Clasificador Naïve Bayes: Una categorización de relatos de fenómenos paranormales.

Romina Nájera Fuentes - A01424411
Humberto Mondragón García - A01711912
Juan Braulio Olivares Rodríguez - A01706880
Edgar Andrey Balvaneda - A01644770

Análisis de métodos de razonamiento e incertidumbre

14 de Septiembre, 2025

Abstract

El internet alberga millones de relatos personales sobre experiencias paranormales, los cuales ofrecen una rica fuente de datos sobre percepciones individuales y creencias colectivas. Este artículo busca aplicar técnicas de Procesamiento de Lenguaje Natural (NLP) para categorizar estas narrativas, utilizando un clasificador Naïve Bayes para diferenciar los relatos del género "Haunted Places" del resto. La metodología consistió en la extracción de datos mediante web scraping del sitio "Your Ghost Story", la transformación del texto a una representación numérica (count vectorizer) y el entrenamiento de cuatro modelos Naïve Bayes distintos. Los resultados muestran que los modelos iniciales fueron ineficaces, clasificando todas las observaciones en una sola categoría. Sin embargo, al ajustar el modelo para asumir una distribución de Poisson en las frecuencias de palabras y aplicar un suavizado de Laplace (con $\alpha = 1$), se obtuvo el mejor rendimiento, alcanzando una exactitud del 73.39% y un F1-score de 0.6116. Se concluye que un clasificador Naïve Bayes, debidamente ajustado a las características de los datos textuales, es una herramienta eficaz para categorizar este tipo de narrativas no convencionales.

Palabras clave: Clasificador Naïve Bayes, procesamiento de lenguaje natural (NLP), web scraping, relatos paranormales, matriz de confusión, suavizado de Laplace.

Introducción

El internet se ha convertido en un espacio para millones de personas que comparten experiencias personales relacionadas con lo paranormal, como avistamientos de ovnis, encuentros de fantasmas, entre otros. Estos testimonios, aparte del carácter anecdótico, representan fuentes con amplias descripciones textuales contenientes de la percepción individual y creencias colectivas de lo inexplicable. En estos casos, la variedad de fenómenos reportados y ambigüedad del lenguaje dificultan la clasificación sistemática de estos relatos.

En el ámbito del análisis de texto y procesamiento de lenguaje natural (NLP), uno de los enfoques utilizados para categorizar los documentos es el clasificador Naïve Bayes, un modelo probabilístico basado en el teorema de Bayes y en la suposición de independencia entre características.

El clasificador Naïve Bayes ha probado ser eficaz en el ámbito de la minería de texto, como han evidenciado diferentes estudios. Por ejemplo, gracias a su habilidad para identificar patrones en el uso de palabras, se ha empleado extensamente en la detección de correos electrónicos no deseados (filtrado de spam). Además, se ha utilizado en análisis de sentimientos, lo cual posibilita diferenciar la polaridad de los puntos de vista en las reseñas de productos o en las redes sociales. Más recientemente, se ha ampliado su uso a la clasificación automática de noticias, la organización de grandes cantidades de datos en foros digitales y la categorización de documentos médicos.

El modelo Naïve Bayes, a pesar de su simplicidad, presenta resultados competitivos en

comparación con algoritmos más sofisticados, particularmente cuando los datos textuales tienen una dimensionalidad elevada. No obstante, en el ámbito de los relatos paranormales, la bibliografía académica es limitada: no hay investigaciones sistemáticas que intenten categorizar este tipo de narraciones. Esto brinda la oportunidad de indagar cómo se pueden implementar las técnicas tradicionales de NLP en un campo escasamente estudiado y con un componente cultural importante.

Este trabajo busca aprovechar este método en la categorización de relatos recopilados a través de web scraping en la página "Your Ghost Story".

Metodología

El web scraping es la técnica de extraer datos de sitios web de manera automatizada, simulando la navegación de un usuario pero procesando el contenido de forma automática. Utilizando esta técnica, se generarán HTTP request con el método GET para obtener las páginas. A partir de ello, la labor se dirigirá a encontrar qué etiquetas html estaban asociadas con la información relacionada al título, a la clasificación de la historia, al lugar en donde la historia se publicó, y a la historia en sí. Para eficientar el tiempo de recopilación de los relatos, se realizará una paralelización de las peticiones. Con la información obtenida, se generará una base de datos para su posterior uso.

Dado que el texto no es directamente interpretable por los algoritmos a utilizar, se requiere transformarlo en una representación numérica. Para esto, se utilizará un proceso conocido como count vectorizer, que genera una matriz de frecuencia de términos, donde cada fila representa uno de los relatos y cada columna una palabra única, de modo que cada celda contiene el número de veces que la palabra aparece en el relato. Además de esta vectorización, también se buscará eliminar las stop words, es decir, palabras que se repiten un gran número de veces pero que en realidad no cargan con gran significado. [2]

Para este trabajo, se utilizará un clasificador Naïve Bayes, un tipo de red bayesiana. Una red bayesiana es un modelo probabilístico gráfico representado por un grafo acíclico dirigido, formalmente definido como $D = (V, E)$, donde cada vértice $v = \{X_1, ..., X_n\}$ representa una variable aleatoria del problema y las aristas E representan dependencias condicionales entre variables. Esta estructura nos permite expresar la distribución conjunta de las variables de la siguiente manera:[1]

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Padres(X_i))$$

Como se mencionó, el clasificador Naïve Bayes es un tipo de red bayesiana, donde solamente existe un nodo raíz, que corresponde a la variable a predecir Y, mientras que todos los nodos hijos que dependen únicamente de Y corresponden a las características de la observación. En este modelo, se hace la suposición de que todas las características son independientes entre sí, esto para simplificar los cálculos. Gracias a esta configuración, se puede

factorizar la distribución general de la siguiente manera:

$$P(Y, X_1, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i|Y)$$

Una observación es que en este modelo, el nodo padre (o el que corresponde a la variable a predecir) siempre es una variable categórica, mientras que los nodos hijos pueden ser tanto categóricos como cuantitativos. En el caso de que un nodo hijo sea categórico, este se distribuye siguiendo una distribución multinomial, mientras que, en el caso cuantitativo, aunque es común que se asuma que estos nodos tienen distribuciones normales, también es posible tener otros tipos de distribución, tanto paramétricos (por ejemplo una distribución tipo Poisson), como no paramétricos (como estimadores tipo Kernel).

Con esto y el teorema de Bayes, puede calcularse la probabilidad de que una nueva observación sea igual a cierta categoría de la siguiente manera:

$$P(Y = k|X = x) = \frac{P(Y = k)f(x|y = k)}{\sum_{k=1}^{K} P(Y = k)f(x|y = k)}$$

Para evaluar el clasificador, se utilizará una matriz de confusión y algunas métricas que pueden calcularse con ella. La matriz antes mencionada es un método de visualización para los resultados de un algoritmo clasificador. Usualmente, el cuadro superior izquierdo contiene la cantidad de verdaderos positivos (TP), es decir, la cantidad de predicciones correctas respecto a la categoría positiva. El cuadro inferior izquierdo muestra la cantidad de falsos positivos (FP), o la cantidad de predicciones que fueron catalogadas como positivas cuando no lo eran. El cuadro superior derecho es el número de falsos negativos (FN), que es la cantidad de predicciones que fueron catalogadas erróneamente como negativas. Finalmente, en el cuadro inferior derecho se encuentran la cantidad de negativos verdaderos (TN), que son la cantidad de predicciones en la categoría negativa catalogadas de manera correcta. [4]

Con la matriz de confusión, es posible calcular ciertas métricas para evaluar el desempeño del clasificador, entre las que se encuentran:

- Exactitud (Accuracy): Proporción de resultados verdaderos positivos (TP) más los verdaderos negativos (TN) entre el número total de predicciones. Aunque esta métrica puede no llegar a ser totalmente informativa cuando la proporción de positivos y negativos es muy desigual.

- Precisión (Precision): Proporción de verdaderos positivos (TP) dividido entre todos la suma de los verdaderos positivos más los falsos positivos. En otras palabras, es el porcentaje de casos positivos detectados.

- Sensibilidad (Recall): Proporción de verdaderos negativos (TN) respecto a todas las predicciones hechas de negativos.

- F1-score: Esta métrica es útil porque resume tanto la precisión como la sensibilización de una sola métrica. Se calcula como $\frac{2*Precision*Recall}{Precision+Recall}$.

Sin embargo, todo lo anterior aplica para categorizaciones binarias, es decir, cuando únicamente existen dos clases de la variable objetivo. Y aunque también existen matrices de confusión para variables con múltiples categorías, se decidió que para este estudio se dicotomizarán las predicciones, entrando únicamente en dos categorías: *Haunted Places* y *Other*

Para implementar el clasificador como tal se utilizarán dos librerías distintas, una de ellas es e1071, que contiene distintos algoritmos de machine learning. También, se hará uso de la librería *naivebayes*, en la que podemos agregar otros parámetros que no se pueden con la función de la primera librería con el objetivo de ver el cambio en las métricas. Por ejemplo, asumir que las variables cuantitativas no se comportan con una distribución normal, sino más bien con una distribución de Poisson, esto pues es necesario recordar que estamos hablando de frecuencias de palabras. También, existe la opción de agregar un Laplace Smoothing que se explicará a continuación.

Por último, cabe mencionar que no todas las palabras aparecen en todas las historias, por lo tanto, nuestra matriz obtenida con la vectorización de las historias tendrá múltiples celdas con ceros, lo que puede provocar problemas a la hora de calcular las probabilidades, por ejemplo, si en los datos de entrenamiento en la categoría *Haunted places* no está presente la palabra *magic*, entonces el algoritmo asumirá que si esta palabra está presente en el relato, no hay probabilidad de entrar en esta categoría, lo cual no es verdad. Para resolver este problema, se probará la utilización de la técnica *Laplace Smoothing*, la cual agrega un término de "suavizamiento" a la fórmula para el cálculo de las probabilidades condicionales [3], como se ve en la Ecuación 1.

$$P(x_i|Y = k) = \frac{N(x_i, Y = k) + \alpha}{N(Y = k) + \alpha p} \tag{1}$$

Aquí, α representa el parámetro parámetro de suavizamiento, $N(x_i, Y = k)$ el número de observaciones que contienen la palabra x_i y están en la categoría k. Mientras que $N(Y = k)$ corresponde al número total de observaciones tal que su variable objetivo es igual a k y p corresponde al número de variables. Con esto, se logra eliminar las probabilidades iguales a cero; sin embargo, es necesario establecer un parámetro α para determinar qué tanto suavizamiento se realizará, por lo que se harán múltiples pruebas usando distintos valores para este parámetro y se utilizará aquel que arroje mejores resultados.

Aplicación

Para la parte del web scraping, se parte de una función simple para la extracción de la información usando el protocolo HTTP GET. Con ello, se dividió la información en base a las etiquetas html previamente definidas (siendo estas el título, el lugar de publicación, el tipo de historia y el relato). De las 28500 llamadas lanzadas a la página, 3007 historias fueron obtenidas, las cuales sirven como base para los posteriores clasificadores.

El clasificador Naïve Bayes a ser generado consta de un nodo raíz, el cual representa las posibles categorías del relato a clasificar. Para determinar las categorías, se decidió es-

coger un género, entre los presentados por la página de "Your Ghost Story", dado que este sea el que tiene una mayor cantidad de relatos entre aquellos que fueron recopilados. Esto resultó en el género de *Haunted Places*, el cual contiene 902 relatos dentro del conjunto de datos. Con ello, cada relato se buscará clasificar entre *Haunted Places* y *Non Haunted Places*.

Posteriormente, las características que dependen de la clasificación del relato son aquellas como las emociones de la narrativa, la frecuencia en uso de palabras, y el lugar en donde se escribió. Como estas dependen de la clasificación, se da un arco de la clasificación a cada características, y estas se suponen independientes entre sí. Con este grafo en mente, se ajustarán los respectivos clasificadores, como descrito en la sección anterior.

El primer clasificador Naïve Bayes se entrenó con la librería e1071. Sin especificar algún parámetro, el modelo clasificó a las 902 observaciones del conjunto de prueba como *Haunted Places*, con lo cual se generó la matriz de confusión de la Tabla 1.

A partir de esta matriz de confusión, el modelo se mide en las 4 métricas establecidas en la metodología. Se tiene un accuracy de 0.3027, una precisión de 0.3027, un recall de 1, y un F1-score de 0.4647. El accuracy que menos de la mitad de las predicciones son correctamente clasificadas, la precisión muestra que solo el 30.27% de las predicciones de *Haunted Places* fueron verdaderamente de dicha categoría, aunque el recall indica que todos los relatos de *Haunted Places* fueron clasificados correctamente. El F1-score, que combina tanto a la precisión como al recall, indica que el modelo tiene un rendimiento como del 46.47%, que indica que se tiene una precisión y recall media. Con estas métricas, se puede esperar que el modelo siempre clasifique correctamente los relatos de *Haunted Places*, pero que clasifique erróneamente a todos los relatos que no son de este género.

Real / **Predicción**	Haunted Places	Non Haunted Places
Haunted Places	273	629
Non Haunted Places	0	0

Tabla 1: Matriz de confusión del modelo 1.

Posteriormente, se generó un clasificador con la librería naivebayes, el cual generó el mismo resultado del clasificador anterior, categorizando a todos los relatos de prueba como *Haunted Places* únicamente. La matriz de confusión es la misma, que se encuentra en la Tabla 2, resulta siendo la misma que la del modelo anterior.

Las métricas obtenidas para el segundo modelo son las mismas que en el primer modelo. Se tiene un accuracy de 0.3027, una precisión de 0.3027, un recall de 1, y un F1-score de 0.4647. La interpretación de estas métricas es la misma que en el modelo anterior, llevando a cabo una clasificación perfecta en todos los relatos con categoría *Haunted Places* y erróneamente a todos los relatos fuera de esta categoría, puesto que todas las observaciones se clasifican en una sola categoría.

Real / **Predicción**	Haunted Places	Non Haunted Places
Haunted Places	273	629
Non Haunted Places	0	0

Tabla 2: Matriz de confusión del modelo 2.

Los primeros dos modelos, aún cuando son ajustados por librerías distintas, muestran un resultado consistente en el set de prueba. Sin embargo, ambos clasifican a todos los relatos como *Haunted Places*, lo cual no está mostrando métricas muy favorables. Esto podría deberse al supuesto de normalidad de las variables continuas.

El tercer modelo, a diferencia de los dos anteriores, deja de asumir una distribución normal para las variables numéricas. En lugar de ello, genera el modelo asumiendo una distribución de Poisson. Con este ajuste, la clasificación de las observaciones de prueba ya no solamente es en Haunted Places, sino que también se clasifican en *Non Haunted Places*. Dichas predicciones generan la matriz de confusión de la Tabla 3.

Utilizando la matriz de confusión, se calculan las 4 métricas del tercer modelo. Este tiene un accuracy de 0.6874, una precisión de 0.4894, un recall de 0.7619 y un F1-score de 0.5960. Esto refleja una tasa de 0.6874 predicciones correctas del modelo, mientras que casi la mitad de las predicciones de relatos catalogados como *Haunted Places* fueron correctamente clasificados, y la otra mitad fue incorrectamente clasificada. La proporción de casos positivos reales identificados fue alta, de 0.7619, y el score que pondera tanto a la precisión como al recall, tiene un valor de casi 0.6, que no es tan alto pero tiende a indicar un buen modelo que a uno malo.

Se puede observar que fuera del recall, las demás métricas mejoran en comparación con los modelos 1 y 2, ya que no todas las observaciones son clasificadas en una sola categoría, por lo que se podría decir que el modelo tiene un mejor ajuste al asumir una distribución de Poisson para las características numéricas.

Real / **Predicción**	Haunted Places	Non Haunted Places
Haunted Places	208	217
Non Haunted Places	65	412

Tabla 3: Matriz de confusión del modelo 3.

En vista de que existen muchos ceros en los datos de entrada del clasificador, algunas de las probabilidades son 0, lo cual afecta en la predicción. Es por ello que utilizar el suavizamiento de Laplace, descrito anteriormente, podría representar una mejora en el modelo.

Para escoger el parámetro de α para este suavizamiento, se simularon múltiples clasificadores, manteniendo el supuesto de distribución Poisson para variables numéricas, con lo

cual se comparó la métrica de accuracy utilizando el set de datos de prueba. En la Tabla 4 se pueden observar los valores de α probados, así como las medidas de accuracy obtenidas, con lo cual se escogió maximizar el accuracy, lo cual sucede con $\alpha = 1$ de este subconjunto de parámetros.

α	**Accuracy**
0.05	0.7217295
0.1	0.7250554
0.3	0.7272727
0.7	0.7283814
1	0.7339246
5	0.7272727
10	0.7084257
50	0.6940133
100	0.695122

Tabla 4: Comparación de α para Laplace smoothing.

Utilizando el parámetro $\alpha = 1$, se generó entonces el cuarto clasificador, cuya matriz de confusión se muestra en la Tabla 5. Con este cuarto modelo, se tiene un accuracy de 0.7339, una precisión de 0.5478, un recall de 0.6923 y un F1-score de 0.6116. La tasa de clasificaciones correctas es del 0.73, que se considera un accuracy alto ya que más del 70% de los datos se clasifica correctamente. La precisión indica que más de la mitad de los relatos categorizados como *Haunted Places* es clasificado correctamente, y el recall muestra que el modelo es bueno identificando los relatos de *Haunted Places*.

Una observación de añadir este parámetro de suavizamiento al clasificador es que los α probados mejoraron todos el accuracy del modelo, en comparación al tercer clasificador, el cual no contaba con suavizamiento. Sin embargo, aún cuando el accuracy mejoró, el recall disminuyó significativamente, mostrando que se clasificaron menos relatos correctamente como *Haunted Places*. A pesar de esta disminución, el aumento de la precisión permitió que el F1-score aumentara también, indicando que el rendimiento del modelo es mejor con el uso del suavizamiento a cuando no se utiliza.

Real / **Predicción**	Haunted Places	Non Haunted Places
Haunted Places	189	156
Non Haunted Places	84	473

Tabla 5: Matriz de confusión del modelo 4.

Las métricas de los 4 modelos se ven condensadas en la Tabla 6.

	Accuracy	**Precision**	**Recall**	**F1-score**
Modelo 1	0.3027	0.3027	1	0.4647
Modelo 2	0.3027	0.3027	1	0.4647
Modelo 3	0.6874	0.4894	0.7619	0.5960
Modelo 4	0.7339	0.5478	0.6923	0.6116

Tabla 6: Métricas de las matrices de confusión.

Conclusión

La ambigüedad del lenguaje y la diversidad de los fenómenos hacen difícil la clasificación sistemática de los relatos, para abordarlo, se construyó un clasificador de Naïve Bayes y técnicas de procesamiento de lenguaje natural.

Las implementaciones iniciales resultaron inadecuadas, ya que clasificaron incorrectamente todas las narrativas en una sola categoría ("Haunted Places"). Esto significó un rendimiento general deficiente, con una exactitud de solo 30.27% y un recall engañoso de 100%.

Notamos una mejora cuando cambiamos el modelo a dejar de asumir normalidad y usar la distribución de Poisson (accuracy de 68.74% y F1-score de 59.60%). Mejorar y cambiar estos supuestos marcaron un buen desempeño, y todavía implementando la técnica de suavizamiento de Laplace, con un parámetro $\alpha = 1$, se obtuvo el mejor modelo, con accuracy del 73.39% y un F1-score de 0.6116.

El proceso de construir este clasificador permite comprender algo esencial que muchas veces se pasa por alto, y es que no basta con realizar un preprocesamiento básico de los datos, entrenar un modelo inicial y proceder a interpretarlo. Es fundamental analizar la naturaleza de los datos y su comportamiento para poder ajustar el modelo de manera adecuada, con lo cual, los resultados pueden acercarse más a lo deseado.

En este artículo, esto se evidenció claramente. Los primeros modelos mostraron limitaciones serias al clasificar todos los relatos en una sola categoría. Fue solo al replantear la distribución de las variables y aplicar técnicas como la asunción de Poisson y el Laplace Smoothing que se logró mejorar sustancialmente el desempeño, alcanzando métricas mucho más equilibradas y útiles.

La revisión constante de los resultados y la naturaleza de los datos es tan relevante como el modelo mismo.

Referencias

[1] Siddhant Bhattarai. *Comprehensive Guide to Probabilistic Reasoning and Bayesian Networks*. Accedido: 14-09-2025. 2025. URL: https://siddhantbhattarai.hashnode.dev/comprehensive-guide-to-probabilistic-reasoning-and-bayesian-networks.

[2] IBM. *Using CountVectorizer for NLP feature extraction*. Accedido: 14-09-2025. 2023. URL: https://www.ibm.com/reference/python/countvectorizer.

[3] Vaibhav Jayaswal. *Laplace smoothing in Naive Bayes algorithm*. Accedido: 14-09-2025. 2020. URL: https://towardsdatascience.com/laplace-smoothing-in-naive-bayes-algorithm-9c237a8bdece/.

[4] Jacob Murel Ph.D. *¿Qué es una matriz de confusión?* Accedido: 14-09-2025. 2024. URL: https://www.ibm.com/mx-es/think/topics/confusion-matrix.