Clasificación de relatos paranormales mediante un clasificador Naïve Bayes

Joshua A. Chaidez Ochoa[1], Luis C. Marrufo Padilla[1], Ángel Esparza Enríquez[1] and Gustavo A. Aguilar Torreblanca[1]

[1] *Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias, Guadalajara, Jalisco*

Abstract—Este documento presenta la implementación de un clasificador Naïve Bayes para la categorización de relatos paranormales. Se extrajeron 1000 textos narrativos para entrenar diferentes clasificadores Naïve Bayes con diferentes distribuciones, también aplicando suavizamiento de Laplace. Los resultados muestran que el modelo multiclase tuvo un desempeño muy bajo (accuracy 0.0067), mientras que al dicotomizar la variable de salida se alcanzó una precisión cercana al 0.62. Esto confirma que, pese a la simpleza del modelo y su supuesto de independencia, puede identificar patrones relevantes en escenarios binarios y proveer un modelo de clasificación decente en dicho escenario. .

Keywords—Naïve Bayes, clasificación de texto, relatos paranormales, análisis de sentimientos, redes bayesianas

I. Introducción

Desde que el internet se volvió accesible para la mayoría de la población mundial, diariamente se genera una cantidad enorme de información: historiales de transporte, imágenes, mensajes de texto, millones de compras, entre otros. Por ello, resulta fundamental administrar de manera rápida, confiable y eficaz toda esta información. En este contexto, los clasificadores Naïve Bayes ofrecen una técnica basada en redes bayesianas con una estructura específica que permite, a partir de las características de diferentes observaciones, estimar cómo debería clasificarse nueva información.

En internet existen comunidades dedicadas a lo paranormal que han generado una gran cantidad de testimonios, relacionados con fantasmas, apariciones o sucesos inusuales difíciles de explicar. En este estudio se emplearán relatos paranormales recopilados de la página "Your Ghost Stories" como base del análisis ya que contiene miles de relatos. Ante esta gran cantidad de información, la clasificación automática de texto permite explorar patrones en la redacción de los mismos. Los clasificadores Naïve Bayes son útiles en estos escenarios por su buen desempeño a pesar de su simplicidad.

El clasificador se formula como una red bayesiana en la que la clase (o etiqueta) se define como nodo padre, mientras que las características, en este caso el contenido del relato, se representan como nodos hijo. El presente trabajo tiene como objetivo implementar y evaluar un clasificador naïve Bayes para categorizar relatos paranormales obtenidos en línea, se incluyen diferentes técnicas relacionadas a los clasificadores como la suavización y se exploran los diversos resultados dentro de las matrices de confusión.

II. Metodología

a. Recopilación y limpieza de los datos

Para construir la base de datos se implementaron técnicas de web scraping sobre la página web "Your Ghost Stories" [1], donde se recopilan historias paranormales. Se realizó la extracción del título, estado, categoría y descripción de cada relato y se almacenó en un archivo CSV utilizando las bibliotecas `rvest` de R.

Para acotar el análisis, se recopilaron solo los relatos con ubicación en Estados Unidos, único país en el que incluye el campo “estado” y es de donde vienen la mayoría de narrativas. El conjunto final reúne 1,000 relatos con tres variables fundamentales: estado, categoría y descripción.

Una vez recolectados los relatos de la página, se limpió el texto. Se transformó todo el contenido a minúsculas y se eliminaron carácteres especiales, números y saltos de línea. Posteriormente, con la librería `tidytext` se realizó la tokenización, es decir, se dividió el texto en palabras. También se eliminaron palabras con longitud menor a dos carácteres y aquellas dentro del diccionario de `stop_words` de la libreria, que contiene palabras comunes que no suelen aportar mucha información en un análisis de texto.

Por último, se ordenaron en una matriz dispersa de frecuencias de palabras por documento para después agregarse a la matriz con el estado y la categoría con base a la llave de cada documento.

b. Análisis de sentimiento

Para contar con un análisis de texto de mayor profundidad, se realizó un análisis de sentimiento empleando el diccionario de `nrc`, el cual le otorga un sentimiento a una palabra dada. Con esto, se realiza una matriz de frecuencias del número de palabras que mostraron un sentimiento dado, para

completar la base de datos que se le alimentará al clasificador Naïve Bayes [2].

c. Construcción del clasificador Naïve Bayes

Se implementaron dos modelos de clasificadores Naive Bayes, empleando las librerías naivebayes y e1071. Se realizó el split en conjuntos de entrenamiento y de prueba siendo 70% y 30% respectivamente, Para reproducibilidad, se utilizó `set.seed(1310)`.

Una vez implementados los modelos se obtuvo su precisión para medir el desempeño de éstos. Se muestra a continuación:

```
Overall Statistics
               Accuracy : 0.0067
```

Se muestra un solo dato ya que la precisión para los modelos de ambas librerías fue la misma. Se puede observar que la precisión del modelo es extremadamente baja, por lo que el modelo no es capaz de predecir correctamente en absoluto. Una razón muy probable por la que la precisión de los modelos sea tan baja es la cantidad de clases que se encuentran en los datos utilizados.

d. Dicotomización del modelo

Se optó por dicotomizar el modelo para poder crear un modelo que pueda predecir entre una clase y el resto, aumentando la precisión y, por tanto, el desempeño del modelo.

Se decidió utilizar la clase "Haunted Places" a partir de que era la clase que contaba con un mayor número de observaciones dentro de la base de datos. Esta clase cuenta con un 425 de 1000 observaciones totales.

Ya una vez teniendo las clases "Haunted Places" y "Others", se realizaron nuevos modelos de clasificación empleando ambas librerías: `naivebayes` y `e1071`.

Modelos tras dicotomizar

```
                Reference
Prediction       Haunted Places Other
  Haunted Places             69    50
  Other                      60   121
               Accuracy : 0.6333
```

Nuevamente ambos modelos muestran exactamente la misma matriz de confusión. Ya los modelos muestran una precisión aceptable, por lo que pueden llegar a ser útiles a la hora de clasificar.

Si bien los modelos ya parecen ser útiles, se asume que las variables con las que se alimentó siguen una distribución normal, aunque puede no ser el caso. También, se optará por utilizar solamente la librería de `naivebayes` para ahorrar tiempo y recursos computacionales ya que ambas están regresando resultados idénticos.

e. Clasificador utilizando distribución Poisson

Se optó por realizar un modelo asumiendo una distribución Poisson en los datos para comparar su desempeño respecto al modelo realizado asumiendo una distribución normal. La matriz de confusión se muestra a continuación:

Modelo utilizando distribución Poisson

```
                Reference
Prediction       Haunted Places Other
  Haunted Places             69    53
  Other                      60   118
               Accuracy : 0.6233
```

f. Modelo utilizando Laplace Smoothing

Podemos aplicar una técnica de suavización conocida como Laplace smoothing en nuestra red con distribución Poisson. Esta técnica consiste en sumar un valor constante (α, usualmente 1) a cada conteo de ocurrencias antes de calcular las probabilidades condicionales. De esta manera, ninguna probabilidad queda en cero aunque un término no aparezca en determinada clase, lo que evita que el clasificador descarte por completo esa clase y mejora la robustez del modelo.

Modelo Poisson suavizado

```
                Reference
Prediction       Haunted Places Other
  Haunted Places             69    50
  Other                      60   121
               Accuracy : 0.6333
```

Se puede observar que el suavizado consigue mejorar el desempeño del modelo usando distribución Poisson.

`Modelo binarizando los datos`

Además de la suavización de Laplace hay otras técnicas que ayudan a un modelo cuando se tienen problemas relacionados con una matriz con exceso de ceros. Una alternativa es utilizar una distribución bernoulli (binarizar los datos): la red solo toma en cuenta la ausencia o presencia de las variables en cada observación [3].

Modelo utilizando distribución Bernoulli

```
                Reference
Prediction       Haunted Places Other
  Haunted Places             71    49
  Other                      58   122
               Accuracy : 0.6433
```

III. ANÁLISIS DE RESULTADOS

Los resultados obtenidos de la matriz de resultados y el accuracy de los diferentes modelos muestran como hay un salto considerable al dicotomizar las clases. Al trabajar con todas las clases el accuracy del modelo es extremadamente bajo, mientras que en los modelos dicotomizados, el accuracy aumenta a alrededor del 0.62 (dependiendo la distribución que usemos), esto demuestra que al simplificar las clases es más sencillo para el modelo hacer clasificaciones oportunas. Respecto al análisis de sentimientos con el diccionario NRC, si bien se integró como parte de las variables de entrada, su impacto en el desempeño del modelo fue marginal. Esto sugiere que las palabras específicas utilizadas en los relatos aportan más al proceso de clasificación que la carga emocional de los textos.

En cuanto a las distribuciones utilizadas, hay ligeras diferencias entre el rendimiento de cada una de ellas. El modelo gaussiano, la distribucion por defecto de este tipo de redes, resulto tener mejores resultados que al usar distribucion Poisson. Sin embargo, al usar la suavizacion de Laplace, Poisson

logra empatar con la Gaussiana, mostrando que usar la suavización es un método efectivo para mejorar la red. Pero la distribución que mejor rendimiento tuvo fue la Bernoulli, la cual tuvo un score de 0.6433, ligeramente mayor que las anteriores pero eso nos indica que es más útil usar la presencia de los tokens en vez de su recurrencia. También el análisis dque implementamos de sentimientos no tuvo un impacto tan siquiera notorio. Dandonos a entender que son aquellas palabras específicas que utilizan los relatos quienes hacen el proceso de clasificación más efectivo que la carga emocional impresa en el texto.

Estas difererencias son muy pequeñas de modelo a modelo pero nos pueden ayudar a entender la estructura de nuestros datos y saber cuáles son los siguientes pasos para mejorar nuestro modelo de clasificación.

IV. Conclusiones

Al dicotomizar la variable de clase, el rendimiento mejora enormemente, con menos categorías disminuye la cantidad de errores, además debido a la gran cantidad de relatos de la categoría dicotomizada, hace sentido hacer este procedimiento. Cabe mencionar que el modelo multiclase inicial falló rotundamente, ya que obtuvo un accuracy extremadamente bajo, en este caso las 20 categorías originales no fueron óptimas para la resolución de el caso. En los modelos posteriores usando clases dicotomizadas, el Naïve Bayes gaussiano alcanzó una accuracy de 0.63, razonable dada la complejidad de los textos, y aunque el Naïve Bayes Poisson podría parecer más adecuado por trabajar con conteos, en la práctica mostró una accuracy ligeramente menor aunque al suavizarlo empató con el gaussiano. Por último, la red que usa Bernoulli fue la que por una ventaja pequeña tuvo un mejor desempeño, consistente con la abundancia de ceros.

Algunas posibles consideraciones para trabajos futuros podrían ser el uso de TF-IDF para identificar términos importantes y únicos de un documento. Es importante denotar que pese a la subjetividad y las limitaciones que implica trabajar con textos de este tipo naïve bayes tuvo unos resultados bastante buenos considerando que es un modelo relativamente simple, pero es un buen punto de partida para tareas de clasificación de texto, en casos más exigentes se deberán implementar modelos que logren capturar la tan compleja naturaleza que son los textos especialmente los narrativos.

References

[1] "Your ghost stories," https://www.yourghoststories.com/, 2006–2025, sitio web con historias reales de fantasmas y reportes paranormales. [Online]. Available: https://www.yourghoststories.com/

[2] J. Silge and D. Robinson, *Text Mining with R: A Tidy Approach*. Sebastopol, CA: O'Reilly Media, 2017. [Online]. Available: https://www.tidytextmining.com/

[3] S. Eyheramendy, D. D. Lewis, and D. Madigan, "On the naive bayes model for text categorization," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, C. M. Bishop and B. J. Frey, Eds., vol. R4. PMLR, 03–06 Jan 2003, pp. 93–100, reissued by PMLR on 01 April 2021. [Online]. Available: https://proceedings.mlr.press/r4/eyheramendy03a.html