Ghostbusters

Diana L. Hernández-Almeida[1], José A. López-Torres[1], Axel D. Luna-Hernández[1] and Jorge E. Nevarez-Soto[1]

[1] *Tecnologico de Monterrey*

Abstract—This study addresses the task of classifying narratives of paranormal phenomena shared in online communities. The objective was to implement and evaluate naïve Bayes classifiers on stories collected through web scraping from the platform *Your Ghost Stories*. The dataset comprised 1,828 narratives labeled into two categories: *Haunted* (823) and *Other* (1,005). Preprocessing included tokenization, stop word removal, and the construction of sparse matrices to represent word frequencies, yielding 6,727 unique terms. Two baseline implementations of naïve Bayes, using the `e1071` and `naivebayes` packages, achieved identical results with an accuracy of 61.6%, precision of 57.1%, recall of 50.2%, and F1-score of 53.4% for the *Haunted* class. Extending the model with a Poisson distribution improved performance to 69.6% accuracy and an F1-score of 60.9%, although it generated warnings due to zero counts. Incorporating Laplace smoothing mitigated these warnings, producing a stable model with 69.2% accuracy and an F1-score of 60.0%. Cross-validation identified an optimal Laplace parameter ($\lambda = 0.2$), confirming the trade-off between predictive power and robustness. Overall, the results demonstrate that naïve Bayes can provide meaningful classification of highly subjective paranormal narratives, and that probabilistic extensions such as Poisson modeling and Laplace smoothing substantially enhance performance and stability.

Keywords—naïve Bayes, text mining, web scraping, paranormal phenomena, classification, NLP

I. Introducción

En internet existen numerosas comunidades en las que las personas comparten experiencias relacionadas con fantasmas, UFOs, brujería, criaturas extrañas y otros sucesos inexplicables. Estas narraciones suelen estar redactadas con descripciones detalladas y un lenguaje subjetivo, lo que dificulta su clasificación sistemática. La diversidad de fenómenos reportados y la ausencia de categorías estandarizadas complican la organización y el análisis de este tipo de contenidos.

En el campo del procesamiento de lenguaje natural (NLP), los modelos probabilísticos han sido ampliamente utilizados en tareas de clasificación de texto[1, 2]. Entre ellos, el clasificador *naïve Bayes* ha destacado por su eficacia al trabajar con datos de alta dimensionalidad y representaciones dispersas. A pesar de la fuerte suposición de independencia condicional entre las variables, distintos estudios han demostrado que sigue siendo una línea base competitiva en el ámbito de la clasificación.

Este estudio aplica un clasificador *naïve Bayes* al análisis de narraciones paranormales recopiladas mediante *web scraping* de la plataforma *Your Ghost Stories*. El conjunto de datos obtenido incluye el título, lugar, tipo de evento y descripción de cada historia. Inicialmente se evaluaron dos implementaciones estándar de *naïve Bayes*, y posteriormente se incorporaron extensiones probabilísticas para enfrentar limitaciones propias de este enfoque. Entre ellas destacan el uso de la distribución Poisson para modelar datos de conteo, la aplicación de suavizamiento de Laplace para manejar la abundancia de ceros y la validación cruzada para la selección óptima de parámetros. De esta forma, el objetivo principal es valorar la capacidad de los modelos *naïve Bayes* y sus variantes para categorizar estas narraciones en clases significativas, resaltando tanto su potencial como las limitaciones al tratar con datos textuales no convencionales y altamente subjetivos.

II. Metodología

La metodología consistió en aplicar técnicas de clasificación supervisada a narrativas textuales, organizadas en cuatro fases principales: recolección de datos, preprocesamiento, representación de los textos en formato matricial y entrenamiento de un clasificador probabilístico.

La recolección de datos se realizó mediante técnicas de *web scraping*, seleccionando un portal especializado en relatos de fenómenos paranormales. Este procedimiento permitió construir una base a partir de historias reales publicadas por usuarias y usuarios, preservando para cada registro atributos básicos como título, lugar, categoría asignada y descripción narrativa.

El preprocesamiento incluyó operaciones comunes en análisis de texto, como normalización de caracteres, eliminación de signos de puntuación, filtrado de palabras vacías y tokenización[3, 4]. Estas transformaciones convirtieron las narraciones en unidades comparables y facilitaron su posterior análisis cuantitativo.

Para la representación matemática de los textos se construyó una matriz documento-término, donde cada relato se corresponde con una fila y cada palabra con una columna. Dada la naturaleza de los datos, esta matriz resultó dispersa, lo cual es típico en tareas de minería de texto[3, 4].

Posteriormente, se entrenaron clasificadores *naïve Bayes*

como modelo base, utilizando tanto la implementación estándar de la librería `e1071` como la del paquete `naivebayes`[5] y `naivebayes` [6].. El desempeño de los modelos se evaluó con métricas estándar en clasificación: exactitud, precisión, recall, puntaje F1 y matriz de confusión.

Con el fin de superar las limitaciones de las versiones básicas, se implementaron tres extensiones metodológicas. En primer lugar, se incorporó la distribución Poisson para modelar de manera más adecuada los datos de conteo presentes en la matriz dispersa. En segundo lugar, se aplicó el suavizamiento de Laplace como estrategia para manejar la abundancia de ceros y reducir el impacto de frecuencias nulas en la estimación de probabilidades. Finalmente, se utilizó validación cruzada para explorar diferentes valores del parámetro de suavizamiento y seleccionar aquel que maximizara el desempeño del modelo de manera estable.

Este diseño metodológico permitió evaluar tanto la capacidad base del clasificador *naïve Bayes* como las mejoras derivadas de ajustes probabilísticos específicos para datos textuales.

III. Aplicación

El procedimiento descrito en la metodología se implementó utilizando los relatos obtenidos de la plataforma *Your Ghost Stories*. El conjunto de datos final incluyó 1,828 narraciones, de las cuales 823 se clasificaron como *Haunted* y 1,005 como *Other*. Esta distribución relativamente balanceada permitió entrenar y evaluar modelos de clasificación sin un sesgo severo hacia una sola categoría.

Como parte del análisis exploratorio, se identificaron las palabras más frecuentes en la base, entre las que destacaron *house*, *time*, *night* y *door*. Estos términos reflejan la recurrencia de contextos domésticos y nocturnos en los relatos paranormales. La Figura 1 muestra las 20 palabras más comunes después del preprocesamiento.

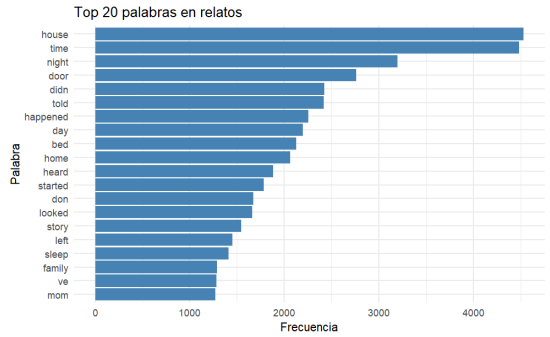

Fig. 1: Top 20 palabras más frecuentes en los relatos.

En cuanto a la modelación, el problema se redujo a una clasificación binaria entre las categorías *Haunted* y *Other*. El conjunto de datos fue dividido en 70% para entrenamiento y 30% para prueba. Se ajustaron dos versiones del clasificador *naïve Bayes*, una con la librería `e1071` y otra con la librería `naivebayes`.

Los resultados se presentan en la Tabla 1. Ambas implementaciones arrojaron métricas idénticas: una exactitud de 61.6%, precisión de 57.1%, recall de 50.2% y F1 de 53.4% para la clase *Haunted*. El desempeño moderado indica que el modelo logra capturar ciertos patrones, pero aún tiende a clasificar incorrectamente parte de los relatos de tipo *Haunted*.

Table 1: Resultados comparativos de las implementaciones estándar del clasificador *naïve Bayes*.

Modelo	Accuracy	Precisión	Recall	F1
e1071::naiveBayes	0.6157	0.5708	0.5021	0.5342
naivebayes	0.6157	0.5708	0.5021	0.5342

Un problema que enfrentan estas primeras implementaciones del clasificador *naïve Bayes* es el supuesto de normalidad en los predictores, lo cual resulta poco adecuado cuando los datos corresponden a conteos. Para abordar esta limitación se probó la distribución Poisson como alternativa.

Table 2: Comparación del modelo estándar frente a la versión Poisson y Poisson con Laplace.

Modelo	Accuracy	Precisión	Recall	F1
e1071::naiveBayes	0.6157	0.5708	0.5021	0.5342
naivebayes:	0.6157	0.5708	0.5021	0.5342
naivebayes::Poisson	0.6958	0.6989	0.5394	0.6089
naivebayes::Poisson_Laplace	0.6812	0.7012	0.4772	0.5679

Los resultados muestran que el modelo con distribución Poisson mejoró considerablemente el desempeño, alcanzando una exactitud de 69.6% y un F1 de 60.9%. Sin embargo, esta versión generó múltiples advertencias debido a la abundancia de ceros en la matriz dispersa. Para contrarrestar este problema se aplicó el suavizamiento de Laplace, lo que eliminó los errores pero redujo ligeramente el desempeño (F1 de 56.8%).

Finalmente, se aplicó validación cruzada para seleccionar el mejor valor del parámetro de suavizamiento.

Table 3: Resultados de validación cruzada con diferentes valores de λ en el suavizamiento de Laplace.

Laplace (λ)	Accuracy promedio
0.1	0.6513
0.2	0.6458
0.3	0.6513
0.4	0.6482
0.5	0.6458
0.6	0.6443
0.7	0.6435
0.8	0.6427
0.9	0.6411
1.0	0.6380

El valor óptimo identificado fue $\lambda = 0.2$. Aunque este ajuste no alcanzó la misma exactitud del modelo Poisson sin suavizamiento, sí proporcionó un desempeño estable y libre de advertencias. Con este ajuste, el modelo obtuvo una exactitud de 69.2% y un F1 de 60.0%, valores lo suficientemente cercanos para considerarse una alternativa robusta frente al problema de los ceros.

IV. Conclusiones

El presente estudio evaluó la aplicación de *naïve Bayes* para categorizar relatos de fenómenos paranormales recopilados mediante *web scraping* de la plataforma *Your Ghost Stories*. El análisis demostró que las versiones estándar del clasificador, implementadas con las librerías `e1071` y `naivebayes`, alcanzaron un desempeño moderado con una exactitud de 61.6% y un F1 de 53.4% para la clase *Haunted*.

La incorporación de la distribución Poisson permitió mejorar sustancialmente los resultados, logrando una exactitud cercana al 70% y un F1 de 60.9%. No obstante, este modelo generó múltiples advertencias debido a la presencia de ceros en la matriz dispersa. Para superar esta limitación se aplicó el suavizamiento de Laplace, que estabilizó el entrenamiento y redujo los errores, aunque con un ligero costo en el desempeño. Mediante validación cruzada se identificó un valor óptimo de $\lambda = 0.2$, el cual ofreció un equilibrio adecuado entre robustez y precisión, con métricas muy cercanas a las del modelo Poisson puro.

En conjunto, los resultados muestran que los clasificadores *naïve Bayes* constituyen una herramienta viable para el análisis de narraciones textuales altamente subjetivas, y que las extensiones probabilísticas pueden mejorar notablemente su rendimiento. Entre las principales limitaciones se encuentra la dependencia del modelo en supuestos simplificadores y la sensibilidad a la representación de los datos.

Como líneas futuras de investigación se sugiere explorar técnicas de representación más avanzadas, como embeddings contextuales[7], así como evaluar clasificadores alternativos capaces de capturar dependencias más complejas en los textos. Estas estrategias podrían complementar los hallazgos aquí presentados y contribuir a un análisis más preciso de fenómenos narrativos no convencionales.

References

[1] H. Zhang, "The optimality of naive bayes," in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 2004, pp. 562–567.

[2] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 Workshop on Learning for Text Categorization*, 1998, pp. 41–48.

[3] H. Wickham, M. Çetinkaya Rundel, and G. Grolemund, *R for Data Science, 2nd Edition*. O'Reilly Media, 2023. [Online]. Available: https://r4ds.hadley.nz

[4] H. Wickham, *stringr: Simple, Consistent Wrappers for Common String Operations*, 2019, r package version 1.5.0.

[5] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2021, r package version 1.7-9.

[6] M. Majka, *naivebayes: High Performance Implementation of the Naive Bayes Algorithm*, 2019, r package version 0.9.7.

[7] M. Scutari, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010.