prediction	Other	Haunted Places
Other	1	0
Haunted Places	337	155

Fig. 5: Matriz de confusión del clasificador Naïve Bayes con `e1071`.

	Reference	
Prediction	Other	Haunted Places
Other	1	0
Haunted Places	337	155

Fig. 6: Matriz de confusión del clasificador Naïve Bayes con `naivebayes`.

de la distribución normal para los nodos hijos, por lo que la distribución de Poisson resulta una mejor alternativa para este tipo de datos, por lo que dentro de la misma función utilizada de naive bayes, modificamos el argumento usepoisson = TRUE. Sin embargo, no hubo mejora en las metricas resultantes, ni en el "accuracy", ni en la matriz de confusión.

Además de lo anterior, otro problema que enfrentamos en este tipo de datos es que existen muchos ceros en la matriz de datos. Una posible solución es utilizar una técnica de suavizamiento conocida como Laplace smoothing, que consiste en modificar el argumento del parámetro de laplace dento de la misma función de naive bayes. Lo que hace este parámetro, es que básicamente asegura que el clasificador Naïve Bayes nunca descarte relatos por la ausencia de una palabra en una clase, lo que es crucial porque nuestros datos son altamente esparsos y con vocabulario extraño.

En este caso, al implementar este argumento **los resultados mejoraron notoriamente**, pero aún era importante obtener el mejor valor para este parámetro, para ellos utilizamos Cross Validation, obteniendo lo siguiente:

laplace	usekernel	adjust
0.5	FALSE	1

Fig. 7: Resultado de Cross Validation

Finalmente, utilizando esto como argumento del mejor valor para laplace, al implementarla en nuestro clasificador obtuvimos nuestros mejores resultados en las métricas:

- **Accuracy**: 0.8600406
- **Precision**: 0.8705234
- **Recall**: 0.9349112
- **F1-score**: 0.9015692
- **Matriz de confusión**:

	Reference	
Prediction	Other	Haunted Places
Other	316	47
Haunted Places	22	108

Fig. 8: Matriz de confusión del mejor modelo Naïve Bayes.

IV. Conclusiones

Este estudio mostró que el clasificador naïve Bayes es una herramienta eficaz para categorizar relatos paranormales obtenidos mediante web scraping. Aunque la clasificación multiclase presentó dificultades por la heterogeneidad de categorías, la recodificación dicotómica mejoró notablemente el desempeño del modelo.

El uso de técnicas como el suavizamiento de Laplace y la distribución de Poisson permitió adaptar el modelo a la naturaleza discreta y dispersa de los datos, reforzando su capacidad predictiva.

En conclusión, los métodos probabilísticos de NLP ofrecen una solución práctica y eficiente para analizar narrativas textuales en contextos no convencionales, y sientan las bases para futuros estudios que integren modelos más complejos o enfoques semánticos.

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, recuperado el 13 de septiembre de 2025.

[2] D. J. E. G. Guillén, "Clasificadores de redes bayesianas," file://C:/Users/user/Downloads/Clasificadores%20de%20redes%20bayesianas%20(3).html, 2025, archivo HTML. Recuperado el 13 de spetiembre de 2025.

[3] *OpenAI, ChatGPT (GPT-5). [Online]. Available: https://chat.openai.com/, 2025.*