Universidad Nacional Autónoma de México

Facultad de Ciencias

Clasificadores basados en prototipos de temas recurrentes

T E S I S

que para obtener el título de:

Licenciado en Ciencias de la Computación

presenta:

Alejandro Martínez Torres

Director de Tesis:

Dr. Adrián Pastor López Monroy

Ciudad Universitaria, Ciudad de México 2022

Para mi padre
que siempre me ha brindado su apoyo incondicional

Agradecimientos

A mi asesor, el Dr. Adrián Pastor López Monroy, quien con su experiencia, conocimiento, confianza e infinita paciencia me guio en esta aventura.

A la Universidad Nacional Autónoma de México, por brindarme mi educación y permitirme conocer a gente tan brillante como apasionada.

A mis amigos, por ayudarme a encontrar mi camino y brindarme alegría en cada día de estos años.

A Katia, mi novia, que me motivó para seguir adelante sin importar los obstáculos.

A mi familia, por siempre creer en mí y apoyarme.

Resumen

Hoy en día, con el auge de las plataformas de *streaming*, cada vez se le da más peso a los sistemas clasificadores en el ámbito cinematográfico. Las plataformas buscan presentar su contenido en grupos óptimos y entendibles para el consumo de sus usuarios. Para ello se necesita de datos que puedan ser usados para asistir en este proceso. En este trabajo se propone el uso de **temas recurrentes** para mejorar las tareas de clasificación. De forma resumida, un tema recurrente es una situación general que ocurre en una variedad de instancias; es un cliché sin la connotación negativa. Para poder hacer uso de los temas recurrentes, creamos un corpus que consta de 2,324 películas y de los 24,756 temas recurrentes que aparecen en estas. La información de los temas recurrentes fue extraída de la *wiki* TV Tropes.

En esta tesis abordamos dos tareas usando temas recurrentes: predicción del género cinematográfico, mostrado en IMDb, y predicción de clasificación por edades, asignada por la *Motion Picture Association of America* (MPAA). Para la predicción del género cinematográfico usamos específicamente los temas recurrentes que están asociados a los 7 géneros básicos. Esta información teórica se encuentra en la *wiki* TV Tropes y relaciona a cada género con un conjunto de temas recurrentes. Para el caso de la clasificación por edades, donde no se cuenta con etiquetas para los temas recurrentes, desarrollamos un método que permite seleccionar los temas recurrentes relevantes para la clasificación por edades y agruparlos en conjuntos de temas recurrentes similares asociados con alguna de las categorías. La base de nuestro método es el uso de la prueba $\boldsymbol{\chi^2}$ de Pearson para seleccionar temas recurrentes relevantes. También usamos técnicas como Word2Vec y K-means para agrupar temas recurrentes relacionados, mitigando el hecho que muchos temas recurrentes aparecen solamente en una o dos películas. Consideramos que uno de los aportes

en esta tesis es el método desarrollado, ya que provee una forma de obtener conjuntos de temas recurrentes asociados a las categorías dadas. Lo cual no solamente es de ayuda a predicciones, sino también para comprender mucho mejor la naturaleza de cada categoría. De igual forma, podría servir para entender mejor clasificaciones previas hechas por seres humanos y encontrar posibles sesgos.

Tanto para la predicción de género cinematográfico como para la predicción de clasificación por edades, usamos la información respectiva para realizar tres predicciones comparando tres algoritmos de clasificación distintos (SVM, Random Forest y KNN). Esto con el fin de tener un panorama más grande sobre la efectividad del uso de temas recurrentes en dichas tareas. Como punto de comparación, realizamos nuevas predicciones con los mismos 3 algoritmos, pero usando los diálogos de películas en tres distintas representaciones. Dichas representaciones utilizan técnicas clásicas en el ámbito como: Bolsa de palabras (BoW, por sus siglas en inglés de *Bag of Words*), Tf-idf y Word2Vec. De igual forma, también comparamos las predicciones realizadas con el método desarrollado, pero con variaciones que simplifican este. Los resultados de la evaluación de predicción usando temas recurrentes son comparables a los que usan diálogos, aunque en un rango menor, mientras que la predicción que utiliza ambas presenta una mejora considerable a cualquiera de las predicciones con los datos individuales. Lo cual muestra que los temas recurrentes no solamente proporcionan información útil para tareas de esta índole, sino que proporcionan información complementaría a la obtenida en otros medios. De tal manera, podemos concluir que los temas recurrentes son una fuente de datos explotable para estos y otros problemas similares, abriendo nuevos horizontes en esta área. Esto con el fin de colaborar en los esfuerzos para satisfacer el voraz apetito de datos de las cada vez más demandantes y prodigiosas Redes Neuronales.

Índice general

Capítulo 1

Introducción

Recientemente, a causa de las plataformas de *streaming*, el uso del aprendizaje de máquina para la clasificación de películas y series ha tenido un gran interés. Las plataformas buscan presentar su contenido en grupos óptimos y entendibles para el consumo de sus usuarios. Estos mismos algoritmos, una vez entrenados correctamente, pueden ser utilizados para clasificar automáticamente información masiva, incluso en productos distintos que cuenten con características (utilizadas para su clasificación) similares. Por ejemplo, los algoritmos basados en texto (inicialmente creados para los diálogos o la descripción de una película) pueden ser utilizados en libros o revistas (adaptando a su contenido), incluso pueden ser usados por un conjunto de obras de distintos medios [25].

Tradicionalmente, para lograr estas clasificaciones o predicciones en general se hace uso de información del personal involucrado en la película (ej. actores, directores y guionistas), información de producción (ej. presupuesto y fecha de estreno), información sobre la narrativa (ej. diálogos) o sobre la audiencia (ej. calificaciones de usuarios a películas). Sin embargo, casos donde se haga uso de elementos teóricos o de análisis de las obras mismas para tareas de clasificación son mucho más escasos. Esto se debe a la complejidad de obtener dichos datos a gran escala; ahora, gracias al internet, se han creado gran variedad de proyectos colaborativos de magnitudes antes inimaginables. Tal es el caso de proyectos como Wikipedia, cuya información abarca prácticamente toda clase de temas, y de *wikis* especializadas en temas específicos. La información provista en estos medios es tan vasta que hace posible su uso en el aprendizaje de máquina, donde se necesita una gran cantidad

de datos para lograr resultados positivos. Algunas de estas *wikis* proveen información sobre películas en el aspecto técnico narrativo a gran detalle y escala.

1.1. TV Tropes y Temas Recurrentes

TV Tropes es una *wiki* de contenido libre que contiene convenciones y análisis de obras de ficción y trabajos creativos de medios como literatura, cómics, videojuegos, publicidad, películas y series [40]. La *wiki* es pública; constantemente se expande y revisa por los usuarios registrados en ella. Es conocida principalmente por sus múltiples artículos de temas recurrentes y la enumeración de obras que los presentan.

En este trabajo un tema recurrente se refiere a los tropos (*tropes*), herramienta narrativa que permite describir una situación que fácilmente puede ser reconocida por la audiencia. Un tropo es la sustitución de una situación particular por un tema más general o de sentido figurado. Es lo que coloquialmente se llama un cliché, pero sin la connotación negativa; es una imagen universalmente identificada e infundida con varias capas de significado contextual que crean una nueva metáfora [28].

Un ejemplo es *Amnesiac Hero*, tema recurrente que hace referencia a cuando el personaje principal tiene amnesia. Este tema recurrente engloba todos los elementos que aparecen en una historia con dicho tema, volviéndose el arquetipo para historias con esta circunstancia. En la Figura 1.1, podemos observar cómo se presenta en la página TV Tropes y en la Figura 1.2 se muestran parte de la variedad de medios en los que se usa este tropo.

Otro ejemplo es *The Eureka Moment* (el momento Eureka), que se refiere a todas las situaciones donde un problema complejo se resuelve en un momento de epifanía. Este tema recurrente ha sido usado múltiples veces por novelistas como Agatha Christie y aparece en una gran variedad de películas como: Duro de Matar, Cuestión de Honor, Hombres de Negro, Una mente brillante y muchas más.

De cierta forma, los temas recurrentes proveen análisis temático de las obras y describen a detalle elementos de la trama. Si se cuenta con suficientes temas recurrentes, es posible

[2] `https://tvtropes.org/pmwiki/pmwiki.php/Main/AmnesiacHero`, 11 de mayo del 2022.
[4] `https://tvtropes.org/pmwiki/pmwiki.php/Main/AmnesiacHero`, 11 de mayo del 2022.

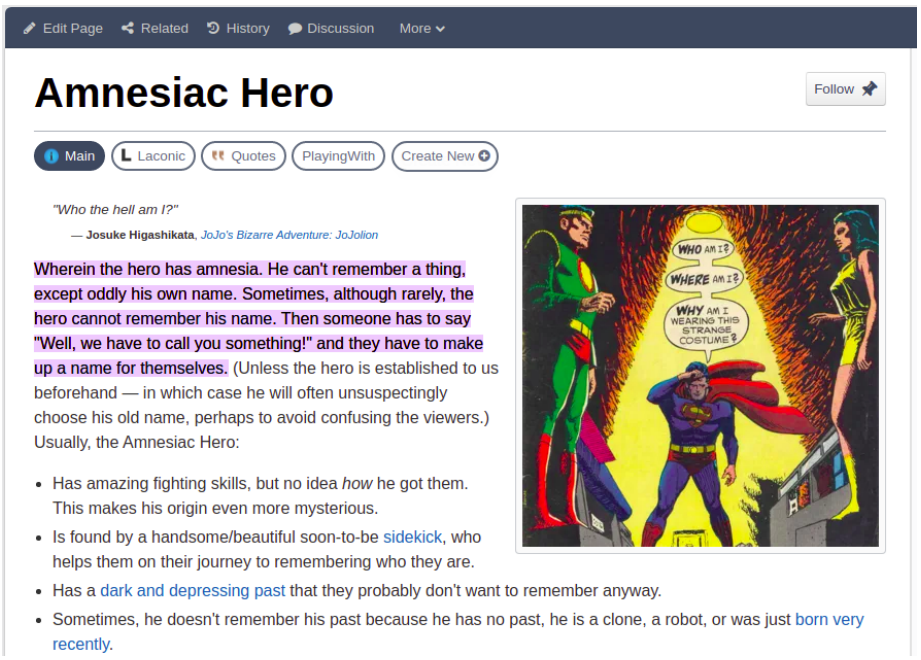

Figura 1.1: Imagen sacada de TV Tropes donde se describe el tema recurrente de *Amnesiac Hero*[2](héroe con amnesia).

obtener una idea general del tono de una película u otro contenido. Además, son elementos frecuentemente presentes en cualquier tipo de narrativa que relacionan una obra con otras a nivel temático. Por ello, planteamos la hipótesis de que pueden ser información útil para la clasificación de películas en múltiples áreas.

1.2. Objetivo

- El objetivo de esta tesis es aprovechar la información provista por los temas recurrentes para lograr mejores resultados en las tareas de clasificación en el ámbito cinematográfico.

En el presente trabajo se busca averiguar si los temas recurrentes pueden asistir en tareas de clasificación en el ámbito cinematográfico. Es de nuestro entender que este tipo de datos no han sido usados en tareas de esta índole en trabajos previos. Proponemos usar los temas recurrentes en dos tareas de clasificación de películas de naturaleza suficientemente distinta para poder observar el efecto que tienen. Las tareas elegidas son: predicción del género cinematográfico, mostrado en IMDb, y predicción de clasificación del público apto

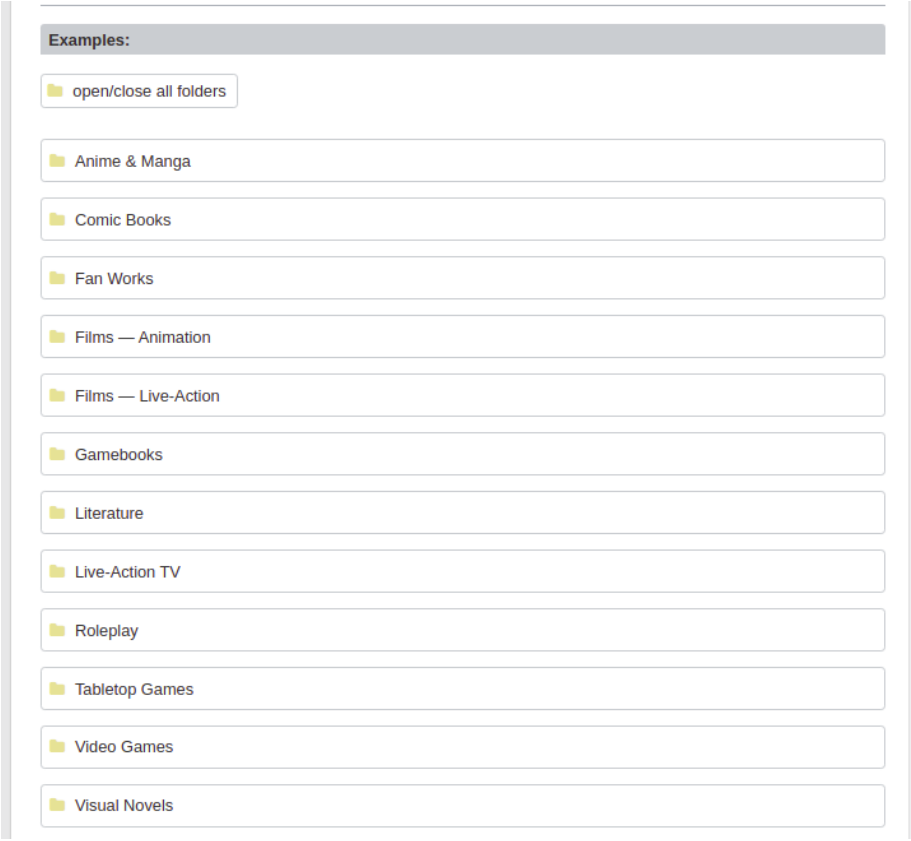

Figura 1.2: Imagen sacada de TV Tropes, en esta sección de la página[4]se muestran múltiples ejemplos donde se presenta el tema recurrente. Cada uno de los folders que se presentan en la imagen contiene una lista de distintos productos pertenecientes al medio correspondiente.

para ver la película acorde a su edad, asignada por la *Motion Picture Association of America* (MPAA). De igual forma, buscamos comparar la información provista por los temas recurrentes con la obtenida en otras fuentes más tradicionales. Esto con el fin de comparar su desempeño y saber si al usar ambos conjuntos de datos se logra una mejora en los resultados o simplemente la información que proveen los temas recurrentes se puede encontrar en otros medios más convencionales. Como punto de comparación, escogimos los diálogos de películas, ya que de cierta forma ambos describen elementos de la trama y dan una idea del tono de la película.

Para lograr esto, primero, buscamos crear un corpus de temas recurrentes. Segundo, diseñar estrategias para extraer la información más relevante de los temas recurrentes (para cada tarea específica). Tercero, realizar las predicciones usando la información de los temas recurrentes. Cuarto, hacer múltiples predicciones usando distintas representaciones

de los diálogos. Quinto, nuevamente, realizar predicciones ahora usando en conjunto la información de los diálogos y de los temas recurrentes. Por último, comparar el resultado de las predicciones realizadas.

1.2.1. Preguntas de Investigación

La pregunta de investigación que se abarca en este trabajo es:

- ¿En qué grado un sistema clasificador de películas se ve beneficiado al agregar información de temas recurrentes y cuál es la mejor forma de representar esta información?

La hipótesis de este trabajo es que podemos aprovechar la información de temas recurrentes presentes en una película para las tareas de predicción del género cinematográfico y predicción de clasificación por edades. Creemos que se pueden usar los temas recurrentes para realizar estas dos tareas y que las predicciones basadas en diálogos también se verán beneficiadas al agregar la información de temas recurrentes, presentando un mejor resultado.

Idea Intuitiva

La razón por la cual consideramos que los temas recurrentes pueden servir en tareas de clasificación de películas es porque toda clasificación divide a las películas acorde a su contenido y, a su vez, los temas recurrentes definen el contenido de una película. De hecho, la mayoría de las clasificaciones en la industria no tienen reglas precisas, están más bien definidas por un conjunto de convenciones o situaciones, que muchas veces pueden ser descritas como temas recurrentes o conjuntos de varios temas recurrentes. Por ejemplo, los géneros cinematográficos suelen definirse más por las emociones que se desea causar en el espectador que por un conjunto de reglas. Las películas buscan causar estas emociones utilizando situaciones similares a las que previamente han producido estas emociones en otras obras. Estas situaciones justamente son temas recurrentes, de tal manera que se puede percibir al género como un conjunto de temas recurrentes enfocados en emitir cierta emoción. Por ello es que decidimos realizar los experimentos y comprobar nuestra hipótesis.

Proceso realizado

El proceso que se realiza en esta tesis es el de describir la construcción de características a base de temas recurrentes, las cuales sean lo suficientemente relevantes para poder ayudar en las tareas de clasificación propuestas. En el primer problema se busca usar la información teórica para identificar y agrupar los temas recurrentes. Para el segundo problema, a falta de recursos teóricos, se busca emular el mismo proceso con métodos de filtración y agrupación. La selección de temas recurrentes se hará con la prueba $\boldsymbol{\chi^2}$ de Pearson. Por otro lado, para la agrupación primero creamos una representación vectorial de cada tema recurrente usando sus descripciones, Word2Vec y las frecuencias de cada palabra. Estas representaciones son agrupadas con el algoritmo K-means creando las características deseadas para el proceso de selección.

Usando estas técnicas y comparando con las predicciones basadas en diálogos, observamos que los temas relevantes pueden ser de gran ayuda para las tareas de clasificación. Agregar esta información ayudó, en promedio, a subir más de 3 puntos el *F1 Score* de las predicciones. Además, las técnicas utilizadas para construir características pueden ser de utilidad para entender categorías no especificadas, así como analizar posibles sesgos.

1.3. Organización de la tesis

El presente trabajo se divide en 6 capítulos. El Capítulo 1 sirve de introducción a la tesis; presentando el objetivo de esta tesis, los conceptos más importantes y una idea general del contenido de este trabajo. En el siguiente capítulo, Capítulo 2, se presenta el soporte teórico, el cual define todos los conceptos necesarios para entender el resto del documento. En el Capítulo 3 presentamos el trabajo relacionado, sin el cual no se hubiera tenido la inspiración para realizar el presente trabajo. En el Capítulo 4 exponemos la metodología que siguieron los experimentos y el trabajo previo necesario; es en este capítulo que se presentan los métodos desarrollados para aprovechar los temas recurrentes. En el Capítulo 5 se encuentra toda la información sobre los experimentos y su evaluación. Finalmente, en el Capítulo 6, presentamos nuestras conclusiones y discutimos el trabajo a futuro.

Capítulo 2

Marco teórico

En este capítulo presentamos los conceptos básicos que fundamentan el presente trabajo. En la Sección 2.1 mostramos diferentes modelos usados para la representación de un texto, desde la representación one-hot hasta los conceptos de encaje de palabras usando Word2Vec [23]. Posteriormente, en las Secciones 2.2 y 2.3 presentamos, respectivamente, el algoritmo K-means para agrupación de elementos y la prueba $\boldsymbol{\chi^2}$ de Pearson para selección de características. Finalmente, la Sección 2.4 está dedicada a describir los tres algoritmos clasificadores utilizados en esta tesis: SVM, Random Forest y KNN.

2.1. Codificaciones y representaciones de palabras

Lamentablemente, las palabras no tienen ningún significado para una computadora. A lo largo de la historia de la computación se han creado distintos modelos para representar las palabras y poder hacer uso de la información provista por estas. Cada modelo tiene distintas ventajas y desventajas, desde la facilidad de crear la representación hasta la cantidad de información que transmite.

Codificación one-hot de palabras

La codificación one-hot [35], una de las representaciones más simples, consiste en hacer una biyección de un conjunto de elementos a un conjunto de secuencias de bits válidos.

Donde cada secuencia de bits debe tener exactamente un bit con valor 1 y el resto con valor 0. Por ejemplo, en la Tabla 2.1, representamos la codificación one-hot del siguiente conjunto de palabras: {de, la, *one-hot*, oración, representación}.

Palabra	**One-hot**
de	00001
la	00010
one-hot	00100
oración	01000
representación	10000

Tabla 2.1: Representación One-hot del conjunto ejemplo.

De igual forma podemos representar la codificación de una oración con la representación one-hot de cada una de las palabras que la forman. Por ejemplo: "Representación *one-hot* de la oración" = **[[10000], [00100], [00001], [00010], [01000]]**

2.1.1. Modelo Bolsa de Palabras

El modelo Bolsa de palabras o BoW por sus siglas en inglés (*Bag of Words*) es un método para representar secuencias de palabras, como oraciones o documentos, con el costo de perder su orden. La idea es tomar la codificación *one-hot* de la secuencia de palabras y sumarla [35]. Usando el ejemplo de la Sección anterior, en la Tabla 2.2 presentamos un ejemplo de varias oraciones en su representación BoW. Podemos observar que la representación BoW de un texto toma en cuenta cada una de las palabras presentes en la secuencia y muestra la frecuencia de cada término usado; sin embargo, se pierde el orden en que se presentan las palabras.

Oración	**BoW**
one-hot	00100
la representación	10010
representación one-hot de la oración	11111
la representación de la oración	11021

Tabla 2.2: Representación BoW de cada oración. Nótese que todas las oraciones están formadas únicamente por palabras pertenecientes al conjunto ejemplo presentado en la Tabla 2.1.

Una ventaja de este modelo es que la representación BoW de documentos similares también será similar de cierta manera, ya que es posible utilizar elementos y métricas que aprovechen el espacio vectorial donde se hallan las representaciones. Por ejemplo, la similitud entre documentos se puede aproximar usando la distancia euclidiana entre sus representaciones. En este caso asumimos que dos documentos son similares si usan las mismas palabras y no tanto por el uso particular de estas. Por ejemplo, dos artículos médicos compartirán más palabras en común que comparadas con un artículo matemático. Es decir, temas similares usan palabras similares, por ende las representaciones BoW de documentos pertenecientes a la misma área serán más similares entre sí que con otros artículos de índole distinta. Sin embargo, es importante tomar en cuenta la cantidad de palabras usadas por cada documento, puesto que un documento de un párrafo de longitud cubrirá mucha menor cantidad de palabras que un documento mucho más extenso como un libro. Por ello, es imperante regularizar los vectores obtenidos con este modelo. Otro factor a tomar en cuenta son las palabras en común de dos documentos. Muchas veces las palabras más usadas son artículos o conjunciones. Esto puede afectar en gran medida la similitud entre dos documentos en su representación BoW, sin que sean similares a entender humano, ya que algunas palabras no proporcionan información del texto en sí, simplemente son palabras comúnmente utilizadas en el lenguaje.

2.1.2. Tf-idf

Tf-idf se refiere a la frecuencia de un término dividida por el número de documentos en los que aparece [19]. La frecuencia de un término es el número de veces que una palabra aparece en un documento, formalmente:

$$\boldsymbol{tf(t,d)} = \# \text{ de veces que aparece la palabra } \boldsymbol{t} \text{ en el documento } \boldsymbol{d} \tag{2.1}$$

Usualmente, BoW representa un documento como un vector que muestra la frecuencia de todos los términos utilizados en un texto. En el modelo de BoW todas las palabras del documento tienen el mismo peso. Por el contrario, en Tf-idf la frecuencia inversa de un documento sirve para dar más relevancia a unas palabras que a otras. Por ejemplo, palabras

que aparecen en todos los documentos no logran otorgar información que diferencie a un documento de otro y por ende no deberían tener la misma relevancia que el resto de las palabras. Considere los artículos que aparecen muy comúnmente en las oraciones, en ese caso los artículos no son tan relevantes como el sujeto o la acción. Lo que busca la frecuencia inversa es darle más valor a las palabras de un texto que son más discriminativas, palabras que puedan ayudar a diferenciar un texto de otro [29]. Para lograr lo anterior se define a la frecuencia inversa de un documento como:

$$\boldsymbol{idf}(\boldsymbol{t}, \boldsymbol{D}) = \log \frac{|\boldsymbol{D}|}{|\boldsymbol{d} \in \boldsymbol{D} : \boldsymbol{t} \in \boldsymbol{d}|} \tag{2.2}$$

Donde $\boldsymbol{D}$ es un conjunto de documentos, $|\boldsymbol{D}|$ es el número de documentos que lo conforman y $|\boldsymbol{d} \in \boldsymbol{D} : \boldsymbol{t} \in \boldsymbol{d}|$ es el número de documentos donde el término $\boldsymbol{t}$ aparece. De esta forma, mientras el término $\boldsymbol{t}$ aparece en más documentos, menos valor tiene su *idf*. Por otro lado, a menor cantidad de documentos mayor será su valor *idf*.

Para calcular el Tf-idf se multiplica la frecuencia de término y la frecuencia inversa de documento:

$$\boldsymbol{tfidf}(\boldsymbol{t}, \boldsymbol{d}, \boldsymbol{D}) = \boldsymbol{tf}(\boldsymbol{t}, \boldsymbol{d}) \times \boldsymbol{idf}(\boldsymbol{t}, \boldsymbol{D}) \tag{2.3}$$

De tal manera, se puede representar que tanto valor tienen cada palabra en cada texto de un conjunto de textos. La representación de un texto usando Tf-idf se puede obtener multiplicando la representación BoW de un documento por un vector donde cada entrada tiene el valor *idf* de cada palabra (en el mismo orden que BoW). Obteniendo un vector que en cada entrada tiene el Tf-idf de un término distinto. Es decir, cada entrada del vector representa la importancia de cada palabra en el texto.

2.1.3. Word2Vec

Actualmente, Word2Vec [24] es una de las técnicas de representación de palabra más populares, esto se debe a su gran efectividad para capturar información sintáctica y semántica del lenguaje. Word2Vec permite interpretar el significado semántico con la distancia entre las representaciones de palabras. Lo cual se logra al relacionar cada palabra con el

contexto en donde es utilizada. Word2Vec crea un representación de las palabras usando una red neuronal [6] que aprende de manera supervisada la relación entre una palabra y su contexto dentro de una ventana (un número determinado de palabras que se encuentran alrededor) [30].

Para entrenar la red neuronal (usada por Word2Vec) se necesita un corpus que conste de texto, donde se haga uso en lenguaje natural de las palabras que se quieren representar. Este corpus se itera obteniendo una palabra objetivo y una ventana de palabras (el contexto de una palabra). Dada una palabra objetivo, una ventana es un número (previamente definido) de palabras que se encuentran al rededor de la palabra objetivo; esto de cierta forma proporciona el contexto en que se usa cada palabra. La red neuronal tiene como objetivo, dado un contexto, predecir la palabra objetivo. Dicha red, representada en la Figura 2.1, consta de:

- Una capa de entrada, que tiene tantas neuronas como palabras diferentes haya en el corpus. Cada neurona es la representación *one-hot* de alguna de las palabras del corpus.
- Una capa oculta que tiene la misma cantidad de neuronas que la dimensión que tendrá la representación de Word2Vec de cada palabra.
- Una capa de salida que tiene tantas neuronas como palabras diferentes en el corpus. Cada una de estas neuronas usa la función de activación *softmax* y su resultado representa si se predice que una palabra pertenece al contexto dado.

Una vez entrenada, la red neuronal ha aprendido la relación entre una palabra y su contexto. Lo cual es de gran utilidad, ya que palabras similares tienen contextos similares. Finalmente, la representación de las palabras se encuentra en la matriz que representa el paso de la capa de entrada a la capa escondida [30].

Para representar un texto con Word2Vec se puede tomar el promedio de la representación en Word2Vec de las palabras utilizadas en el texto. También se puede representar un texto utilizando la representación de palabras Word2Vec y el Tf-idf de un texto. Para esto se saca el promedio del producto del idf, la representación Word2Vec y la frecuen-

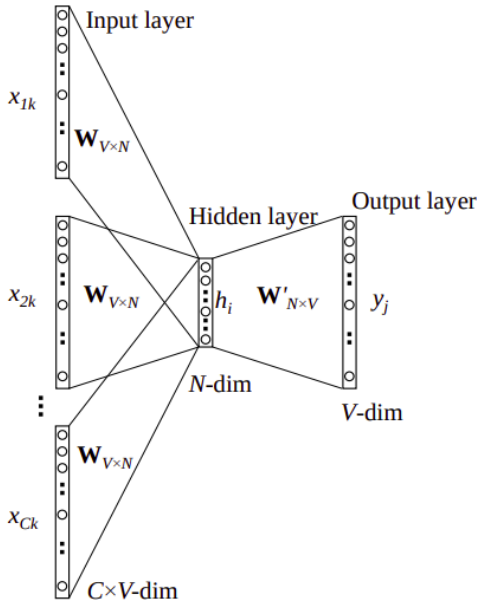

Figura 2.1: Imagen representativa de la red neuronal usada para Word2Vec usando la arquitectura explicada, CBOW (por sus siglas en inglés, *Continuous bag-of-word model*) [30]

cia de cada palabra usada en el texto. Esta última representación permite darle el valor correspondiente a cada palabra y simultáneamente relacionar temáticamente.

2.2. K-means

K-means es un algoritmo de agrupamiento de observaciones. Lo que se busca es colocar en el mismo grupo las observaciones similares. Para definir similitud se usan las características de las observaciones. Donde observaciones con características similares deberían estar en el mismo grupo. Para agruparlas se asigna a cada observación un vector que representa una codificación de sus $\boldsymbol{d}$ características. Posteriormente, se agrupan estos vectores en *clusters* que definen los grupos de las observaciones.

Una forma de representar un *cluster* de vectores es mediante el centroide. El centroide es el vector promedio de todos los vectores dentro del *cluster*. Dado un *cluster* de vectores $\boldsymbol{V}$, su centroide $\boldsymbol{\mu}$ se calcula de la siguiente forma:

$$\vec{\boldsymbol{\mu}}(\boldsymbol{V}) = \frac{\mathbf{1}}{|\boldsymbol{V}|} \sum_{\vec{\boldsymbol{v}} \in \boldsymbol{V}} \vec{\boldsymbol{v}} \tag{2.4}$$

El algoritmo *K-means* busca construir los *clusters* de tal manera que la distancia entre el centroide de un *cluster* y la de los vectores que lo integran sea lo menor posible.

2.2.1. Proceso

Dado un conjunto de $\boldsymbol{n}$ elementos, *K-means* busca agrupar los $\boldsymbol{n}$ elementos en $\boldsymbol{k}$ grupos de observaciones similares, donde $\boldsymbol{k} \leq \boldsymbol{n}$ [42]. Para agrupar las observaciones, primero se les asigna un conjunto de $\boldsymbol{n}$ vectores de dimensión $\boldsymbol{d}$ que los represente: $\{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots, \boldsymbol{x_j}, \ldots, \boldsymbol{x_n}\}$. Posteriormente, usando los vectores resultantes se selecciona de manera aleatoria $\boldsymbol{k}$ centroides para generar un conjunto $\boldsymbol{S}$ de $\boldsymbol{k}$ *clusters*. Realizando iterativamente el siguiente algoritmo se mejoran los centroides que definen los *clusters*:

1. Para cada vector de observaciones $\boldsymbol{x_j}$ se calcula el centroide $\boldsymbol{\mu_i}$ más cercano. Seleccionando específicamente uno en caso de empate.

2. Cada vector $\boldsymbol{x_j}$ se asocia con solamente un *cluster* $\boldsymbol{S_i}$, representado por su centroide $\boldsymbol{\mu_i}$, con el cual tuvo la menor distancia.

3. Una vez asignados todos los vectores de observaciones y habiendo redefinido los *clusters* de $\boldsymbol{S}$, se recalculan los nuevos centroides.

Este proceso se realiza hasta que ningún vector de observaciones cambie de *cluster* o se llegue a cierto número de iteraciones, definido previamente. Como se mencionó anteriormente, el objetivo del algoritmo es agrupar las observaciones en $\boldsymbol{k}$ *clusters* tal que se minimice la distancia de cada vector a su centroide. Es decir, seleccionar el mejor conjunto de *clusters* para $\boldsymbol{S}$ tal que:

$$\underset{\boldsymbol{S}}{\arg\min} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{x_j} \in \boldsymbol{S_i}} ||\boldsymbol{x_j} - \boldsymbol{\mu_i}||^{\mathbf{2}} \tag{2.5}$$

2.3. Prueba χ^2 de Pearson

Poseer una gran cantidad de características en un conjunto de observaciones puede llegar a ser problemático al momento de hacer predicciones con modelos. Esto se debe a que distintas características proveen distinto grado de información y puede haber casos donde ciertas características solo metan ruido al modelo; es decir, no provean información útil. Un método para selección de características es la prueba $\boldsymbol{\chi^2}$ de Pearson. Esta mide discrepancia entre una distribución observada y una esperada teóricamente. Esto sirve para probar independencia de dos variables entre sí. Para ello se calcula la distribución asumiendo que ambas variables son independientes y se compara con la distribución observada (la que se presenta en la realidad). Dicha comparación se realiza con la fórmula que se muestra a continuación:

$$\boldsymbol{\chi^2} = \sum_{i}^{n} \frac{(\text{ observada }_{\boldsymbol{i}} - \text{teórica}_{\boldsymbol{i}})^{\boldsymbol{2}}}{\text{teórica}_{\boldsymbol{i}}} \tag{2.6}$$

Se puede observa que a mayor similitud entre ambas distribuciones más se aproxima $\boldsymbol{\chi^2}$ a 0. Y cuanto mayor sea $\boldsymbol{\chi^2}$, menos probable se vuelve la hipótesis que generó la distribución teórica.

Para rechazar o aceptar la hipótesis se usan las tablas de significación estadística de la distribución de probabilidad gamma, de las cuales $\boldsymbol{\chi^2}$ es un caso particular [39]. De igual manera, se usan los grados de libertad de las observaciones y características. Sin embargo, para la selección de características en modelos de predicción esto último no es esencial.

La prueba $\boldsymbol{\chi^2}$ de Pearson se pueden usar como criterio de selección de características, ya que indica en qué medida la distribución observada y una teorizada difieren. Se puede teorizar que una característica y lo que se intenta predecir son independientes y ver que tan cierta resulta esta hipótesis. Específicamente, se toma cada característica $\boldsymbol{c_i}$ y la característica a predecir $\boldsymbol{c_p}$, se calcula $\boldsymbol{\chi^2}$ asumiendo que son variables independientes. Posteriormente, se seleccionan las $\boldsymbol{c_i}$ que tengan los valores más bajos de $\boldsymbol{\chi^2}$, ya que son las características que menor probabilidad de independencia con $\boldsymbol{c_p}$ tienen. Las características a seleccionar pueden depender de un número específico de características deseadas o por un umbral de significancia estadística [27].

2.4. Clasificadores

En el aprendizaje supervisado se usa un modelo para "aprender" la relación entre ejemplos de entrada y su salida correspondiente. Las tareas que usan como datos de entrenamiento un vector con su correspondiente salida para poder predecir futuras instancias, son conocidas como problemas de aprendizaje supervisado. Cuando la tarea es asignar cada vector a un número finito y discreto de categorías, a esto se le denomina un problema de clasificación [8]. Se llaman clasificadores a los algoritmos que buscan resolver problemas de clasificación. Como se mencionó previamente, estos clasificadores usan los datos de entrenamiento para intentar descifrar como los datos de entrada se relacionan con las categorías correspondientes. Una vez que un clasificador termina su entrenamiento, es capaz de intentar predecir la clase a la que pertenecen nuevos vectores de entrada. De tal manera que un clasificador entrenado permite obtener la categoría a la que pertenecen los datos otorgados con cierto nivel de certeza, tarea que es de gran utilidad.

2.4.1. Máquinas de vectores de soporte (SVM)

Las máquinas de vectores de soporte, comúnmente llamadas SVM por sus siglas en inglés (*Support Vector Machines*), son un conjunto de algoritmos de aprendizaje supervisado usados para tareas de clasificación, regresión o detección de valores atípicos [26]. Para lograr que un SVM clasifique en categorías se provee, para la etapa de entrenamiento, un conjunto de datos etiquetados que son representados como vectores de la misma dimensión. Posteriormente, SVM construye un conjunto de hiperplanos que separen de manera óptima las diferentes instancias de categorías [11], creando un criterio para separar estos datos de entrenamiento. En la Figura 2.2 se muestra un ejemplo de como un hiperplano separa los datos en dos distintas clases. Estos mismos criterios de separación son usados para clasificar futuros datos no etiquetados y obtener su posible clasificación. SVM es uno de los clasificadores más comúnmente utilizados y de los más efectivos [6].

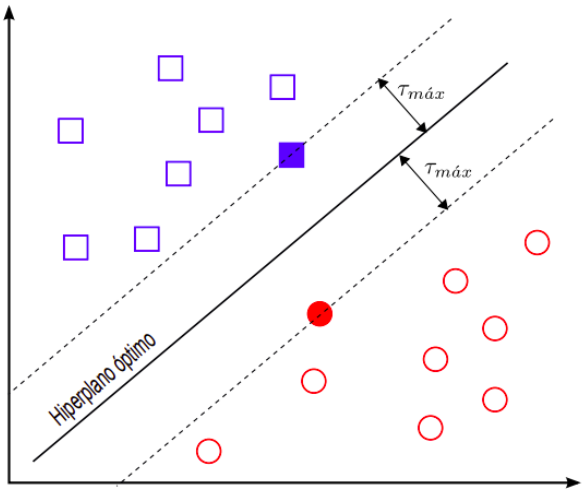

Figura 2.2: Imagen de un hiperplano que separa los datos en dos distintas clases de manera óptima [36]

2.4.2. Random forest

Un árbol predictor es un modelo predictivo de aprendizaje supervisado. Dada su simplicidad e inteligibilidad, es una de los algoritmos de aprendizaje de máquina más populares [41]. Entre sus muchos usos, puede ser utilizado para tareas de clasificación. Para lograr esto, primeramente, se entrena al árbol predictor con datos etiquetados para que establezca los parámetros de decisión. Dado un conjunto etiquetado de observaciones con ciertas características, un árbol predictor lo divide en subconjuntos disjuntos con base en una de las características que poseen las observaciones. Este proceso se repite hasta que cada elemento pertenece a la misma categoría que el resto de los elementos en el subconjunto o ya no haya observaciones por las cuales se pueda dividir un subconjunto, que mejoren la precisión. Las observaciones elegidas para dividir son seleccionadas en cada etapa para maximizar la ganancia de información [33].

Un *Random Forest* se compone de un conjunto de árboles predictores. Los cuales fueron construidos con subconjuntos aleatoriamente seleccionados de los datos de entrenamiento. *Random Forest* junta y promedia las decisiones de cada árbol para la predicción final. De este modo, se evitan problemas de sobre-ajuste con los datos de entrenamiento y se mejora la exactitud de la predicción [26].

2.4.3. K vecinos más próximos (KNN)

El algoritmo de los K vecinos más próximos o KNN, por sus siglas en inglés (*k-nearest neighbors*, es un método de clasificación supervisada. El concepto principal de este método es: Dado un conjunto etiquetado de puntos en un plano (que representan un conjunto de observaciones) y un nuevo punto a etiquetar. Predecir la etiqueta del nuevo punto usando la etiquetas de los $\boldsymbol{k}$ vecinos más cercanos al nuevo punto [26]. Se puede usar distancia Euclidiana para determinar los $\boldsymbol{k}$ vecinos más cercanos u otro tipo de medida. A pesar de la simplicidad de este método, es bastante efectivo para tareas complejas de clasificación como identificación de dígitos escritos a mano [26]. En la Figura 2.3 se puede observar un ejemplo de los $\boldsymbol{k}$ vecinos más cercanos (en distancia euclidiana) con $\boldsymbol{k = 3}$ y con $\boldsymbol{k = 7}$. En caso que se decidiera la clase de un punto, considerando únicamente a cuál clase pertenecen la mayoría de sus vecinos más cercanos, el nuevo punto pertenece a la clase A si $\boldsymbol{k = 3}$; por otro lado, si $\boldsymbol{k = 7}$, el punto pertenece a la clase B.

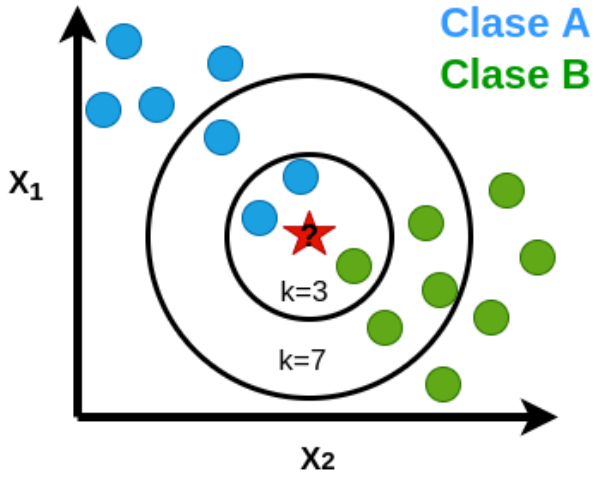

Figura 2.3: Ejemplo de KNN donde se busca clasificar a la estrella con base en sus vecinos más cercanos.

Capítulo 3

Trabajo Relacionado

A nuestro saber, este es el primer trabajo que usa temas recurrentes para la predicción de características en el ámbito cinematográfico. Si bien los datos contenidos en la página TV Tropes ya han sido explorados y reconocidos como una fuente de datos de utilidad [14, 15], el uso de los temas recurrentes en el aprendizaje de máquina no había sido investigado. Lo más cercano en esta área es un trabajo reciente donde el enfoque está en la predicción de temas recurrentes [10]. Por lo tanto, creemos que este es el primer trabajo que usa los temas recurrentes para la predicción del género cinematográfico y la clasificación por edades.

3.1. Género cinematográfico y clasificación por edad como características

La predicción de la clasificación por edades y del género cinematográfico, son tareas que ya han sido exploradas anteriormente [32, 34]. Esto se debe a que estas características a su vez sirven para predecir otros datos de vital importancia en la industria. Por ejemplo, la clasificación por edades ha sido utilizada para predecir las ganancias de una película [18]. De igual forma, el género cinematográfico es un recurso utilizado en varias tareas, como la predicción de calificaciones de usuarios [20]. De hecho, la predicción de calificaciones de usuarios es un tema altamente explorado que muestra el tipo de datos comúnmente

utilizados para predicciones en el ámbito cinematográfico. Por ejemplo, se han usado datos como: calificaciones de otros usuarios, el género cinematográfico, el año en que se estrenó una película o los directores, escritores y actores involucrados en su realización [16, 12, 4]. Otras tareas exploradas que usan datos similares son la predicción de éxito [7], ganancias esperadas [2] o que tan bien recibida será una película [31], este último también utiliza los diálogos.

3.2. Clasificadores basados en diálogos

En esta tesis, la predicción basada en diálogos es usada como base de comparación por inspiración de otros trabajos previos. Por ejemplo, el uso de los diálogos para la predicción de la clasificación por edades (asignada por la MPAA) ya había sido realizado anteriormente en otro artículo [32]. Un trabajo similar es la predicción de la violencia presentada en una película, haciendo uso del guion cinematográfico [22]. Por otro lado, predicciones del género cinematográfico habían sido realizadas usando la sinopsis o resumen de una película [5, 17]. Lo cual intuimos que podía ser remplazado por los diálogos de películas, ya que, en cierta medida, los diálogos proporcionan una gran parte de la trama y dan una idea del tono usado en la película. Si bien los trabajos previos que realizan predicciones en estos campos usan algoritmos de predicción mucho más avanzados, ultimadamente muestran que datos útiles para las predicciones pueden ser extraídos de los diálogos de películas o fuentes similares de información.

3.3. Técnica de prototipado

Finalmente, la técnica utilizada para agrupar temas recurrentes fue inspirada en un artículo para detectar depresión [3]. En dicho trabajo, se agrupan palabras asociadas a emociones usando FastText [9] y propagación por afinidad [37]. En la Figura 3.1 se logra apreciar un diagrama del proceso realizado para agrupar palabras asociadas a emociones y en la Figura 3.2 se pueden apreciar algunos ejemplos de estas agrupaciones. Aunque en este trabajo el objetivo de este proceso era crear subdivisiones de emociones más generales,

yendo de lo general a lo particular; la misma idea puede utilizarse para ir de lo particular a lo general, agrupando conceptos específicos en uno más global. Por ello, de manera similar, nosotros usamos Word2Vec y K-means para agrupar los temas recurrentes y crear un prototipo de supra-temas recurrentes, por así decirlo, como se verá posteriormente.

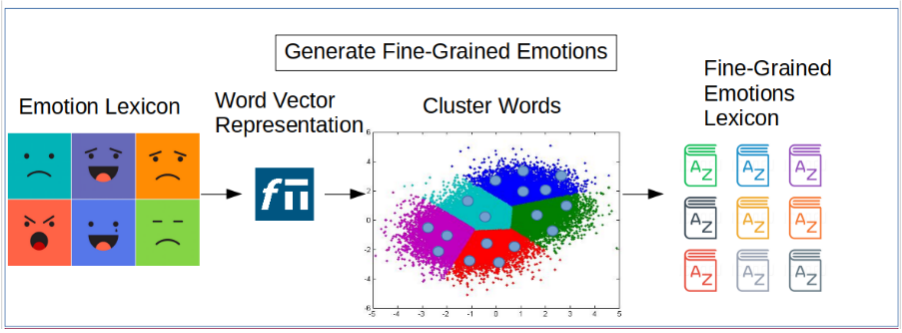

Figura 3.1: Diagrama que muestra el proceso usado para agrupar vocabulario de emociones generales a un conjunto más específico [3].

Anger			**Joy**		
anger1	***anger2***	***anger3***	***joy1***	***joy2***	***joy3***
abomination	growl	battle	accomplish	bounty	charity
fiend	growling	combat	achieve	cash	foundation
inhuman	thundering	fight	gain	money	trust
abominable	snarl	battler	reach	reward	humanitarian
unholy	snort	fists	goal	wealth	charitable
Surprise			**Disgust**		
surprise1	***surprise2***	***surprise3***	***disgust1***	***disgust2***	***disgust3***
accident	art	magician	accusation	criminal	cholera
crash	museum	wizard	suspicion	homicide	epidemic
disaster	artwork	magician	complaint	delinquency	malaria
incident	gallery	illusionist	accuse	crime	aids
collision	visual	sorcerer	slander	enforcement	polio

Figura 3.2: Ejemplo de grupos de sub-emociones generados por el proceso realizado [3].

Capítulo 4

Metodología Propuesta

En este capítulo explicamos la metodología seguida para la realización de esta tesis. En la Sección 4.1, detallamos la primera contribución; el corpus utilizado en este trabajo y los pasos seguidos para su construcción. Una vez establecido el conjunto de datos con el que se trabaja, en la Sección 4.2 mencionamos una deficiencia que este presenta y establecemos una estrategia para intentar aprovechar al máximo los temas recurrentes. En la Sección 4.3, explicamos el proceso realizado para la creación de un vector que representa la información de temas recurrentes de una película relacionados con los géneros cinematográficos. Posteriormente, en la Sección 4.4, describimos el método desarrollado para crear un vector que represente la información de los temas recurrentes asociados a la clasificación por edades y las posibles variantes que se podrían llevar a cabo. Finalmente, en la Sección 4.5, explicamos el proceso realizado tanto en las predicciones base, las cuales solo usan la información de diálogos, como las predicciones efectuadas para medir la utilidad de los temas recurrentes en las tareas de clasificación elegidas.

4.1. Diseño y Construcción del Corpus

Para llevar a cabo el presente trabajo, usamos dos principales fuentes de información. La primera son los datos recopilados para la realización del artículo *Age Suitability Rating: Predicting the MPAA Rating Based on Movie Dialogues* [32]. La segunda, el compendio de datos recopilados de la *wiki* TV Tropes, datos que después de ser extraídos les realizamos

una estructuración para poderlos usar adecuadamente. De los datos obtenidos dividimos el conjunto de datos en dos partes. El 70 % de los datos lo usamos para el análisis de datos y el entrenamiento de clasificadores. Y el 30 % restante lo utilizamos para evaluar los resultados obtenidos en la fase de experimentación.

4.1.1. Descripción de metadatos y diálogos

En el artículo *Age Suitability Rating* [32] se usa la información de un total de 5,562 películas. La información con la que se cuenta son los diálogos de estas películas y sus metadatos. Entre los metadatos de estas películas se encuentra información como: año en que la película fue estrenada, director de la película, géneros cinematográficos a los que pertenece (de acuerdo con IMDb) y la clasificación de la película. La clasificación de la película fue asignada por *The Motion Picture Association of America* (MPAA), la cual busca proporcionar a los negocios y a los padres una guía para decidir si una película es apta o no para un menor de edad.

Es importante mencionar dos aspectos de los metadatos del conjunto de datos. La primera, los géneros cinematográficos no son mutuamente excluyentes, en general una película tiene múltiples géneros. Y la segunda, las categorías presentes en la clasificación de las películas según la MPAA son G, PG, PG-13, R Y NC-17; sin embargo, la categoría NC-17 solo cuenta con 9 películas, por lo cual no son tomadas en cuenta. También es importante resaltar la disparidad de los datos, ya que más de la mitad de las películas tienen la clasificación R.

4.1.2. Extracción de datos Temas Recurrentes

Para poder hacer uso de los temas recurrentes en cada película, lo primero que necesitamos es poder acceder a estos datos de una forma sencilla y confiable. Para ello creamos una base de datos que contiene la información relevante de TV Tropes para este trabajo. Para la recolección de datos tomamos 3 tipos de artículos: Artículos de películas, que cuentan con una descripción y análisis de la película, además de un lista de temas recurrentes que aparecen en ella (como se muestra en las Imágenes 1.1 y 1.2). Artículos sobre temas

recurrentes, que cuentan con una descripción del tema y los elementos que en general lo conforman, así como una lista de obras en los que se usa este tema recurrente (según los usuarios de TV Tropes). Y artículos que enlistan temas recurrentes pertenecientes a una categoría; en particular, se usó el listado de temas recurrentes de cada género universal (Acción/Aventura, Drama, Comedia, Horror, Misterio, Romance y Thriller). La Figura 4.1 y la Figura 4.2 presentan cómo se muestra el listado de temas recurrentes en TV Tropes del género de Drama y Horror, respectivamente.

Figura 4.1: Imagen sacada de TV Tropes que muestra un listado de todas los temas recurrentes asociados al género de Drama[2].

Usando la herramienta de *TropeScraper* [15] se logró obtener un listado de películas y los temas recurrentes que se encuentran en ellas. La intersección entre el listado de películas presentes en TV Tropes y del conjunto de datos de diálogos y metadatos es de 2,536 películas, las cuales presentan 24,372 temas recurrentes diferentes. En promedio, cada pe-

[2] `https://tvtropes.org/pmwiki/pmwiki.php/Main/DramaTropes`, 11 de mayo del 2022.
[4] `https://tvtropes.org/pmwiki/pmwiki.php/Main/HorrorTropes`, 11 de mayo del 2022.

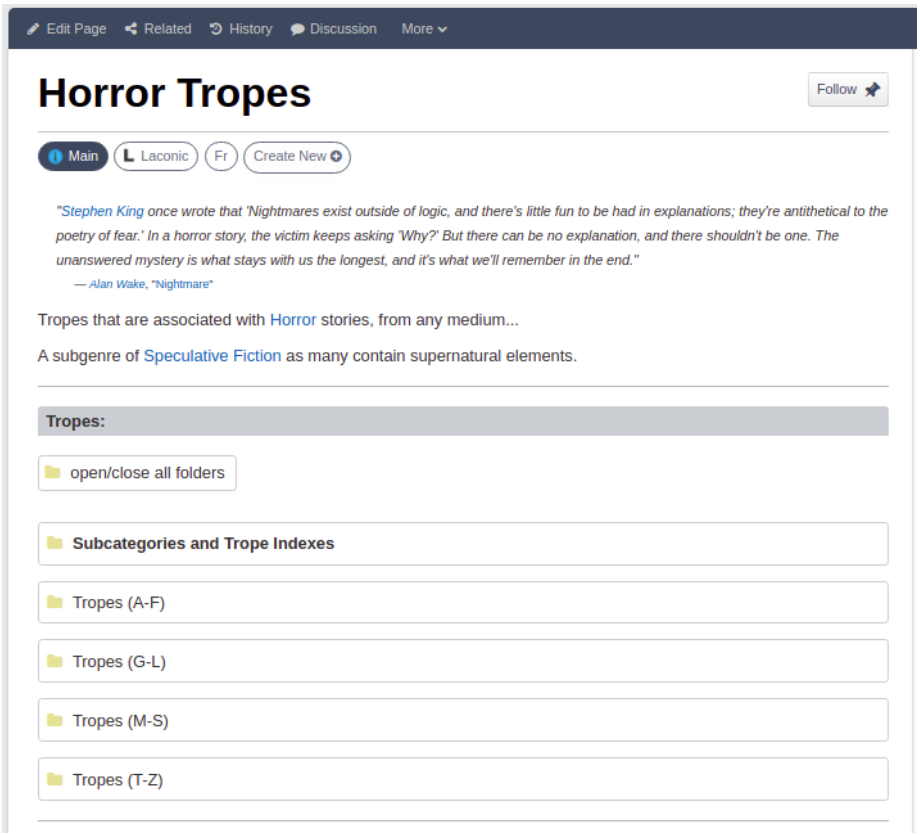

Figura 4.2: Imagen sacada de TV Tropes que muestra distintos folders donde están por orden alfabético los temas recurrentes asociados al género de Horror[4].

lícula presenta 107 temas recurrentes registrados; sin embargo, muchos temas recurrentes solo aparecen en una o dos películas. Por medio de técnicas de *web Scraping* [38] logramos extraer la descripción de los más de 24 mil temas recurrentes. De igual forma, usando los artículos de temas recurrentes por género, creamos un listado temas recurrentes comúnmente encontrados en los géneros universales. La lista de géneros presentados en la base de datos de diálogos y metadatos contempla una mayor variedad de géneros, pero los 7 géneros universales se encuentran dentro de estos. También un total de 2,271 películas presenta al menos uno de estos géneros, lo cual permite usar estos datos de forma adecuada.

Con la información mencionada previamente y creando identificadores únicos para cada uno de estos datos, logramos crear una base de datos que relaciona la información correctamente. La base de datos realizada nos permite acceder y explorar de manera sencilla los datos recolectados, tarea que antes presentaba un grado de dificultad para nada trivial y ha sido explorada en otros trabajos previos [14]. Creemos que la base de datos creada podría

llegar a ser de gran utilidad para futuros trabajos. La Figura 4.3 resume los 3 principales conjuntos de datos extraídos de TV Tropes utilizados en este trabajo. Los cuales son: La relación entre películas y temas recurrentes, los temas recurrentes asociados a un género y las descripciones de un tema recurrente.

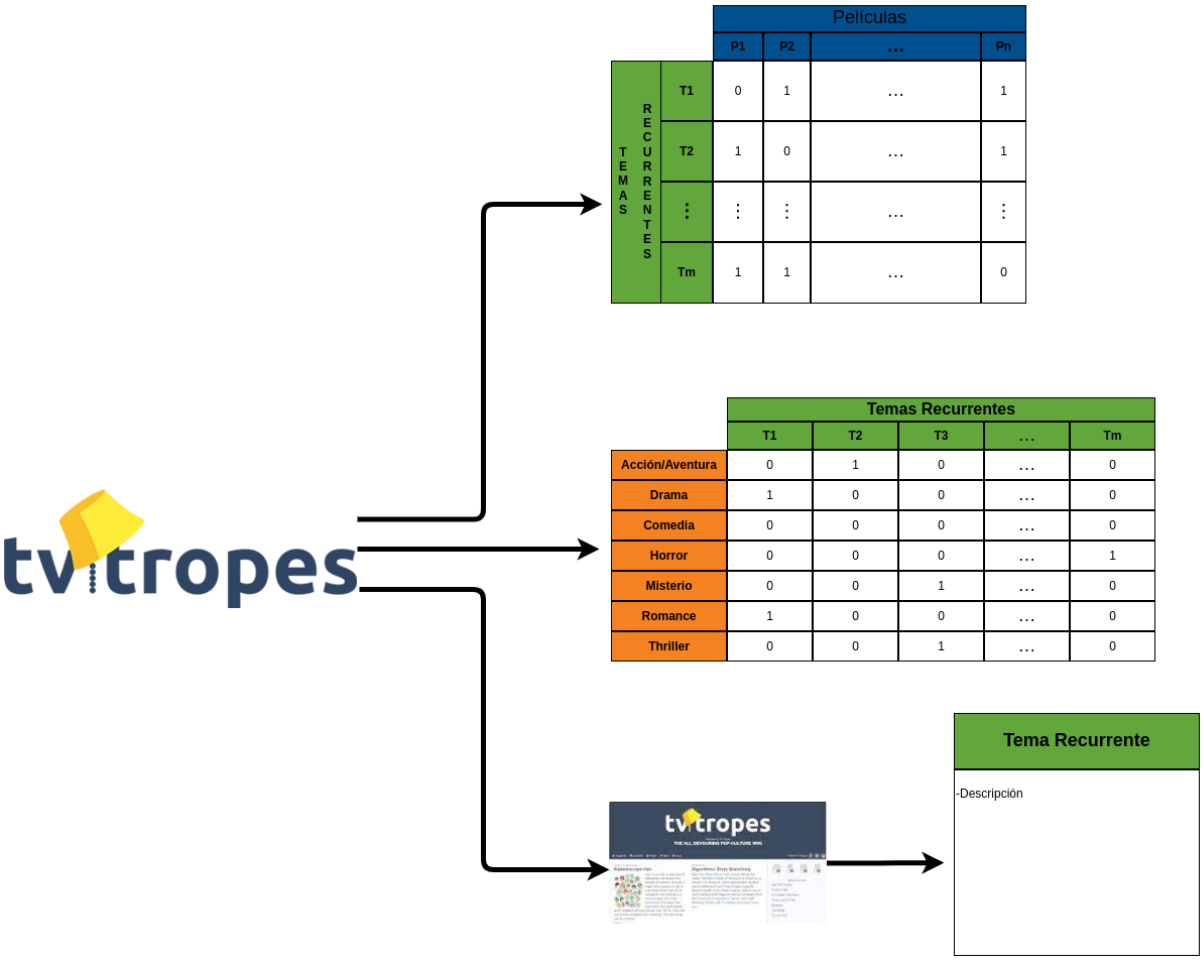

Figura 4.3: Datos extraídos de TV Tropes ya procesados. El primero es una matriz de películas y temas recurrentes, donde 1 indica la presencia del tema recurrente y 0 su ausencia. El segundo una matriz de temas recurrentes y géneros, donde 1 indica que están asociados y 0 que no lo están. Y finalmente, la descripción en texto de cada tema recurrente.

4.2. Estrategia de representación

Aunque la cantidad de datos en nuestro corpus es bastante cuantitativa, un problema es la frecuencia de cada tema recurrente. Como se mencionó con antelación, contamos

con 24,372 temas recurrentes que aparecen en 2,536 películas. En promedio, cada tema recurrente aparece en solamente 11 películas. Si bien se puede encontrar un tema recurrente específicamente ligado a una categoría de una clasificación hecha, es poco probable que el tema recurrente aparezca en la mayoría de las películas. Más aún, la frecuencia de los temas recurrentes presentada en nuestro corpus no siempre refleja la realidad.

Aunque el trabajo colaborativo hecho en TV Tropes de análisis y detección de temas recurrentes es bastante exhaustivo, al ser hecho por personas no es perfecto. Lo anterior se debe a la extensión de datos, ya que es casi imposible que el conjunto de personas trabajando en el análisis de cierta película conozcan a la perfección el total de temas recurrentes existentes y por ello no sean capaces de detectar la presencia de todos ellos. Lo mismo en el caso del grupo de personas trabajando en el artículo de un tema recurrente, es bastante difícil que conozcan sobre toda obra literaria que presenta el tema recurrente. Al tener presente lo anterior, se puede entender que los temas recurrentes de cada película en nuestro corpus son los temas recurrentes más claramente presentes en la película, más no los únicos.

Sin embargo, a pesar de las deficiencias anteriores, los temas recurrentes tienen la propiedad de ser agrupables y jerarquizables. De hecho, muchos temas recurrentes son casos particulares de un tema recurrente más general. Por ello, en este trabajo proponemos un método sencillo, pero efectivo, para proporcionar los datos relacionados con los temas recurrentes a un clasificador. Buscamos agrupar los temas recurrentes en conjuntos de temas similares. En este escenario, cada grupo se convierte en una especie de supra-tema recurrente o "meta" tema recurrente que agrupa a varios bajo una nueva etiqueta o ID. De esta manera, el conjunto de temas recurrentes tienen una frecuencia considerablemente mayor a la de los elementos que lo integran.

En el presente trabajo, denominamos **prototipo de tema recurrente** a un conjunto de temas recurrentes que engloban las situaciones particulares de cierta categoría de películas (ej. conjunto de temas recurrentes asociados a un género cinematográfico). Elegimos este nombre, puesto que la generalización y abstracción del conjunto de temas recurrentes crean un propuesta de tema recurrente que define la categoría, donde las instancias del prototipo son los mismos temas recurrentes del conjunto. Consideramos que identificar

prototipos de temas recurrentes será de gran ayuda para las tareas de clasificación. Posteriormente, abarcaremos a detalle métodos para generar prototipos de temas recurrentes y algunos de sus posibles usos.

4.3. Representación de temas recurrentes por género cinematográfico

El género cinematográfico sirve para la clasificación y comercialización de las películas. Los 7 géneros cinematográfico que manejamos para la tarea elegida son: Acción/Aventura, Drama, Comedia, Horror, Misterio, Romance y Thriller. Para la predicción del género cinematográfico se propone usar los temas recurrentes que están asociados a un género. Extrajimos esta información de TV Tropes y está basada en la teoría literaria. Un tema recurrente está asociado a un género cinematográfico si este aparece constantemente en obras (películas) pertenecientes al género o está directamente asociado por su contenido.

4.3.1. Validación de temas recurrentes por género cinematográfico

Primeramente, queremos validar el listado de temas recurrentes asociados a un respectivo género cinematográfico. Para ello, realizamos un análisis de datos y averiguamos el porcentaje de películas que presentan dichos temas recurrentes. Del total de 2,536 películas en nuestra base de datos, 2,271 de ellas presenta al menos una ocurrencia de uno de los 4,294 temas recurrente asociados a alguno de los 7 géneros cinematográficos básicos. La Figura 4.4 muestra la distribución de los temas recurrentes por género. Por estas razones, concluimos que existe cierta viabilidad en limitarnos a usar los temas recurrentes seleccionados.

De igual forma, también consideramos la posibilidad que, aunque teóricamente los temas recurrentes estén asociados a un género cinematográfico, en la práctica algunas películas pertenecientes a un género no presenten temas recurrentes asociadas a este. Este fenómeno podría atribuirse a que los géneros vayan cambiando y se adapten a la época. Por ejemplo, las películas de terror de los años 30, que generalmente presentan un monstruo,

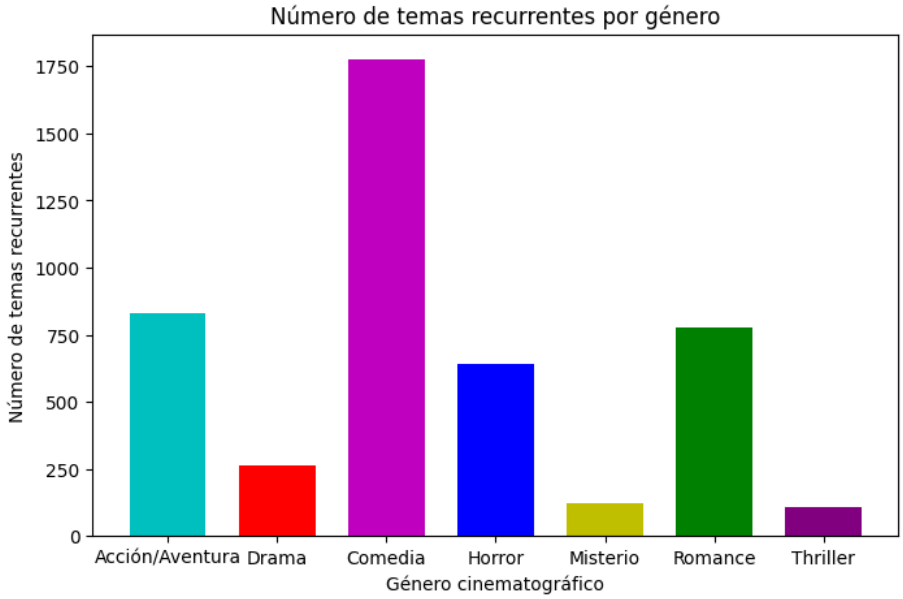

Figura 4.4: Histograma de temas recurrentes asociados a un género.

son sumamente diferentes a las de los 80, las cuales pertenecen principalmente al subgénero del *Slasher* [21]. Si estas diferencias llegan a ser lo suficientemente significativas, podría haber una disonancia entre que constituye un género específico. Por ello, en la Tabla 4.1 contabilizamos el número de películas de cada género y el porcentaje de ellas que presentan temas recurrentes del respectivo género. En esta, podemos observar que la mayoría de las películas perteneciente a un género presentan un tema recurrente asociando a dicho género. De hecho, a excepción del Drama y el Misterio, más del **85 %** de las películas presentan un tema recurrente del género al que pertenecen. Por lo cual, consideramos que los grupos de temas recurrentes extraídos son un recurso viable para la predicción del género cinematográfico

4.3.2. Prototipado y representación

El conjunto de temas recurrentes relacionados con un género cinematográfico forman el prototipo de tema recurrente de dicho género. Así que ya tenemos el prototipo de tema recurrente por género. La forma propuesta de usar los datos es haciendo un conteo por película del número de temas recurrentes de cada género, como se puede observar en la Figura 4.5. La forma más sencilla de obtener estos datos para todas las películas es

Género	Número de películas por Género	Porcentaje de películas con temas recurrentes del género
Acción/Aventura	742	92.99
Drama	1222	69.55
Comedia	901	94.56
Horror	400	94.50
Misterio	338	52.36
Romance	482	90.66
Thriller	860	86.97

Tabla 4.1: Número de películas de cada género y el porcentaje de ellas que presenta temas recurrentes del respectivo género.

multiplicando las dos matrices obtenidas con los datos de TV Tropes (4.3), como se muestra en la Figura 4.6. Lo cual nos permite representar a cada película como un vector que en cada entrada indica cuántos temas recurrentes de cada género tiene. La idea intuitiva es tener una representación que indique cuánto de cada género cinematográfico tiene cada película.

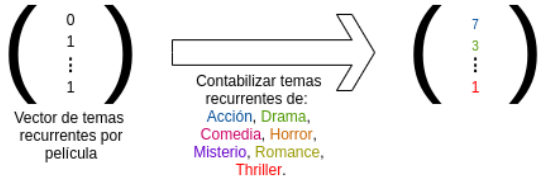

Figura 4.5: Dado un vector que indica los temas recurrentes presentes en una obra, podemos contabilizar el número de temas recurrentes presentes de cada género.

Figura 4.6: Multiplicación de la matriz de Género - Temas recurrentes por Temas recurrentes - Películas, obteniendo el número de temas recurrentes por género de cada película.

4.4. Representación de temas recurrentes por clasificación de edades

A diferencia de los géneros cinematográficos, no hay un compendio de temas recurrentes asociadas a cada una de las categorías designadas por la MPAA. Al igual que en muchos problemas de clasificación de películas, no se tienen definiciones específicas entre que temas recurrentes están asociados con cierta categorías. Lo que buscamos en esta sección es representar la información relevante de los temas recurrentes para la tarea de predicción de clasificación por edades. Creemos que el prototipado de temas recurrentes puede asistirnos en esta tarea, encontrando una buena representación, y adicionalmente ayudarnos a profundizar en el entendimiento de cada categoría.

4.4.1. Vector de temas recurrentes

Un primer acercamiento es usar un vector de 1's y 0's por película. Donde el 1 indica la presencia de un tema recurrente y el 0 la ausencia; como se observa en la Figura 4.7 y se mostró con antelación en la 4.3. Al vector resultante le aplicamos una normalización L1 para evitar sesgos en el número de temas recurrentes por película. Así obtenemos una representación fácil de computar y presenta la información sobre todos los temas recurrentes de una película; sin embargo, también tiene varios defectos. Un primer defecto es el tamaño del vector en comparación a la información provista. En promedio, cada película tiene 107 temas recurrentes, comparado con más 24 mil temas recurrentes en total. Y muchos de estos temas recurrentes no tienen relación alguna con la clasificación deseada, lo cual mete una gran cantidad de ruido. Un segundo problema son los temas recurrentes que no aparecen con suficiente frecuencia, los cuales podrían llevar al clasificador a un conclusión errónea por falta de información. Finalmente, tiene el defecto que muchos temas recurrentes que son muy similares no serán asociados directamente.

Figura 4.7: Matriz de Temas recurrentes - Películas. Cada columna sería la representación de temas recurrentes por película (después de ser normalizada).

4.4.2. Creación de prototipos generales

Para solucionar los problemas mencionados anteriormente, realizamos un proceso de filtrado y agrupación. Primero, tomamos únicamente los temas recurrentes arriba del cuantil 0.9 de frecuencia, para que todos los temas recurrentes que se usen no sean poco frecuentes (menos de 3 ocurrencias). Posteriormente, utilizamos la Prueba χ^2 de Pearson para seleccionar que temas recurrentes son buenas características para determinar la clasificación de la MPAA. Con el proceso de filtrado terminamos con 600 temas recurrentes para ser usados. La Figura 4.8 ilustra el proceso realizado.

Para poder asociar en la misma categoría los temas recurrentes similares hicimos un proceso de prototipado. Como se ilustra en la Figura 4.9, primero tomamos la definición de cada tema recurrente y la representamos como la suma ponderada de la representación de cada palabra en su definición, representada en Word2Vec y ponderada con la frecuencia inversa de documento (donde el documento es la descripción del tema recurrente en el compendio de descripciones de temas recurrentes). Una vez obtenida esta representación, hicimos un K-means para separar en 15 grupos diferentes los temas recurrentes, como se observa en la Figura 4.10. Cada uno de estos clusters es un prototipo de tema recurrente. Por ello, agrupamos los temas recurrentes en una Matriz de Prototipos y Temas recurrentes, este proceso se ilustra en la Figura 4.11. De esta manera, logramos conseguir los Prototipos y los temas recurrentes que lo integran.

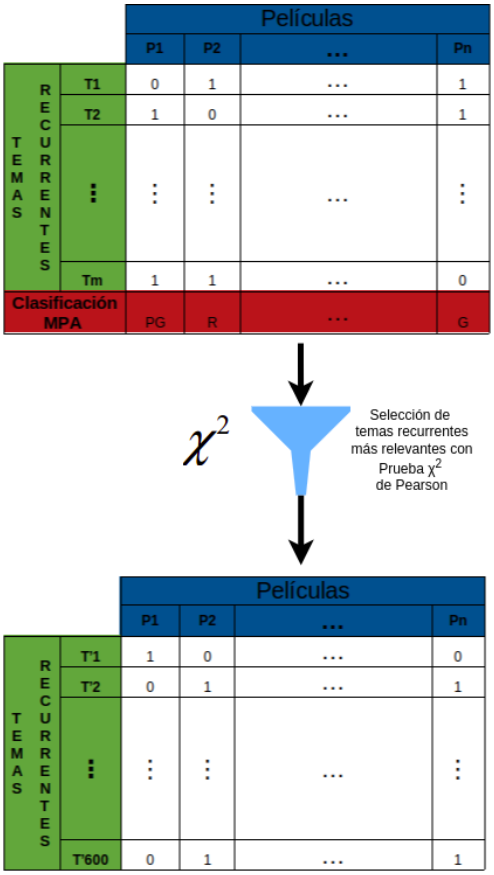

Figura 4.8: Se mete la clasificación MPAA de cada película para poder hacer uso de la Prueba $\boldsymbol{\chi^2}$ de Pearson y seleccionar los temas recurrentes más relevantes.

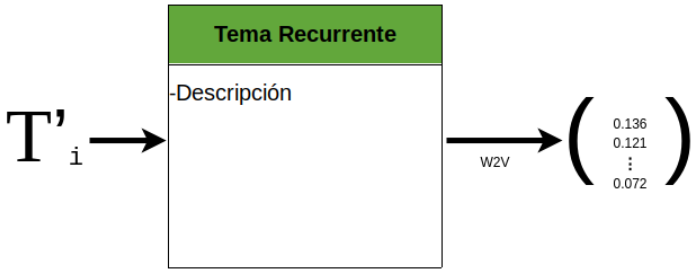

Figura 4.9: Para cada tema recurrente que fue seleccionado en 4.8 se toma su descripción y se representa en un vector. Este vector es la suma ponderada (por la frecuencia inversa) de la representación Word2Vec de cada palabra en su descripción.

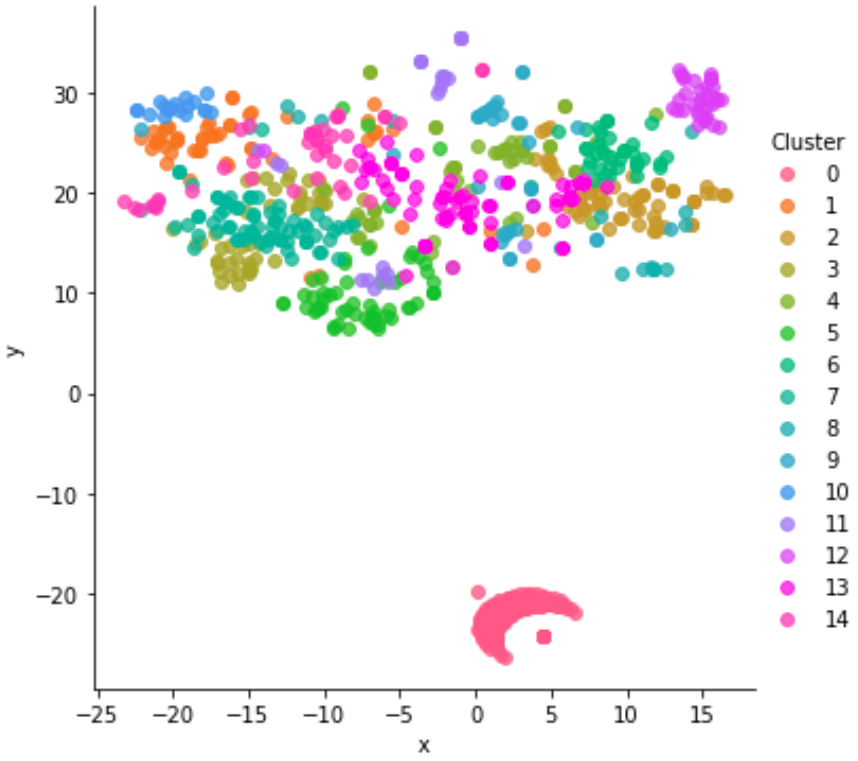

Figura 4.10: Imagen de los 15 clusters de temas recurrentes. La imagen muestra una representación de los datos hecha con la herramienta de TSNE de Sklearn al reducirse a dos dimensiones las coordenadas de los temas recurrentes.

4.4.3. Creación de prototipos por categoría

Si bien con los prototipos logramos agrupar a los temas recurrentes, estas agrupaciones no están asociadas a una clasificación específica, como el caso de los temas recurrentes por género cinematográfico. Lo cual podría ser de utilidad para la tarea de predicción de clasificación por edades y podría ser útil para entender más a fondo las clasificaciones. Para solucionar esto realizamos un paso intermedio después de la filtración, proceso descrito en la sección anterior y en la Figura 4.8. Una vez seleccionadas los temas recurrentes con los que vamos a trabajar, separamos estos en categorías, una por cada clasificación. En estas categorías solo aparecen los temas recurrentes que exclusivamente están en películas con esa clasificación y una categoría extra donde están los temas recurrentes que aparecen en películas con distintas clasificaciones. Es decir, si un tema recurrente aparece en solamente películas con clasificación R, entonces está en la categoría relacionada con la clasificación R, pero si aparece en una película con clasificación R y en otra con clasificación PG se

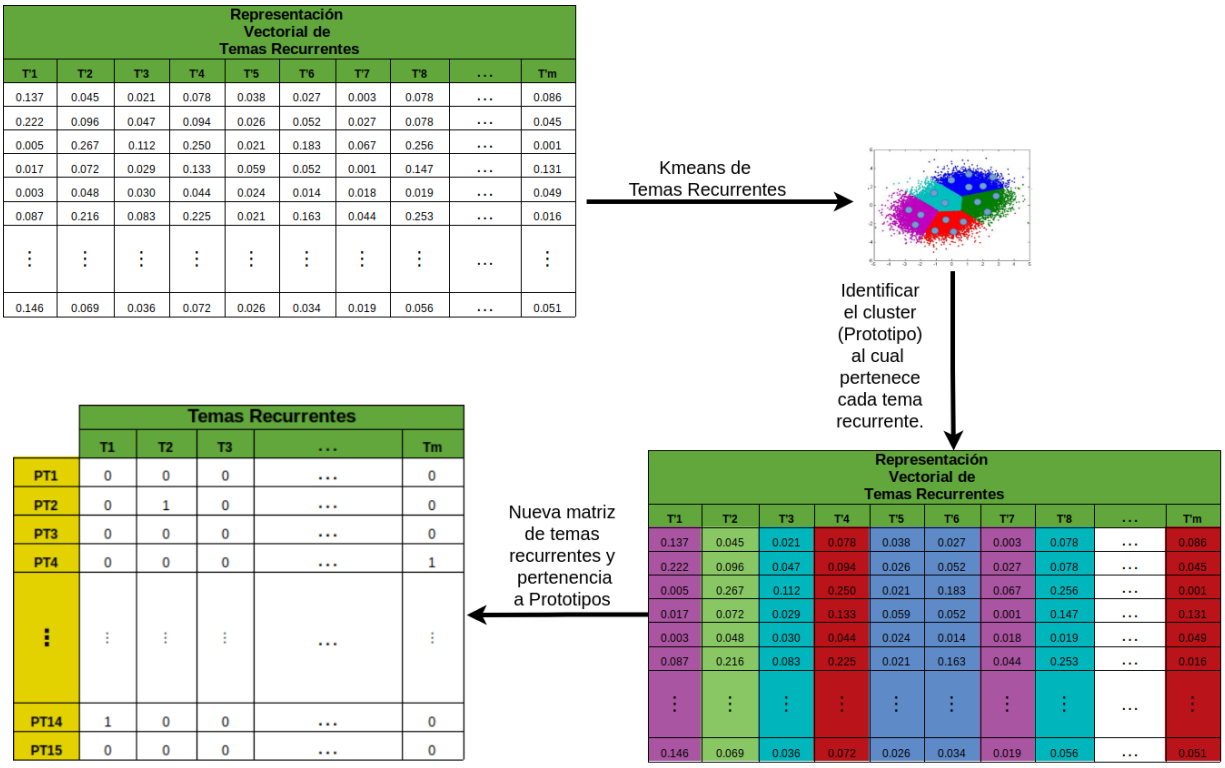

Figura 4.11: Utilizando la representación vectorial de los temas recurrentes obtenidos en 4.9, hacemos K-means para separar en 15 clusters, cada cluster será un prototipo de tema recurrente. Una vez identificados a que cluster pertenece cada Tema recurrente, creamos una matriz de Prototipos y Temas recurrentes donde 1 representa pertenencia y 0 su no pertenencia.

va a la categoría extra, la cual llamamos Mixta. Creamos la sección Mixta, ya que estas categorías son mutuamente exclusivas, a diferencia de los géneros cinematográficos. La Figura 4.12 muestra la distribución de temas recurrentes por estas nuevas categorías.

Posteriormente a esta separación, llevamos a cabo exactamente el mismo proceso de prototipado, mencionado con previamente en la sección anterior, para cada una de las categorías asociadas a una clasificación. Las Figuras 4.16, 4.15, 4.14, 4.13 y 4.17 muestran los clusters formados por los temas recurrentes de las categorías: G, PG, PG-13, R y Mixta. Después de obtener los prototipos por categoría, juntamos todos los prototipos en una sola matriz, como se muestra en la Figura 4.18. De tal manera que obtenemos los prototipos y los temas recurrentes que los integran, pero cada uno de estos prototipos está relacionado con una clasificación (G, PG, PG-13 y R). Lo cual es de ayuda para comprender que constituyen estas clasificaciones.

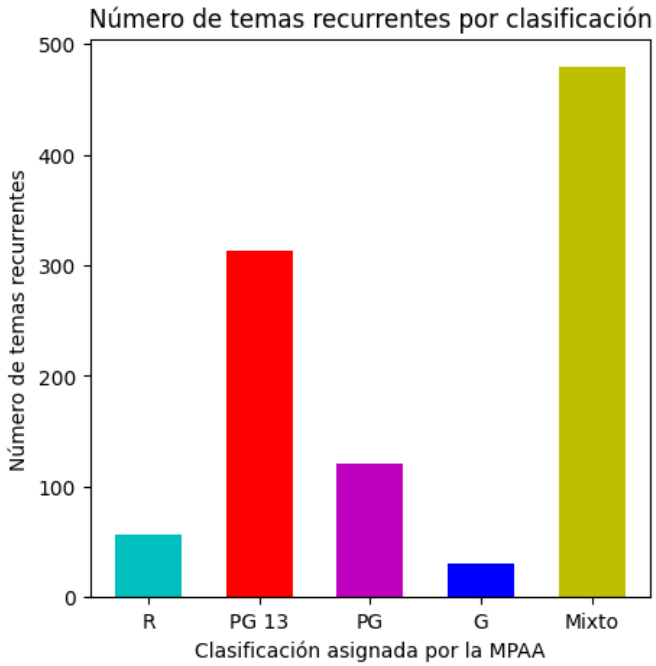

Figura 4.12: Histograma que indica el número de temas recurrentes por categoría.

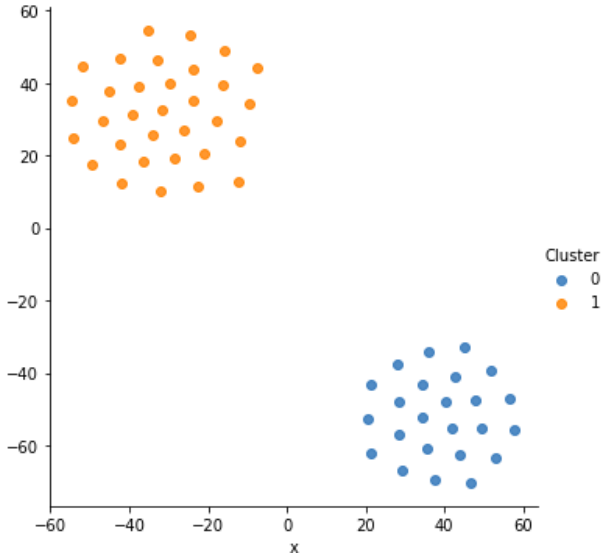

Figura 4.13: Imagen de la representación de los datos hecha con TSNE. Muestra los 2 clusters generados de los temas recurrentes con clasificación R exclusivamente.

4.4.4. Representación vectorial de una película

Una vez que ya obtuvimos un prototipado efectivo, podemos usar la misma técnica que con los géneros cinematográficos: Contabilizar el número de temas recurrentes por prototipo en cada película. Y representar una película con un vector que en cada entrada

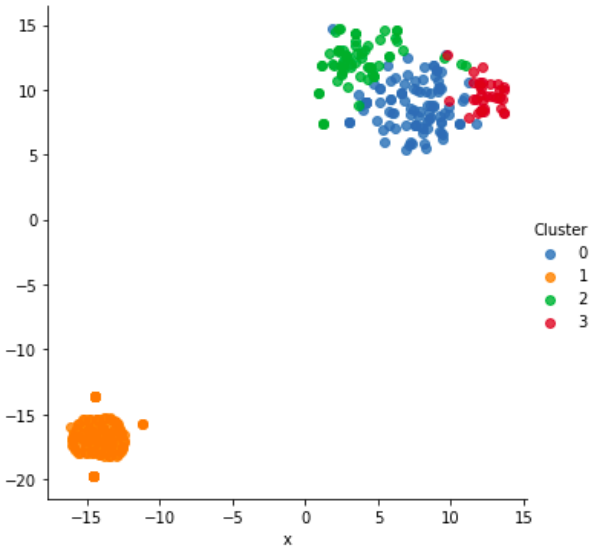

Figura 4.14: Imagen de la representación de los datos hecha con TSNE. Muestra los 4 clusters generados de los temas recurrentes con clasificación PG-13.

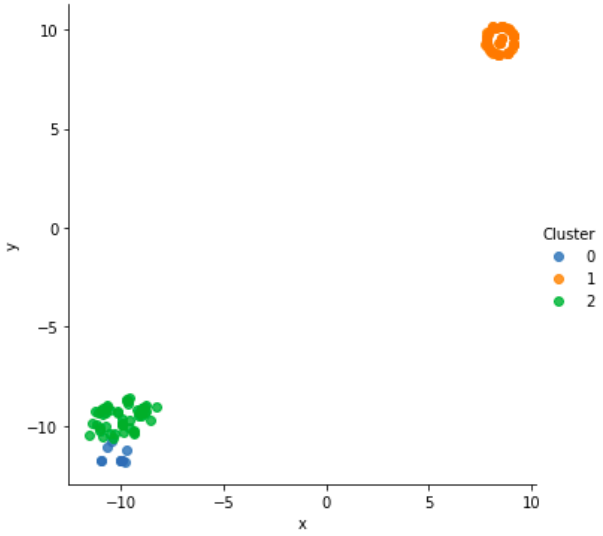

Figura 4.15: Imagen de la representación de los datos hecha con TSNE. Muestra los 3 clusters generados de los temas recurrentes con clasificación PG.

indica con cuántos temas recurrentes de un prototipo cuenta. La forma más fácil de realizar esto es multiplicando la matriz obtenida en 4.11 con la matriz de temas recurrentes y películas, como se muestra en la Figura 4.19. La representación final de cada película no es más que la normalización del vector que en cada entrada tiene cuantos temas recurrentes de cada prototipo presenta la película; es decir, la normalización de las columnas de la matriz resultante en la Figura 4.19. Todo el proceso de inicio a fin se indica en la Figura

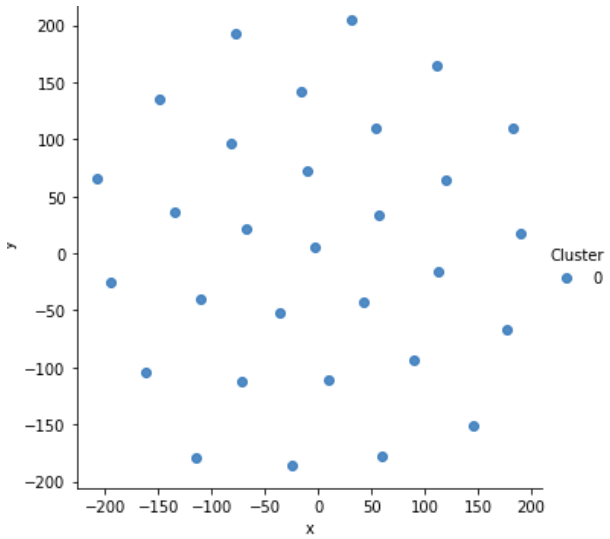

Figura 4.16: Imagen de la representación de los datos hecha con TSNE. Muestra el único cluster formado con los temas recurrentes con clasificación G. Solo se generó un grupo, ya que la cantidad de temas recurrentes pertenecientes a esta categoría es menor que el resto.

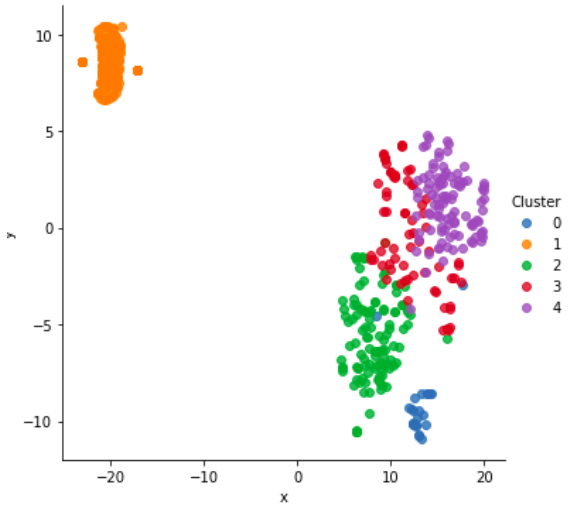

Figura 4.17: Imagen de la representación de los datos hecha con TSNE. Muestra los 5 clusters generados de los temas recurrentes que no pertenecen a una sola clasificación.

4.20.

Similarmente, para obtener la representación de Prototipos por Categoría solo tenemos que multiplicar la matriz obtenida en 4.18 por la matriz de temas recurrentes y películas, como se muestra en la Figura 4.21. Al hacer esto obtenemos cuántos temas recurrentes de cada prototipo por categoría aparecen en cada película.

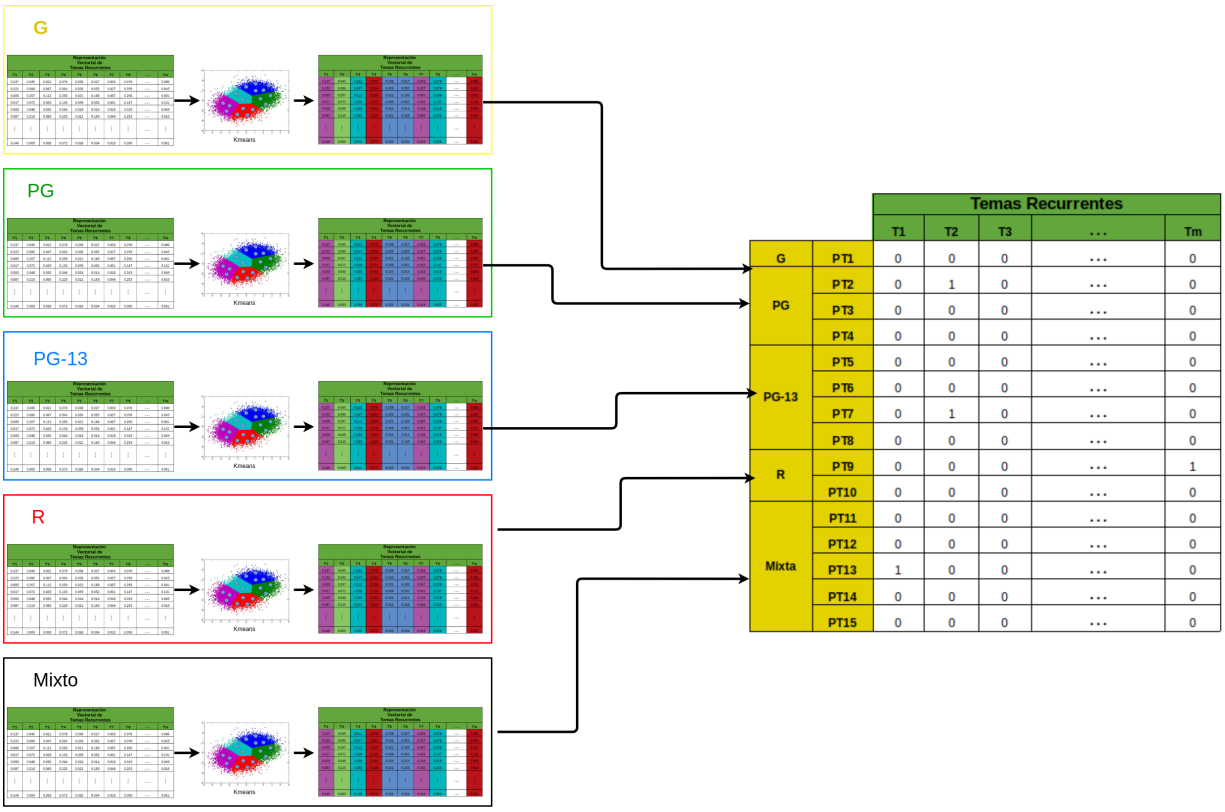

Figura 4.18: Después de identificar los distintos prototipos en cada categoría, se juntan todos en una matriz de Prototipos y temas recurrentes, donde 1 indica pertenencia y 0 no pertenencia.

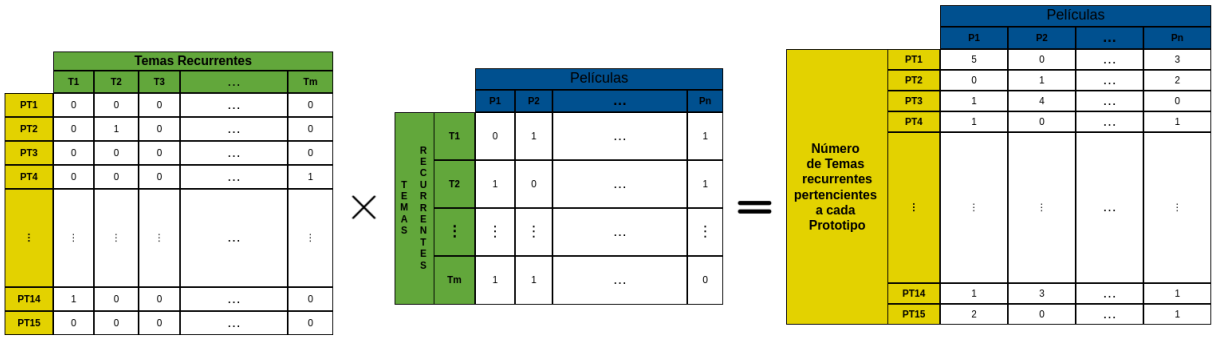

Figura 4.19: Multiplicación de la matriz Prototipos - Temas recurrentes por Temas recurrentes - Película, obteniendo la matriz Prototipos - Películas que en cada entrada indica cuantos temas recurrentes de dicho prototipo contiene la película.

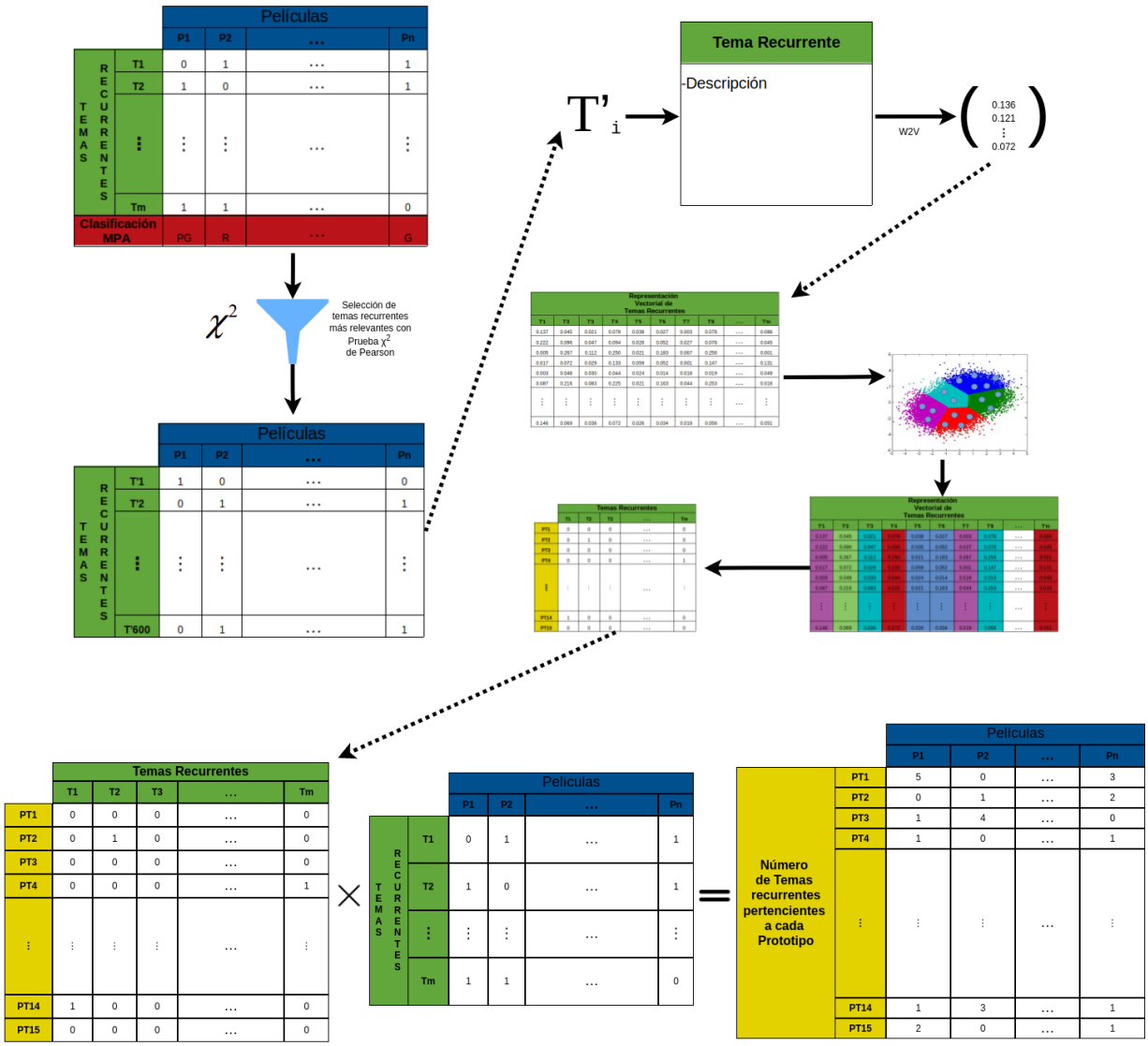

Figura 4.20: Proceso realizado para la creación de la representación de películas por prototipos. En esta se muestra la relación entre las Figuras 4.8, 4.9, 4.11 y 4.19

4.5. Estrategia de clasificación

Para tener una base sobre la cual evaluar la efectividad de los temas recurrentes en las tareas de clasificación, primeramente usamos los diálogos de películas para predecir el género cinematográfico y la clasificación por edades. Realizamos 9 predicciones por tarea, donde usamos 3 algoritmos clasificadores: SVM, Random Forest y KNN, y 3 representaciones distintas de los diálogos, cada una utiliza una técnica clásicas distinta de representación: Bolsa de palabras, Tf-idf y Word2Vec (usamos el promedio de la suma de

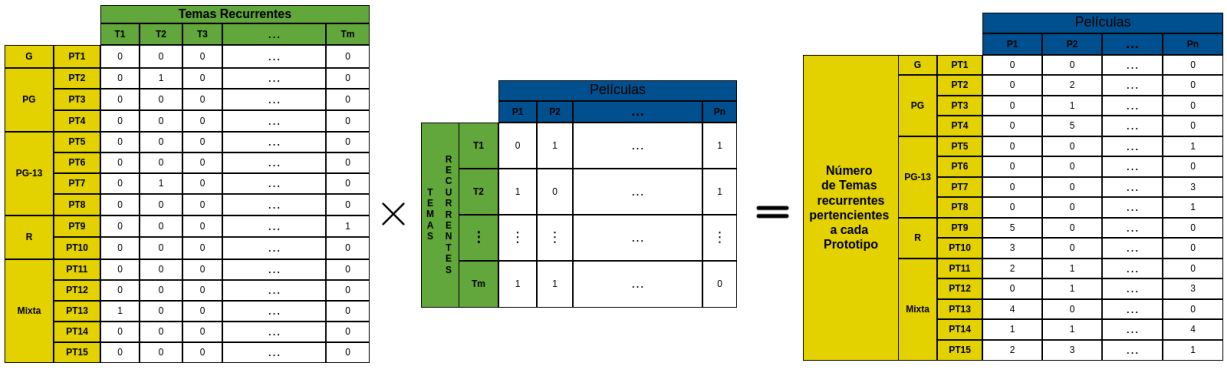

Figura 4.21: Multiplicación de la matriz Prototipos por Categoría - Temas recurrentes por Temas recurrentes - Película, obteniendo la matriz Prototipos por Categoría - Películas que en cada entrada indica cuantos temas recurrentes de dicho prototipo contiene la película.

las representaciones Word2Vec de los diálogos). Todo esto con el fin de tener un panorama más grande y poder generalizar los resultados. De tal manera que podemos comparar qué tanto puede beneficiar añadir la información de temas recurrentes a las predicciones hechas con los diálogos en general y no solo aludir a un caso particular. E igualmente, tener un abanico más grande con el cual comparar los resultados de las predicciones que solamente utilizan temas recurrentes.

Finalmente, utilizando las respectivas representaciones de los temas recurrentes relevantes para cada tarea, realizamos nuevas clasificaciones usando los mismos 3 algoritmos: SVM, Random Forest y KNN. Comparando las predicciones realizadas solo con diálogos, solo con temas recurrentes y haciendo uso de ambas. Para hacer uso de ambas unimos en un solo vector la información de ambos vectores, poniendo en las primeras entradas lo de un vector y en las últimas entradas lo del otro, como se muestra en la figura 4.22, luego realizamos una normalización L2. Esto con el fin de observar cómo la información provista por los temas recurrentes se desempeña en comparación a otra y si agregar esta información a la información de los diálogos beneficia en el resultado final de la clasificación o simplemente no provee utilidad alguna.

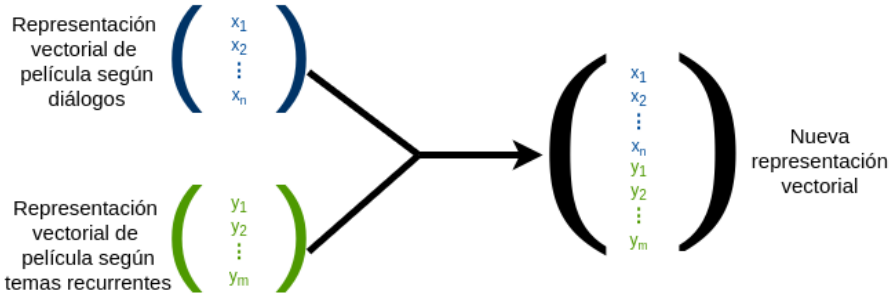

Figura 4.22: Para juntar los dos vectores de información basta con montar sus normalizaciones una encima de la otra. Obteniendo un vector con varias entradas que posteriormente serán analizados por los clasificadores.

Capítulo 5

Experimentos y Resultados

En este capítulo presentamos una evaluación de los métodos propuestos a través de varios experimentos realizados. En la Sección 5.1 presentamos los experimentos relacionados con la predicción de los 7 géneros cinematográfico básicos. Por otro lado, en la Sección 5.2 presentamos los experimentos relacionados con la predicción de la clasificación por edades, asignada por la MPAA. En ambas secciones presentamos las evaluaciones de los experimentos y concluimos que ambas tareas se benefician del uso temas recurrentes.

5.1. Género cinematográfico

El objetivo de los experimentos realizados en esta sección es averiguar si el uso de temas recurrentes puede ayudar en la predicción del género cinematográfico. Como base de comparación para el análisis, realizamos métodos tradicionales de clasificación usando diálogos de películas. Como se especifica en el capítulo anterior, realizamos 3 representaciones para los diálogos usando: Bag of Words, Tf-idf y Word2Vec. Igualmente, usamos 3 algoritmos de clasificación distintos: SVM, Random Forest y KNN. Primeramente, evaluamos la predicción de cada uno de los 3 algoritmos clasificadores con cada una de las 3 representaciones de los diálogos. Lo anterior se hace con el objetivo de obtener un panorama más amplio del efecto que tiene, en la predicción, agregar los temas recurrentes.

Para obtener la representación de los temas recurrentes por género, contabilizamos en cada película el número de temas recurrentes de cada género. Realizamos 3 predicciones,

usando los 3 algoritmos clasificadores y la representación de temas recurrentes por género. Después, a cada una de las 3 representaciones de diálogos le agregamos la información de la representación de temas recurrentes por género, como se especifica en el capítulo anterior. De tal manera obtenemos un vector que representa en sus primeras entradas la información de los temas recurrentes y en el resto la información de los diálogos. Usando estos nuevos vectores realizamos nuevas predicciones con cada uno de los 3 algoritmos clasificadores.

5.1.1. Evaluación

Clasificador	Evaluación	Temas Recurrentes por Género	Diálogos Representación Bag of Words	Diálogos Representación Tf-idf	Diálogos Representación Word2Vec
SVM	Exactitud	57.88	80.08	**82.37**	80.37
	F1 Score	40.19	69.45	**71.05**	69.75
Random Forest	Exactitud	58.14	80.94	80.80	**82.95**
	F1 Score	36.87	57.77	55.62	**68.26**
KNN	Exactitud	51.92	78.79	73.49	**80.65**
	F1 Score	27.37	55.95	23.23	**65.11**

Tabla 5.1: Evaluación de las predicciones realizadas con temas recurrentes por género cinematográfico y de las predicciones realizadas con las distintas representaciones de los diálogos. Se presenta en negritas los mejores resultados por clasificador.

En la Tabla 5.1 presentamos la evaluación de las predicciones realizadas con la información de temas recurrentes y las realizadas con la información de los diálogos. Al comparar las evaluaciones, podemos observar que las predicciones de temas recurrentes no tienen una evaluación tan buena como las basadas en diálogos. Sin embargo, la Tabla 5.2, la cual compara las predicciones hechas con diálogos contra las de diálogos y temas recurrentes, muestra que las mejores evaluaciones son las que usan ambos conjuntos de datos. En la tabla podemos observar que la predicción del género cinematográfico se beneficia de agregar la información provista por los temas recurrentes sin importar el algoritmo clasificador o la representación de diálogos usada. La mejora es poca en la mayoría de los casos de exactitud, pero más significativa en el *F1 Score* (ambas medidas efectuadas con micro-promedio). Podemos notar una mejora notable en los casos que cuentan con una calificación inicial muy baja. La exactitud mejora en promedio 1.31 puntos y el *F1 Score* 4.21 puntos. Esto

Representación Diálogos	Evaluación	Diálogos	Diálogos + Temas Recurrentes por Género
Clasificador SVM			
Bag of Words	Exactitud	80.08	80.50
	F1 Score	69.45	70.59
Tf-idf	Exactitud	82.37	**82.37**
	F1 Score	71.05	**71.33**
Word2Vec	Exactitud	80.37	81.37
	F1 Score	69.75	71.17
Clasificador Random Forest			
Bag of Words	Exactitud	80.94	81.36
	F1 Score	57.77	58.91
Tf-idf	Exactitud	80.80	81.37
	F1 Score	55.62	57.76
Word2Vec	Exactitud	82.95	**85.23**
	F1 Score	68.26	**73.68**
Clasificador KNN			
Bag of Words	Exactitud	78.79	79.64
	F1 Score	55.95	57.23
Tf-idf	Exactitud	73.49	79.2
	F1 Score	23.23	44.08
Word2Vec	Exactitud	80.65	**81.79**
	F1 Score	65.11	**71.53**

Tabla 5.2: Evaluación de las predicciones previa y posteriormente a la agregación de temas recurrentes por género. Se presenta en negritas los mejores resultados por clasificador.

confirma nuestra hipótesis, mostrando que los temas recurrentes proveen nueva información útil para la predicción del género cinematográfico. En la Tabla 5.3 presentamos en un solo lugar la evaluación de todas las predicciones del género cinematográfico realizadas.

Diferencias entre género asignado y género percibido

La percepción de un género cinematográfico no está definida de forma específica y no siempre es universal. Muchas veces una película es percibida de forma distinta a la que se tenía planeada por los mismos creadores o el equipo de marketing. Un ejemplo son las

Representación Diálogos	Clasificador	Evaluación	Diálogos	Diálogos + Temas Recurrentes por Género
Sin diálogos	SVM	Exactitud	-	57.88
		F1 Score		40.19
	Random Forest	Exactitud		58.14
		F1 Score		36.87
	KNN	Exactitud		51.92
		F1 Score		27.37
Bag of Words	SVM	Exactitud	80.08	80.50
		F1 Score	69.45	70.59
	Random Forest	Exactitud	80.94	81.36
		F1 Score	57.77	58.91
	KNN	Exactitud	78.79	79.64
		F1 Score	55.95	57.23
Tf-idf	SVM	Exactitud	82.37	82.37
		F1 Score	71.05	71.33
	Random Forest	Exactitud	80.80	81.37
		F1 Score	55.62	57.76
	KNN	Exactitud	73.49	79.2
		F1 Score	23.23	44.08
Word2Vec	SVM	Exactitud	80.37	81.37
		F1 Score	69.75	71.17
	Random Forest	Exactitud	82.95	85.23
		F1 Score	68.26	73.68
	KNN	Exactitud	80.65	81.79
		F1 Score	65.11	71.53

Tabla 5.3: Evaluación de las predicciones del género cinematográfico realizadas. Se presentan los resultados de las Tablas 5.1 y 5.2.

películas con subtramas románticas, en las cuales gran parte de la trama está relacionado con el género de Romance, pero no son consideradas como tal (ej. Shrek, Deadpool, Rocky, etc.). Esto se debe principalmente al público al que están dirigidos y cierta connotación negativa al género de Romance.

Otro ejemplo de películas con características de un género al cual formalmente no pertenecen, son las películas de estilo serie B. Serie B es un tipo de cine comercial de bajo presupuesto; donde, muchas veces, la película tiene un tono cómico fársico que no fue contemplado originalmente. Una instancia de esto es la película *Machete*, dirigida por

Robert Rodríguez. Las predicciones que usan temas recurrentes asignan la película *Machete* al género de comedia. Que si bien no es el género asignado por la producción, es el género percibido por la audiencia.

Si bien una gran cantidad de las predicciones realizadas con temas recurrentes no fueron las correctas, creemos que algunas de estas podrían ser consideradas acertadas en una revisión más profunda sobre el género cinematográfico. Esto se especifica no con el fin de justificarnos, sino para resaltar que es posible abrir una discusión sobre lo que constituye o debería constituir un género cinematográfico a partir de las discrepancias en las predicciones y pudiendo justificar la pertenencia o no pertenencia basada en los temas recurrentes que se presentan en una obra. Posibilidad que se vuelve mucho más compleja cuando se usan datos más ambiguos para las tareas de clasificación, como lo serían palabras o personas. En otras palabras, el uso de temas recurrentes hace mucho más inteligibles las predicciones realizadas y permite a las personas observar posibles anomalías en la clasificación original.

5.2. Clasificación por edades

El objetivo de los experimentos realizados en esta sección es averiguar si el uso de temas recurrentes puede ayudar en la predicción de la clasificación por edades, asignada por la MPAA. Como base de comparación para el análisis, realizamos métodos tradicionales de clasificación usando diálogos de películas. Como se especifica en el capítulo anterior, realizamos 3 representaciones para los diálogos usando: Bag of Words, Tf-idf y Word2Vec. Igualmente, usamos 3 algoritmos de clasificación distintos: SVM, Random Forest y KNN. Primeramente, evaluamos la predicción de cada uno de los 3 algoritmos clasificadores con cada una de las 3 representaciones de los diálogos.

Para obtener la representación de los temas recurrentes relevantes a la clasificación por edad, realizamos el proceso de prototipado por categorías, descrito en el capítulo anterior. Realizamos 3 predicciones, usando los 3 algoritmos clasificadores y la representación de prototipos por categoría. Después, a cada una de las 3 representaciones de diálogos le agregamos la información de prototipos por categoría, como se especifica en el capítulo

anterior. De tal manera obtenemos un vector que representa en sus primeras entradas la información de los temas recurrentes y en el resto la información de los diálogos. Usando estos nuevos vectores realizamos nuevas predicciones con cada uno de los 3 algoritmos clasificadores para ser comparadas con las predicciones que solo usaron la información provista por los diálogos.

De igual forma, comparamos las predicciones del prototipado por categorías con versiones más simples de la representación de temas recurrentes. Tales como el prototipado simple (sin la división por categorías) y el uso del vector completo de temas recurrentes. Para ambas versiones realizamos predicciones a secas y con las representaciones de diálogos. Esto con el fin de poder hacer comparaciones uno a uno con el prototipado por categorías y descubrir si este último presenta alguna ventaja en relación con sus contrapartes. De esta forma, tenemos un idea de la efectividad del sistema desarrollado para la selección de temas recurrentes por categoría, ilustrado en la tarea de predicción de clasificaciones por edades.

Para la predicción de clasificación por edades, en un principio, redefinimos la clasificación por edades, dividiendo únicamente en dos grupos, apto para menores y no apto para menores. Esto separando entre las películas con clasificación R y todas las demás. La razón de esta división se debe a que es la separación más importante y la que mejor divide los datos en grupos del mismo tamaño. Después de realizar los experimentos con únicamente estas dos clasificaciones, realizamos nuevamente los experimentos ahora con las posibles clasificaciones: G, PG, PG-13 y R. Aunque no haya tantas películas de una clasificación como de otra, consideramos que los resultados de las predicciones son lo suficientemente buenos como para ser presentados.

5.2.1. Evaluación de predicciones de clasificación Mayores-Menores de edad

En la Tabla 5.4 presentamos la evaluación de las predicciones realizadas con las distintas representaciones de temas recurrentes. En esta podemos ver que la que mejor se desempeña es la representación de prototipos por categoría. De hecho, en la Tabla 5.5 (la cual compara

Clasificador	F1 de Temas Recurrentes	F1 de Prototipos	F1 de Prototipos por Categoría
SVM	76.66	63.75	**77.12**
Random Forest	76.00	70.03	**78.01**
KNN	73.28	65.69	**75.36**

Tabla 5.4: Evaluación de predicciones de clasificación por edades (Menores y Mayores de edad) realizadas con las distintas representaciones de los temas recurrentes que se discutieron en la Sección anterior. Se presenta en negritas los mejores resultados por clasificador.

Clasificador	F1 de Prototipos por Categoría	F1 de Diálogos Representación Bag of Words	F1 de Diálogos Representación Tf-idf	F1 de Diálogos Representación Word2Vec
SVM	77.12	80.54	**83.22**	77.36
Random Forest	78.01	**86.91**	85.64	75.55
KNN	**75.36**	66.27	72.83	68.62

Tabla 5.5: Evaluación de predicciones de clasificación por edades (Menores y Mayores de edad) realizadas con los prototipos por categoría y realizadas con las distintas representaciones de los diálogos. Se presenta en negritas los mejores resultados por clasificador.

las predicciones hechas con los prototipos por categoría con las predicciones hechas con diálogos), podemos observar que los resultados obtenidos por los prototipos por categoría son comparables con los de los diálogos, incluso llegando a tener el mejor resultado en uno de los clasificadores. Sin embargo, más relevante son los resultado que se presentan en la Tabla 5.6, la cual presenta los resultados de las predicciones basadas en diálogos previa y posteriormente a agregar las distintas representaciones de temas recurrentes. Por un lado, podemos observar que agregar las otras representaciones de temas recurrentes puede tanto afectar como beneficiar el resultado de las predicciones. Por el otro lado, podemos observar que, salvo en un caso, agregar la información de los prototipos por categoría mejoró la predicción de la clasificación por edades. El aumento es bastante significativo, en algunos casos es mayor a 5 puntos y el único caso en que se vio afectado negativamente es menor a un punto. Así mismo, los mejores resultados por clasificador fueron logrados haciendo uso de los prototipos por categoría y alguna de las representaciones de los diálogos. Por todo ello, creemos que en general el uso adecuado de los temas recurrentes puede mejorar

Representación Diálogos	F1 Diálogos	F1 Diálogos + Temas Recurrentes	F1 Diálogos + Prototipos	F1 Diálogos + Prototipos por Categoría
	Clasificador SVM			
Bag of Words	80.54	80.54	81.63	81.63
Tf-idf	83.22	77.66	82.90	**87.22**
Word2Vec	77.36	77.25	76.92	82.59
	Clasificador Random Forest			
Bag of Words	86.91	84.89	84.69	86.69
Tf-idf	85.64	85.54	84.61	**87.74**
Word2Vec	75.55	74.65	78.78	82.22
	Clasificador KNN			
Bag of Words	66.27	66.76	66.95	66.37
Tf-idf	72.83	71.35	71.06	77.52
Word2Vec	68.62	68.45	66.94	**77.74**

Tabla 5.6: Evaluación de predicciones de clasificación por edades (Menores y Mayores de edad) previa y posteriormente a la agregación de las distintas representaciones de temas recurrentes. Se presenta en negritas los mejores resultados por clasificador.

las predicciones de las clasificaciones por edades. En la Tabla 5.7 presentamos en un solo lugar la evaluación de todas las predicciones de la clasificación por edades realizadas.

5.2.2. Evaluación de predicciones de clasificación por edades

En la Tabla 5.8 presentamos la evaluación de las predicciones realizadas con las distintas representaciones de temas recurrentes. De igual forma, en la Tabla 5.9 comparamos las predicciones hechas con los prototipos por categoría con las predicciones hechas con diálogos. Al igual que en la evaluación anterior, podemos ver que la predicción realizada usando únicamente la información de los prototipos por categoría es bastante aceptable (comparada con las predicciones que solo usan diálogos). De hecho, en ambas tablas podemos observar que la predicciones realizadas con los prototipos por categoría son las que mejor se desempeñan en 2 de los 3 clasificadores. La Tabla 5.10 muestra los resultados de las predicciones basadas en diálogos previa y posteriormente a agregar las distintas representaciones de temas recurrentes. Nuevamente, en la mayoría de los casos, agregar

Representación Diálogos	Clasificador	F1 Diálogos	F1 Diálogos + todos los Temas Recurrentes	F1 Diálogos + Prototipos	F1 Diálogos + Prototipos por Categoría
Sin diálogos	SVM	-	76.66	63.75	**77.12**
	RF		76.00	70.03	**78.01**
	KNN		73.28	65.69	**75.36**
Bag of Words	SVM	80.54	80.54	**81.63**	**81.63**
	RF	**86.91**	84.89	84.69	86.69
	KNN	66.27	66.76	**66.95**	66.37
Tf-idf	SVM	83.22	77.66	82.90	**87.22 ***
	RF	85.64	85.54	84.61	**87.74 ***
	KNN	72.83	71.35	71.06	**77.52**
Word2Vec	SVM	77.36	77.25	76.92	**82.59**
	RF	75.55	74.65	78.78	**82.22**
	KNN	68.62	68.45	66.94	**77.74 ***

Tabla 5.7: Evaluación de todas las predicciones de clasificación por edades (Menores y Mayores de edad) realizadas. Se presentan los resultados de las Tablas 5.4, 5.5 y 5.6. Se muestran en negritas la mejor calificación por fila y los elementos con * muestran el mejor resultado por clasificador.

Clasificador	F1 de Temas Recurrentes	F1 de Prototipos	F1 de Prototipos por Categoría
SVM	**66.76**	47.99	65.18
Random Forest	60.17	57.44	**67.04**
KNN	63.03	52.86	**63.61**

Tabla 5.8: Evaluación de predicciones de clasificación por edades (categorías: G, PG, PG-13 y R) realizadas con las distintas representaciones de los temas recurrentes que se discutieron en la Sección anterior. Se presenta en negritas los mejores resultados por clasificador.

la información de los prototipos por categoría mejoró la predicción de la clasificación por edades (a diferencia de las otras representaciones que tiene efectos tanto positivos como negativos). En promedio se presenta un aumento de más de 3 puntos. Los mejores resultados por clasificador fueron logrados haciendo uso de los prototipos por categoría y alguna de las representaciones de los diálogos (una distinta en cada clasificador). Por todo ello, concluimos que el uso adecuado de los temas recurrentes puede mejorar las clasificaciones por edades y que no se puede agregar la información de los temas recurrentes de forma tri-

Clasificador	F1 de Prototipos por Categoría	F1 de Diálogos Representación Bag of Words	F1 de Diálogos Representación Tf-idf	F1 de Diálogos Representación Word2Vec
SVM	65.18	70.77	**71.63**	60.17
Random Forest	**67.04**	65.04	65.04	61.03
KNN	**63.61**	53.15	58.88	58.16

Tabla 5.9: Evaluación de predicciones de clasificación por edades (categorías: G, PG, PG-13 y R) realizadas con los prototipos por categoría y realizadas con las distintas representaciones de los diálogos. Se presenta en negritas los mejores resultados por clasificador.

Representación Diálogos	**F1 Diálogos**	**F1 Diálogos + Temas Recurrentes**	**F1 Diálogos + Prototipos**	**F1 Diálogos + Prototipos por Categoría**
		Clasificador SVM		
Bag of Words	70.77	70.91	72.92	**73.21**
Tf-idf	71.63	68.91	54.58	70.04
Word2Vec	60.17	65.08	53.43	64.89
		Clasificador Random Forest		
Bag of Words	65.04	67.04	66.33	66.81
Tf-idf	65.04	64.89	65.18	**67.76**
Word2Vec	61.03	59.31	59.74	65.91
		Clasificador KNN		
Bag of Words	53.15	52.43	54.01	54.15
Tf-idf	58.88	62.75	53.15	64.18
Word2Vec	58.16	64.04	52.86	**64.46**

Tabla 5.10: Evaluación de predicciones de clasificación por edades (categorías: G, PG, PG-13 y R) previa y posteriormente a la agregación de las distintas representaciones de temas recurrentes. Se presenta en negritas los mejores resultados por clasificador.

vial, ya que agregar estos datos de manera poco procesada o con una agrupación no óptima pude dar resultados cruzados. En la Tabla 5.11 presentamos en un solo lugar la evaluación de todas las predicciones de la clasificación por edades (con las posibles clasificaciones: G, PG, PG-13 y R).

Representación Diálogos	Clasificador	F1 Diálogos	F1 Diálogos + Temas Recurrentes	F1 Diálogos + Prototipos	F1 Diálogos + Prototipos por Categoría
Sin diálogos	SVM	-	**66.76**	47.99	65.18
	RF		60.17	57.44	**67.04**
	KNN		63.03	52.86	**63.61**
Bag of Words	SVM	70.77	70.91	72.92	**73.21 ***
	RF	65.04	**67.04**	66.33	66.81
	KNN	53.15	52.43	54.01	**54.15**
Tf-idf	SVM	**71.63**	68.91	54.58	70.04
	RF	65.04	64.89	65.18	**67.76 ***
	KNN	58.88	62.75	53.15	**64.18**
Word2Vec	SVM	60.17	**65.08**	53.43	64.89
	RF	61.03	59.31	59.74	**65.91**
	KNN	58.16	64.04	52.86	**64.46 ***

Tabla 5.11: Evaluación de todas las predicciones de clasificación por edades (categorías: G, PG, PG-13 y R) realizadas. Se presentan los resultados de las Tablas 5.8, 5.9 y 5.10. Se muestran en negritas la mejor calificación por fila y los elementos con * muestran el mejor resultado por clasificador.

5.2.3. Consideraciones éticas

Como consecuencia de realizar el prototipado por categoría, obtuvimos conjuntos de temas recurrentes asociados a ciertas categorías. Los temas recurrentes asociados a una categoría proveen un mejor entendimiento de lo que constituye dicha categoría. De igual forma, dichos temas recurrentes proveen una herramienta de análisis sobre la clasificación de películas en cierta categoría. A continuación presentamos algunas observaciones sobre los temas recurrentes asociados a la clasificación R (no apta para menores).

Al realizar el proceso de prototipado para la categoría R se dividieron en dos grupos los temas recurrentes pertenecientes únicamente a dicha clasificación. Los dos grupos en sí dan una gran idea de que se considera no apto para menores. Uno de los prototipos principalmente se refiere a temas recurrentes sobre la violencia, algunos ejemplos son: Gorn, que se refiere a violencia extrema. Slasher movie, Slasher film y Serial killer; que se refieren a películas de terror sobre asesinos. Final guy y Final girl, que hacen referencia a los últimos sobrevivientes de un grupo de cada género. Y otros temas recurrentes más específicos sobre violencia. El otro prototipo hace, en su mayoría, referencia a temas recurrentes

sobre contenido sexual como: Male frontal nudity, A date with Rosie Palms, Death by Sex y otros. Sin embargo, uno de los temas recurrentes que también se encuentra en este conjunto es Queer Media[1], que hace referencia a obras que presentan un enfoque positivo a miembros o relaciones de la comunidad LGBT [1]. Esto se refiere a películas sobre la vida y experiencias de la comunidad LGBTQ+ y no sobre elementos de carácter sexual.

Lo primero a notar es que el tema recurrente Queer Media está asociado a la categoría R porque solo aparecía en películas con dicha clasificación en nuestro conjunto de datos. Si bien esto podría ser una coincidencia, también puede que muestre un sesgo en la industria. Existen otros trabajos donde los temas recurrentes ya han sido utilizados para buscar sesgos en la industria cinematográfica [13]. Creemos que el proceso de asociar temas recurrentes a una categoría podría ayudar a detectar sesgos e identificar de manera más clara que define a dicha categoría en la práctica.

De igual forma, creemos que es importante que al realizar el proceso de prototipado por categorías se revisen los temas recurrentes asociados a cada categoría, ya que una asociación incorrecta puede llevar a acciones de discriminación en la práctica. Muchas veces al mantener los datos en una forma abstracta se puede propiciar la generación de sistemas clasificadores que cometan actos discriminatorios. Y ya sea que la causa sea un conjunto de datos no lo suficientemente extenso o que el algoritmo clasificador aprenda las decisiones hechas previamente por personas con esos sesgos, las consecuencias pueden ser muy graves. Los temas recurrentes son datos que tienen una descripción más concisa para el ser humano y podemos opinar si el uso de alguno es éticamente correcto. De tal forma que si se considera que un tema recurrente está indebidamente asociado a una clasificación, puede ser removido para no ser usado, acción que resulta difícil ejercer con datos más abstractos.

[1] `https://tvtropes.org/pmwiki/pmwiki.php/Main/QueerMedia`, 11 de mayo del 2022.

Capítulo 6

Conclusiones

Los experimentos realizados en este trabajo comprobaron nuestra hipótesis. Tanto la predicción del género cinematográfico, como la predicción de la clasificación por edades, mostraron una mejoría significativa al usar la información de temas recurrentes. Los resultados de predicción del género cinematográfico sugieren que se pueden usar los temas recurrentes asociados teóricamente a ciertas categorías. Por otra parte, el resultado de la predicción de clasificación por edades señala que en caso que no existan temas recurrentes asociados a categorías, se puede usar el método desarrollado en esta tesis para la selección y agrupación de temas recurrentes por categorías. Dicho sistema también proporciona un método para asociar una categoría con elementos accesibles e inteligibles, los temas recurrentes, que describen o ejemplifican dicha categoría. Lo cual es de utilidad para analizar los estándares y las reglas implícitas que siguen las clasificaciones hechas por personas. Las cuales heredan los sistemas de aprendizaje de máquina al usar estos datos como base.

Asimismo, al poder presentar una mejora en las dos tareas de predicción propuestas, en cada una de sus instancias con distintos algoritmos clasificadores y distintas representaciones del texto, podemos teorizar que posiblemente también otras tareas de predicción en el ámbito cinematográfico se beneficiaría del uso de la información provista por los temas recurrentes. Por lo cual, cumplimos el objetivo principal de esta tesis y al mismo tiempo podemos proponer una nueva fuente de información para otras tareas en el ámbito. Durante este trabajo desarrollamos una base de datos que relaciona temas recurrentes con películas. La cual puede ser un recurso valioso para el desarrollo e investigación de nuevos

proyectos en el área.

6.1. Trabajo a futuro

Como trabajo a futuro, proponemos realizar algunas tareas mencionadas en la Sección de Trabajo Relacionado haciendo uso de temas recurrentes. Principalmente, proponemos su uso para la tarea de predicción de calificaciones por usuario. Posiblemente, el uso de temas recurrentes y el sistema creado, descrito en esta tesis, podrían ser de utilidad para crear grupos de temas recurrentes que son del interés de cada usuario. Igualmente, reconocemos que una de las limitantes del trabajo desarrollado son las asociaciones limitadas entre películas y temas recurrentes. Ya que, si bien los datos de TV Tropes son bastante extensos, son desarrollados por personas y el área a abarcar es bastante extensa. Por ello, creemos que un buen trabajo a futuro sería la detección de temas recurrentes en una película, usando información como el guion cinematográfico y el director de una película.

En caso que se desee colaborar para cualquier trabajo a futuro relacionado con el tema u la obtención de los recursos aquí dispuestos, favor de comunicarse al correo: alejnadro-martinezt@hotmail.com

Bibliografía

[1] Queer media, tv tropes.

[2] Apte, N., Forssell, M., and Sidhwa, A. Predicting movie revenue. *CS229, Stanford University* (2011).

[3] Aragón, M. E., López-Monroy, A. P., González-Gurrola, L. C., and Montes, M. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019), pp. 1481–1486.

[4] Arora, G., Kumar, A., Devre, G. S., and Ghumare, A. Movie recommendation system based on users' similarity. *International journal of computer science and mobile computing 3*, 4 (2014), 765–770.

[5] Battu, V., Batchu, V., Gangula, R. R. R., Dakannagari, M. M. K. R., and Mamidi, R. Predicting the genre and rating of a movie based on its synopsis. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation* (2018).

[6] Bengio, Y., Goodfellow, I., and Courville, A. *Deep learning*, vol. 1. MIT press Massachusetts, USA:, 2017.

[7] Bhave, A., Kulkarni, H., Biramane, V., and Kosamkar, P. Role of different factors in predicting movie success. In *2015 International Conference on Pervasive Computing (ICPC)* (2015), IEEE, pp. 1–4.

[8] Bishop, C. M., and Nasrabadi, N. M. *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[9] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics 5* (2017), 135–146.

[10] Chang, C.-H., Su, H.-T., Hsu, J.-H., Wang, Y.-S., Chang, Y.-C., Liu, Z. Y., Chang, Y.-L., Cheng, W.-F., Wang, K.-J., and Hsu, W. H. Situation and behavior understanding by trope detection on films. In *Proceedings of the Web Conference 2021* (2021), pp. 3188–3198.

[11] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[12] Funk, S. Netflix update: Try this at home, 2006.

[13] Gala, D., Khursheed, M. O., Lerner, H., O'Connor, B., and Iyyer, M. Analyzing gender bias within narrative tropes. *arXiv preprint arXiv:2011.00092* (2020).

[14] García-Ortega, R. H., Merelo-Guervós, J. J., Sánchez, P. G., and Pitaru, G. Overview of pictropes, a film trope dataset. *arXiv preprint arXiv:1809.10959* (2018).

[15] García-Ortega, R. H., Sánchez, P. G., and Merelo-Guervós, J. J. Tropes in films: an initial analysis. *arXiv preprint arXiv:2006.05380* (2020).

[16] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM 35*, 12 (1992), 61–70.

[17] Hoang, Q. Predicting movie genres based on plot summaries. *arXiv preprint arXiv:1801.04813* (2018).

[18] Joshi, M., Das, D., Gimpel, K., and Smith, N. A. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2010), pp. 293–296.

[19] Jurafsky, D., and Martin, J. H. Speech and language processing, chapter chapter 19: Vector semantics. *Prentice Hall, 3rd edition. Draft of August 24* (2015), 2015.

[20] Lim, Y. J., and Teh, Y. W. Variational bayesian approach to movie rating prediction. In *Proceedings of KDD cup and workshop* (2007), vol. 7, Citeseer, pp. 15–21.

[21] Maddrey, J. *Nightmares in red, white and blue: The evolution of the American horror film*. McFarland, 2010.

[22] Martinez, V. R., Somandepalli, K., Singla, K., Ramakrishna, A., Uhls, Y. T., and Narayanan, S. Violence rating prediction from movie scripts. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), vol. 33, pp. 671–678.

[23] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[24] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[25] Nawar, A., Toma, N. T., Al Mamun, S., Kaiser, M. S., Mahmud, M., and Rahman, M. A. Cross-content recommendation between movie and book using machine learning. In *2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT)* (2021), IEEE, pp. 1–6.

[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[27] Plackett, R. L. Karl pearson and the chi-squared test. *International Statistical Review/Revue Internationale de Statistique* (1983), 59–72.

[28] Rizzo, M. *The art direction handbook for film.* Taylor & Francis, 2005.

[29] Robertson, S. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation* (2004).

[30] Rong, X. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* (2014).

[31] Shafaei, M., López-Monroy, A. P., and Solorio, T. Exploiting textual, visual, and product features for predicting the likeability of movies. In *The Thirty-Second International Flairs Conference* (2019).

[32] Shafaei, M., Samghabadi, N. S., Kar, S., and Solorio, T. Age suitability rating: Predicting the mpaa rating based on movie dialogues. In *Proceedings of The 12th Language Resources and Evaluation Conference* (2020), pp. 1327–1335.

[33] Shalev-Shwartz, S., and Ben-David, S. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[34] Simões, G. S., Wehrmann, J., Barros, R. C., and Ruiz, D. D. Movie genre classification with convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)* (2016), IEEE, pp. 259–266.

[35] Stevens, E., Antiga, L., and Viehmann, T. *Deep learning with PyTorch.* Manning Publications, 2020.

[36] Suárez, E. J. C. Tutorial sobre máquinas de vectores soporte (svm). *Tutorial sobre Máquinas de Vectores Soporte (SVM)* (2014), 1–12.

[37] Thavikulwat, P. Affinity propagation: a clustering algorithm for computer-assisted business simulations and experiential exercises. In *Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference* (2008), vol. 35.

[38] Viveros-Jiménez, F., Sanchez-Perez, M. A., Gómez-Adorno, H., Posadas-Durán, J.-P., Sidorov, G., and Gelbukh, A. Improving the boilerpipe algorithm for boilerplate removal in news articles using html tree structure. *Computación y Sistemas 22*, 2 (2018), 483–489.

[39] Wackerly, D. D., Muñoz, R., Humbertotr, J., et al. *Estadística matemática con aplicaciones*. No. 519.5 W3. 2010.

[40] Whitford, L. Tv tropes. *Reference Reviews* (2015).

[41] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. Top 10 algorithms in data mining. *Knowledge and information systems 14*, 1 (2008), 1–37.

[42] Yedla, M., Pathakota, S. R., and Srinivasa, T. Enhancing k-means clustering algorithm with improved initial center. *International Journal of computer science and information technologies 1*, 2 (2010), 121–125.