
Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

# Extracción de información de películas a través de subtítulos utilizando atributos sintácticos y semánticos en español

Tesis presentada para obtener el título de
Licenciado en Ciencias de la Computación

Alejandro Daniel Masseroli

Director: Lic. Pablo Brusco

Codirector: Dr. Edgar Altszyler

Buenos Aires, 2016

# EXTRACCIÓN DE INFORMACIÓN DE PELÍCULAS A TRAVÉS DE SUBTÍTULOS UTILIZANDO ATRIBUTOS SINTÁCTICOS Y SEMÁNTICOS EN ESPAÑOL

El *procesamiento de lenguaje natural* es un área que combina la inteligencia artificial y la lingüística para permitir la comprensión automática de información expresada en lenguaje humano. En el contexto de esta tesis, lo aplicamos al estudio de diálogos de películas.

El objetivo principal de este trabajo es analizar en qué medida el género de una película se relaciona no sólo con la trama sino también con la estructura gramatical y el contenido emocional de sus diálogos, tomando como representación de los mismos sus subtítulos en español.

Con este fin, estudiamos de qué manera el género está asociado a la estructura gramatical de los diálogos de las películas, sin tomar en cuenta el contenido de aquello que se está diciendo. Luego, en una segunda etapa analizamos cómo la relación existente entre el género y las emociones transmitidas en el contenido de los diálogos.

Para realizar nuestros experimentos, extrajimos atributos a partir de subtítulos que capturan aspectos de la estructura gramatical y del contenido emocional que luego fueron utilizados en clasificadores automáticos que infieren el género de películas a partir de esta información.

El análisis realizado nos permite concluir que el género de una película tiene se relaciona considerablemente tanto con la gramática de sus diálogos como con las emociones que estos transmiten.

**Palabras claves:** Aprendizaje Automático, Procesamiento de Lenguaje Natural, Clasificación Automática de Texto, Análisis Automático de Emociones.

# FILM INFORMATION EXTRACTION USING SYNTACTIC AND SEMANTIC FEATURES FROM SPANISH SUBTITLES

*Natural Language Processing* is an area that combines both artificial intelligence and linguistics to pursue the automatic understanding of information expressed in human language. In the context of this thesis, we apply it to the study of film subtitles.

Our main goal is to analyze to what extent the genre of a film is related not only to the plot but also to the grammatical structure and emotional content of its dialogues, taking its Spanish subtitles as its representation.

To this end, we study how genre is related to the grammatical structure of the film dialogues, without taking into account the content of what is being said. Then, in a second stage, we analyze the relationship between genre and the emotions transmitted in the content of the dialogues.

In order to perform our experiments, we extracted features from the subtitles that capture aspects of grammatical structure and emotional content. Then they were fed into automatic classifiers that infer the genre of films from this information.

The analysis allows us to conclude that the genre of a film is substantially related both to the grammar of its dialogues and to the emotions it transmits.

**Keywords:** Machine Learning, Natural Language Processing, Automatic Text Classification, Automatic Emotion Analysis.

## AGRADECIMIENTOS


- A mis viejos Daniel y Liliana, por ayudarme y acompañarme en todo el camino recorrido hasta hoy. Sin su apoyo y esfuerzo nunca hubiera llegado a ser lo que soy.

- A mi hermano Pablo, por demostrarme con el ejemplo los frutos del estudio y el sacrificio.

- A mi compañera de facultad y de vida, Lore, por estar a mi lado y bancarme tanto. Por contagiarme día a día sus ganas de avanzar y mejorar en todo lo que hace y más.

- A los amigos que me llevo de la facultad, con quiénes compartimos momentos de dificultad y de mucha alegría. En especial a Andrés, Ivan, Fran, Lore y Pablo, mis compañeros de cursadas, estudio, tps, salidas, viajes y tantas cosas más.

- A mis directores de tesis Pablo y Edy, también incluidos en el punto anterior, que me dieron el empujón final y me acompañaron y aconsejaron tan bien.

- A mis amigos y amigas de siempre, que siguen a pesar del tiempo y se bancaron que desaparezca los últimos meses.

- A mis tíos y primos Ana, Miguel, Leandro y Silvina por hacerme conocer el mundo de exactas y ayudarme a definirme a estudiar esta carrera que tanto disfruté (y sufrí).

- A Agustín Gravano y Facundo Carrillo, que me ayudaron a sentar las ideas originales de la tesis y me facilitaron recursos útiles para su confección.

- A los jurados Santiago Figueira y Ramiro Gálvez, por su buena predisposición para la corrección de la tesis.

- A la gente de opensubtitles[1], que amablemente exportó el corpus de subtítulos utilizado para el trabajo.


[1] `http://www.opensubtitles.org/`

*A Mamá, Papá, Papo y Lore*

# Índice general

# 1. INTRODUCCIÓN

## 1.1. Definición del problema

Las películas son concebidas desde sus inicios como un medio de comunicación en los que se transmite un mensaje cargado de emociones y sensaciones a percibir por el espectador. En ese sentido, las películas pueden ser clasificadas según su género cinematográfico, el cual la enmarca tanto en su contenido como en su estructura, permitiéndole al espectador contar con una información mínima de su guión.

Es evidente que el género de una película se relaciona tanto con la trama como con el tipo de acontecimientos que tienen lugar en su desarrollo. Lo que no resulta obvio es la relación que existe entre el género y la gramática[1] y las características estructurales de sus diálogos. En este contexto, el objetivo principal del presente trabajo es analizar en qué medida el género de una película se encuentra asociado no sólo al contenido sino también a la gramática de sus diálogos.

Para lograr este objetivo, estudiaremos características que puedan obtenerse a partir de los diálogos para comprender las diferencias entre cada género. Debido a la restricción de no contar con el libreto original, utilizaremos los subtítulos como la representación más cercana de la cual disponemos para los diálogos. Dada esta restricción, intentaremos distinguir de la mejor manera posible el género de una película utilizando sólo características que extraeremos a partir del texto presente en los subtítulos. Luego, analizaremos qué aspectos fueron los que dieron lugar a la distinción lo cual nos permitirá elaborar conclusiones sobre la forma en que se desarrollan los diálogos según el género.

En general, los subtítulos pueden corresponderse tanto a trascripciones del diálogo en el mismo idioma o a traducciones de éste en un idioma distinto al original. En el primer caso, los subtítulos están principalmente destinados a personas sordas o hipoacúsicas, y tienen el fin de comunicar todo lo que acontece en la película, por lo que no sólo incluyen el diálogo sino también otros detalles que narran la escena en cuestión. En el segundo caso, el objetivo de los subtítulos es el de transmitir los diálogos que tienen lugar en la película para que puedan ser entendidos por los hablantes del idioma al que se traducen. Por este motivo, junto con la intención de realizar un trabajo en nuestro propio idioma para aportar a la comunidad, utilizaremos subtítulos en el idioma español para analizar películas cuyo idioma original es, en la mayoría de los casos, el inglés. La desventaja de este enfoque es que los subtítulos han pasado por un proceso de traducción, lo que puede ocasionar que se pierda información de los diálogos originales, o que incluso se agreguen variaciones propias de la traducción en cuestión.

Al basarnos únicamente en atributos de texto, las técnicas que utilizaremos serán de *clasificación de texto*, perteneciendo a los campos del procesamiento del lenguaje natural y del aprendizaje automático. El área de clasificación de texto es un campo de investigación que ha crecido a pasos agigantados desde la aparición masiva de documentos en internet.

[1] La gramática comprende, entre otros aspectos, la estructura de las palabras y las maneras en las que estas se enlazan en una oración

En líneas generales, en este tipo de tareas se propone clasificar un documento según un conjunto predefinido de categorías. En una de las áreas relevantes dentro del campo, el objetivo es clasificar dichos documentos según el género narrativo al que pertenecen, tarea intrínsecamente ligada a nuestro trabajo.

## 1.2. Trabajo Previo

Mucho esfuerzo se ha puesto en el reconocimiento automático de propiedades de películas y, en particular, en la detección del género de las mismas. Gran parte de estos trabajos han utilizado características visuales y acústicas para dicha tarea. Uno de los precursores de los estudios de este tipo fue el trabajo de Fischer, Lienhart y Effelsberg (1995), dónde se utilizan atributos audiovisuales para clasificar videos en las categorías: noticiero, carrera de autos, partido de tenis, publicidad y dibujo animado. Con el mismo objetivo, pero sobre distintos conjuntos de géneros, en Yuan, Song y Shen (2002) y Yuan y col. (2006) se utilizan atributos puramente visuales con las técnicas de clasificación de árboles de decisión y SVM respectivamente.

En otros trabajos, se combinan los atributos audiovisuales de las películas con los extraídos del texto de sus subtítulos. En particular, muchos se centran en el estudio de subtítulos en el idioma español. Éste es el caso de Helmer y Ji (2012), dónde se utilizan atributos extraídos del video y de los subtítulos en español de trailers de películas, para clasificar su género y su rating MPAA.[2] Los autores utilizan un corpus de 312 trailers, de los cuáles sólo 100 tienen su subtítulo asociado, por lo que sugieren como trabajo futuro repetir sus experimentos sobre un corpus más robusto.

A diferencia de los anteriores, también se han utilizado solamente subtítulos como fuente de información. Por ejemplo en Katsiouli, Tsetsos y Hadjiefthymiades (2007) se extraen automáticamente categorías de subtítulos en español de un corpus de documentales, y proponen repetir el procedimiento para películas de cualquier género. En nuestro trabajo nos centraremos en este tipo de atributos, pero para un corpus de tamaño considerable, compuesta por subtítulos de películas de varios géneros.

Un punto que tienen en común los trabajos detallados anteriormente es que se enfocan principalmente en intentar clasificar los géneros de la mejor manera posible, mientras que el análisis del contenido de los diálogos que produce esta diferenciación ha sido dejado de lado. En nuestro trabajo, utilizaremos la clasificación automática de los géneros como una herramienta que nos permitirá profundizar en este análisis.

## 1.3. Estructura de la tesis

En el capítulo 2 describiremos el conjunto de técnicas del área de aprendizaje automático que utilizamos en la búsqueda de cumplir el objetivo propuesto. Más adelante, en el capítulo 3, explicaremos el proceso seguido en el transcurso del trabajo, partiendo desde la obtención inicial de recursos y llegando hasta las pruebas de confianza y desempeño sobre el clasificador construido. En el capítulo 4, plantearemos las hipótesis de nuestros experimentos y analizaremos los resultados obtenidos. Finalmente, en el capítulo 5, dare-

[2] `https://en.wikipedia.org/wiki/Motion_Picture_Association_of_America`

mos un cierre al trabajo comentando una serie de conclusiones finales sobre los resultados alcanzados, y plantearemos posibles formas de continuar nuestro trabajo.

# 2. TÉCNICAS UTILIZADAS

El objetivo de este trabajo es analizar en qué medida se relaciona el género de una película tanto con la trama como con la gramática de sus diálogos. En particular, se estudiará qué características relevantes pueden ser extraídas de los subtítulos de las películas para identificar su género. Una herramienta computacional que se puede utilizar para abordar el estudio planteado es la del aprendizaje automático.

El *aprendizaje automático* es una rama de la inteligencia artificial que se dedica al desarrollo de técnicas que se utilizan para inferir modelos que explican comportamientos a partir del estudio de casos conocidos. Luego, dichos modelos permiten estudiar o generalizar el conocimiento deseado sin la necesidad de ser programados explícitamente.

Entre los diferentes enfoques existentes dentro del aprendizaje automático, se encuentra la técnica del *aprendizaje supervisado*. Su objetivo es construir automáticamente modelos que estimen una función a partir de un conjunto de datos de entrada sobre los cuales se conocen sus respectivas salidas. Esto permite predecir la salida para otros datos sobre los cuales es desconocida. Cuando la salida de dicha función es de tipo categórico, se trata de un modelo de *clasificación*. La mayor parte de estos modelos utilizan como entrada vectores de valores numéricos. Por este motivo, a partir de los datos, deben extraerse características que contengan la mayor cantidad de información posible para distinguir a cada una de las clases. Exhibimos en la figura 2.1 un diagrama sobre el esquema que aplicaremos a nuestra tarea particular.

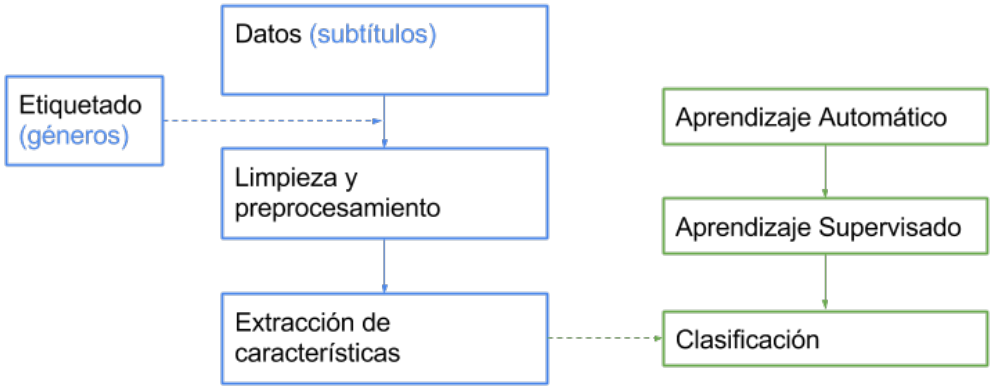


*Fig. 2.1:* Esquema de clasificación.

De la definición explicitada previamente, se desprende que la tarea de predecir el género de las películas a través de sus subtítulos, es de clasificación.

En las próximas secciones detallamos la extracción de atributos, y fundamentamos la elección del algoritmo de clasificación utilizado.

## 2.1. Extracción de atributos

Como mencionamos anteriormente, las técnicas de aprendizaje supervisado se alimentan de vectores de valores numéricos que representan a cada instancia. El proceso de obtener dichos valores es denominado *extracción de atributos* (o extracción de características principales) y en general consta de uno o más procesos que se encargan de analizar cada instancia (en este caso cada subtítulo) y devolver una o más características del mismo.

A continuación comentamos los aspectos técnicos y conceptuales más relevantes que permitieron luego implementar cada uno de los extractores que fueron utilizados.

### 2.1.1. POS tagging

Para el estudio de atributos gramaticales nos basaremos fuertemente en el POS tag de las palabras de los subtítulos. El POS (Part Of Speech) tagging, o etiquetado gramatical, consiste en asignar a cada palabra de un texto, la categoría gramatical correspondiente. En su versión más simple, cada palabra puede etiquetarse según su función en el texto en una de las siguientes categorías: adjetivo, adverbio, artículo, conjunción, interjección, preposición, pronombre, sustantivo y verbo. Exhibimos en la figura 2.1 un ejemplo de etiquetado gramatical.

| **Oración** | **A donde vamos no necesitamos caminos** | | | | | |
|---|---|---|---|---|---|---|
| **Palabras** | A | donde | vamos | no | necesitamos | caminos |
| **POS tag** | *preposición* | *pronombre* | *verbo* | *adverbio* | *verbo* | *sustantivo* |

*Tab. 2.1:* Ejemplo de etiquetado gramatical.

Existen múltiples algoritmos (llamados *taggers*) confeccionados para realizar esta tarea. Nosotros utilizaremos una versión desarrollada por el Instituto de Lingüística Computacional de la Universidad de Stuttgart[1], descrito en (Schmid 2013) y (Schmid 1995). Dada una frase en español, el tagger devuelve una de 74 etiquetas posibles para cada una de las palabras dentro de dicha frase. Exhibimos la lista de etiquetas de POS tag en español completa junto con su descripción, en el Apéndice Etiquetas de POS tag en español.

#### Verbos dinámicos y estáticos

Existe una subclasificación posible para los verbos en las categorías *dinámicos* y *estáticos*, dependiendo de si refieren o no a acciones. Por ejemplo, el verbo "correr" es dinámico, mientras que el verbo "entender" es estático.

La implementación de POS tagger utilizada en nuestro trabajo incluye etiquetas sobre distintos tipos de verbos. Sin embargo, no provee información acerca de si los verbos son dinámicos o estáticos. Para conseguir dicha clasificación, utilizamos un recurso externo[2] del que extrajimos una lista con 58 de los verbos estáticos más frecuentes del idioma español. Si un verbo está presente en dicha lista, será anotado como *verbo estático*; si no, supondremos que no lo es y será anotado como *verbo dinámico*.

[1] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[2] `https://bibliotecadeinvestigaciones.wordpress.com/ingles/stative-verbs-los-verbos-de-estado/`

### 2.1.2. Emociones

Para el análisis de emociones de la segunda parte del trabajo, extraeremos los atributos de emociones utilizando el recurso "EmoLex".[3] (Mohammad y Turney 2013) Este consiste en una lista de 14,182 palabras en inglés, que luego fueron traducidas automáticamente al idioma español utilizando el traductor de google.[4] Cada una de estas palabras está asociada a ninguna, una o más de ocho emociones básicas (alegría, tristeza, enojo, miedo, disgusto, sorpresa, confianza y anticipación) y a una, o ninguna, de las polaridades positiva y negativa. Al ser este un recurso concebido en inglés y posteriormente traducido al español con un mecanismo automático, contiene varias palabras repetidas con distintos valores de emociones y polaridad. En estos casos, se decidió agrupar en la palabra todas las emociones y valores de polaridad presentes en cada una de sus apariciones.

## 2.2. Modelo de clasificación

Suponiendo que ya contamos con la traducción de los datos a vectores numéricos, tendremos los elementos necesarios para comenzar con la tarea de clasificación. De todas formas, queda una decisión por tomar: debemos decidir si nos interesa construir un modelo que explique cada género por separado (por ejemplo a través de modelos generativos) o queremos construir un modelo que permita encontrar fronteras en el espacio en que están representados los datos, tales que permitan separar los distintos géneros. En este caso, optamos por utilizar un modelo discriminativo: *Random Forest*.

Random Forest es un tipo de modelo discriminativo basado en la combinación de múltiples árboles de decisión. A continuación presentaremos un breve resumen de estas técnicas y veremos cómo utilizar estos modelos para cumplir los objetivos planteados.

### 2.2.1. Árboles de decisión

Un *árbol de decisión* es un modelo de clasificación automática cuyo objetivo es encontrar automáticamente barreras o cortes en el espacio de los vectores de atributos, tales que separen instancias según su clase tanto como sea posible. Con este fin, dado un conjunto de datos de entrenamiento etiquetado, se comienza buscando en el espacio de atributos la combinación de $< atributo, corte >$ que mejor permita discriminar las instancias de cada clase según alguna de las métricas de evaluación existentes (daremos más detalles sobre ellas en los siguientes párrafos).

En la figura 2.2 exhibimos un conjunto de datos de ejemplo en donde contamos sólo con los valores de dos atributos. Se puede ver que estos datos pertenecen a dos clases distintas (círculos y cruces), que podemos intentar separar utilizando árboles de decisión. Para ello, nos gustaría encontrar sectores en la figura tales que delimiten de la mejor manera posible las distintas clases.

Siguiendo el primer paso que describimos anteriormente, buscamos el corte inicial más adecuado, que en este caso podría estar dado por la condición $cantidad\ de\ palabras = 50$.

[3] `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`
[4] `https://translate.google.com/`

Esta condición de corte del espacio será la raíz del árbol, y se generará a partir de ella una rama para cada sector delimitado por esa condición.

Una vez separado el espacio según el primer corte, se repite el mismo procedimiento para cada sector creado en donde se examinan únicamente las instancias delimitadas por el sector resultante. Este proceso se repite hasta que el corte a realizar ya no mejore significativamente la distinción de clases del nodo superior según nuestra métrica.

En este punto, sólo resta determinar la clase que asignaremos a cada sector. Volviendo a nuestro ejemplo, puede verse que cada sector contiene una clase que presenta mayor cantidad de instancias que el resto. Al finalizar el proceso, cada hoja del árbol se corresponde con un sector del espacio de atributos al que habrá que asignarle una clase. La clase de cada sector se suele identificar como aquella que contiene mayor cantidad de instancias.

En general, alcanzar el punto donde continuar haciendo cortes en el espacio ya no mejora la discriminación de clases, es una manera de que el algoritmo detenga la construcción del árbol. Sin embargo, también es usual configurar otra manera de detener el crecimiento del modelo, que consiste en introducir una profundidad máxima más allá de la cuál no se le permitirá crecer.

Nuevamente en la figura, indicamos como quedan delimitadas las clases a devolver en cada sector para el caso de ejemplo.

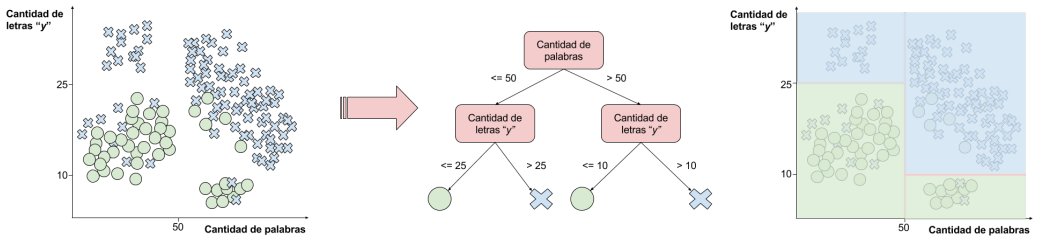


*Fig. 2.2:* Ejemplo de árbol de decisión con su delimitación del espacio de atributos.

Una vez construido nuestro modelo, queremos utilizarlo para clasificar una instancia de clase desconocida. Para esto, comenzamos obteniendo su vector de atributos correspondiente. Luego este vector es alimentado al modelo, evaluando las condiciones desde la raíz y siguiendo la rama dada por sus valores hasta llegar a una hoja, que se corresponderá con la clase a predecir. Esto equivale a determinar en qué cuadrante del espacio de vectores de atributos, delimitado por el árbol de decisión, se ubica la instancia a clasificar.

Según la implementación y el tipo de tarea, hay distintos parámetros para configurar en el modelo de los árboles de decisión que puede ser interesante estudiar. Entre los más relevantes se encuentran:

1. La *profundidad* máxima que se le permitirá alcanzar al árbol, introduciendo un nuevo criterio de parada en la construcción del modelo.

2. Un *peso asociado a cada clase* que utilizará el clasificador con el objetivo de compensar problemas de desbalanceo. Dependiendo la implementación, puede modificar las decisiones de donde cortar o de que clase considerar para cada sector.

3. Una *métrica de evaluación de atributos* para evaluar la calidad de un atributo y elegir el par $< atributo, corte >$ que se utilizará en cada corte. Entre las métricas más comúnmente utilizadas se encuentran "ganancia de información" e "impureza de Gini". La primera se basa en la entropía o información aportada por los datos a cada lado del corte. La segunda, en cambio, se basa en la probabilidad de asignar de manera aleatoria la etiqueta incorrecta a una instancia, asumiendo la distribución de las etiquetas del sector del corte al que pertenece. Para mayores detalles sobre las métricas evaluación, ver Raileanu y Stoffel (2004). En este trabajo, utilizaremos la métrica *impureza de Gini*, que es la métrica por defecto de la implementación utilizada.[5]

Una propiedad esencial de los árboles de decisión en el contexto de nuestro trabajo, es que es posible extraer del modelo la información aportada por cada atributo a la clasificación. Aquellos que aportan la mayor cantidad de información, son los que permiten discriminar mejor entre las distintas clases y, por lo tanto, los más importantes para la tarea de clasificación planteada. Para un atributo dado, la importancia se calcula combinando la información aportada en cada nodo en el que aparece, medido a través de la métrica de evaluación detallada en el ítem 3.

Los árboles de decisión presentan también algunas desventajas conocidas. Son propensos a *sobreajustar*, es decir, el modelo podría describir los datos de entrenamiento pero no generalizar correctamente. Esto puede ocasionar que su desempeño en datos no vistos durante el entrenamiento sea significativamente peor que en otras instancias. Por otra parte, también pueden presentar *alta varianza*, lo que significa que una leve variación en los datos de entrenamiento pueden cambiar en forma considerable el modelo obtenido.

Con el correr del tiempo, nuevas técnicas fueron desarrolladas con el fin de alcanzar mayor poder de generalización sin perder la simplicidad y transparencia que proporciona la técnica de árboles de decisión. Entre estas técnicas se encuentran los llamados *modelos de ensamble*: modelos conformados por múltiples clasificadores utilizados de manera tal que el resultado final de la clasificación será algún tipo de combinación de los resultados parciales de cada clasificador. En ese trabajo utilizaremos un algoritmo de ensamble que está basado en la combinación de múltiples árboles de decisión: *Random Forest*.

### 2.2.2. Random Forest

*Random forest*, presentado en Breiman (2001), se basa en la combinación de múltiples árboles de decisión. Cada uno de estos árboles será construido de manera independiente utilizando una técnica conocida como "boostrapping", que consiste en eligir al azar una muestra con reposición de la misma cantidad de instancias que el conjunto de datos de entrenamiento original. De esta manera, cada árbol es construido sobre un conjunto de instancias levemente diferente. La decisión final de la clasificación estará definida por una combinación de los resultados de cada árbol del bosque. En el trabajo de Breiman (2001), se devuelve la clase elegida por la mayoría de dichos árboles. En la implementación que utilizamos, el modelo devolverá la clase con mayor media de probabilidad entre todos los

[5] `http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier`

árboles.[6]

Otro aspecto importante del modelo, es que en cada división de un nodo, la elección del atributo de separación no se realiza sobre todos los atributos disponibles, sino sobre un subconjunto de ellos tomado al azar. Esto contribuye a que todos los árboles del bosque sean distintos entre sí, y no haya un pequeño conjunto de atributos dominantes cerca de las raíces de todos.

Para definir un modelo de Random Forest, junto con la configuración de sus árboles internos mediante los parámetros presentados en la sección 2.2.1, deben disponerse otros dos aspectos importantes:

- La *cantidad de árboles* que se utilizarán.
- El número de atributos a considerar en el momento de buscar el mejor corte para cada nodo.

Es necesario detallar que el método tiene un mecanismo para, una vez entrenado el modelo, analizar cuáles fueron los atributos que permitieron separar las clases de la mejor manera posible. Existe más de una manera de calcular esta importancia de los atributos. Nosotros trabajaremos con la más usual, que consiste en obtener la importancia de cada atributo de manera análoga a la explicada en la sección 2.2.1, pero promediada entre las apariciones del atributo en cuestión en todos los árboles del bosque.

Esta última propiedad es la que nos lleva a decidirnos por utilizar Random Forest cómo nuestro algoritmo de clasificación frente a otros algoritmos del estado del arte, ya que facilitará el análisis de la importancia de cada atributo utilizado en la clasificación. Además, como mencionamos anteriormente, Random Forest tiene un alto poder de generalización en tareas relacionadas a la estudiada en este trabajo. Por ejemplo, en el trabajo de Helmer y Ji (2012) ya mencionado en la sección 1.2, Random Forest es la técnica que alcanza mejores resultados para predecir el ranking MPAA, y sólo es superada por SVM para la predicción del género. Otro ejemplo es el trabajo de Meyer, Leisch y Hornik (2003), donde realizan comparaciones de desempeño entre distintos modelos de clasificación sobre una amplia variedad de conjuntos de datos, resultando Random Forest uno de los algoritmos con mejor desempeño general. En las siguientes secciones detallaremos cómo podemos medir su capacidad predictiva.

## 2.3. Evaluación del modelo

Durante las secciones previas se habló acerca del poder de generalización, o *desempeño*, de los distintos modelos. Se trata de entender qué tan bien funcionó el modelo a la hora de realizar una clasificación. A continuación presentamos un esquema ampliamente utilizado para la evaluación de clasificadores.

[6] `http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.predict`

### 2.3.1. Validación cruzada y K-folds

Es deseable que al realizar una estimación del desempeño de un clasificador, esta se asemeje lo más posible al que se obtendría utilizándolo sobre datos nuevos. Una técnica para obtener una estimación con estas características consiste en separar un porcentaje de los datos de desarrollo para validar los modelos. Estos datos son elegidos de manera aleatoria para evitar tener un conjunto sesgado por cualquier orden intrínseco en ellos.

Sin embargo, los resultados así evaluados estarán atados a la forma en que son separados los datos. Una forma de minimizar las posibilidades de que el azar empeore la evaluación es utilizar la *validación cruzada de k-folds*. Este enfoque plantea el siguiente esquema:

1. Desordenar los datos aleatoriamente.
2. Separarlos en k subconjuntos del mismo tamaño, llamados folds.
3. Para i $\in \{1..k\}$:
   - 3.1. Entrenar sobre todos los folds menos el $i$ utilizando los datos junto a sus etiquetas.
   - 3.2. Evaluar sobre el fold $i$.
4. Obtener un resultado promediando o uniendo de alguna manera los resultados parciales de cada fold.

En nuestro trabajo, utilizaremos una configuración de *validación cruzada de 10 folds*. Cabe destacar que la técnica de validación cruzada debe ser acompañada con alguna métrica a utilizar para la evaluación de la clasificación, como puede verse en el paso 3.2. del esquema. A continuación detallaremos sus aspectos más relevantes.

### 2.3.2. Métricas

Las métricas exhibidas a continuación son funciones que, dada una clasificación, devuelven un valor numérico entre 0 y 1 que indica qué tan parecido a la realidad es el resultado arrojado por el modelo. Mientras más cercano a 1 sea dicho valor, mejor será la clasificación evaluada. Veamos, entonces, algunas de las métricas más ampliamente utilizadas:

#### Accuracy

*Accuracy* es el porcentaje de datos clasificados correctamente. La limitación que tiene esta métrica es que no provee información acerca de qué tipo de errores comete el modelo.

Para detallar otras que mejoren este aspecto, nos basaremos en la matriz de confusión de la clasificación binaria (ver Tabla 2.2):

| | **Positivo (predicho)** | **Negativo (predicho)** |
|---|---|---|
| **Positivo (real)** | *tp* | *fn* |
| **Negativo (real)** | *fp* | *tn* |

*Tab. 2.2:* Matriz de Confusión para Clasificación Binaria

### F-score

Utilizando la información de la matriz de confusión, podemos definir las métricas *precision* y *recall*:

$$Precision = \frac{tp}{tp + fp} \qquad Recall = \frac{tp}{tp + fn}$$

*Precision* puede entenderse como la fracción de instancias positivas, dentro de todas las que el sistema clasificó como positivas.

Por otro lado, *Recall* representa la fracción de instancias que el modelo clasificó como positivas, dentro del universo de instancias positivas reales.

Existe un compromiso entre precision y recall al momento de evaluar un sistema: es posible aumentar el recall tanto como se quiera, sacrificando precision. En el caso extremo, si se clasifican todas las instancias como *positivas*, se obtiene un recall perfecto (con valor 1) pero generalmente a costa de una muy baja precision.

Es por esto que existe otra métrica más robusta que tiene en cuenta la información de ambas. Esta es llamada *F-score* y se calcula con la siguiente fórmula:

$$F_\beta = (1 + \beta) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$$

En particular, con $\beta = 1$:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$F_1$ es una de las métricas más comúnmente utilizada y se entiende como la media armónica de Precisión y Recall del modelo. Comúnmente, tiene a estar más cerca de aquella de las dos que presente el menor valor.

### AUC

La métrica AUC (área bajo la curva) está definida a partir de la curva ROC (Receiver Operating Characteristic). Esta no es un valor numérico sino una representación gráfica de como varía la relación entre False Positive Rate (FPR) y True Positive Rate (TPR) al desplazar el umbral de discriminación. Este *umbral de discriminación o distinción* es, en

el contexto de Random Forest, el límite inferior de probabilidad de pertenencia a la clase positiva, a partir del cuál el modelo clasifica la instancia con esa clase. Definimos entonces:

$$TPR = Recall = \frac{tp}{tp + fn} \qquad FPR = \frac{fp}{fp + tn}$$

Ambos valores se distribuyen entre 0 y 1, y se espera que aumenten a medida que se disminuye el umbral de discriminación. Cuanto mejor sea el algoritmo de clasificación, mayor diferencia habrá entre TPR y FPR (con $TPR > FPR$).

Es usual contrastar la clasificación contra el caso aleatorio, caracterizado con una relación del tipo $FPR \approx TPR$ de la manera exhibida en la figura 2.3.

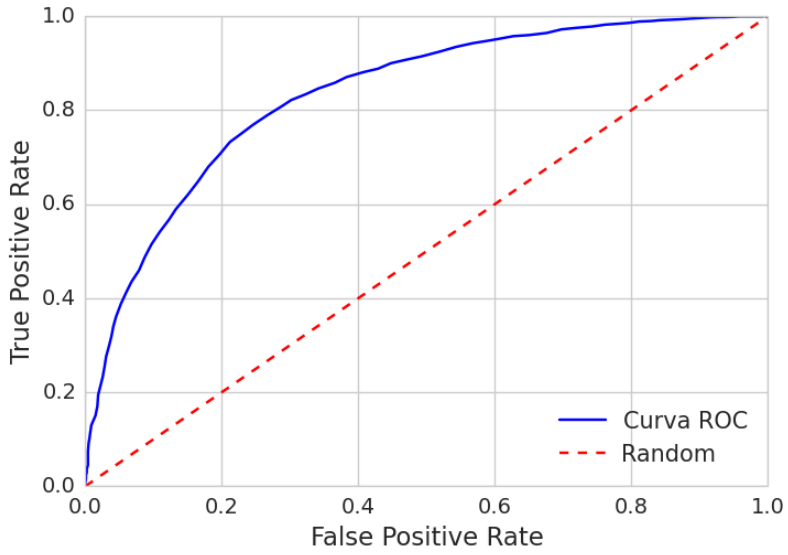


*Fig. 2.3:* Ejemplo de curva ROC.

El $AUC$ es una métrica para colapsar toda la información provista por la curva ROC del sistema en un único valor numérico entre 0 y 1, para facilitar la comparación entre distintos modelos. Se obtiene computando el área del espacio comprendido entre la curva ROC y los bordes inferiores del gráfico. Suele utilizarse el valor del área bajo la curva de un sistema aleatorio (0.5) como *baseline* de comparación.

# 3. DESARROLLO

En este punto, contamos con todos los conocimientos necesarios para comenzar con el estudio planteado. Nos proponemos, en una primera etapa, estudiar cuáles son los atributos gramaticales que explican de mejor manera las diferencias entre géneros y en qué medida permiten predecirlo. Luego, en una segunda etapa, realizaremos un análisis similar sobre las emociones que pueden ser detectadas en los subtítulos. En las secciones siguientes, detallaremos la línea de trabajo que seguimos durante el transcurso del proyecto, que nos permitirá realizar los experimentos que describiremos en el capítulo 4.

Comenzaremos, en la sección 3.1, explicando la manera en que obtuvimos el conjunto de subtítulos de películas junto con sus respectivos géneros para entrenar nuestro clasificador. En la sección 3.2 presentamos detalles sobre la elección de géneros sobre los que trabajaremos. En la sección 3.3, puntualizaremos los atributos gramaticales a utilizar a lo largo de todo el trabajo, y los atributos de emociones que incorporaremos al análisis en la segunda mitad del trabajo. En la sección 3.4 nos concentraremos en la construcción de nuestros clasificadores. Allí, analizaremos cómo configurar los parámetros para obtener el mejor desempeño posible y realizaremos una serie de pruebas para ganar confianza sobre el funcionamiento del sistema, para finalmente poder utilizarlo en el estudio deseado.

## 3.1. Obtención de corpus de subtítulos con anotación de géneros

Utilizamos como corpus, el conjunto de subtítulos en español de la página *opensubtitles*.[1]. Cada película cuenta con múltiples subtítulos disponibles. En nuestro caso particular, elegimos mantener un único representante para cada uno, prefiriendo siempre aquel que cuente con mayor cantidad de descargas y haya sido cargado en la página por un usuario de mayor puntaje.

Junto con los subtítulos, contamos con un identificador unívoco que lo vincula con su detalle en la página *imdb*.[2] Dentro de esta información, contamos con los *géneros* asociados a cada película.

## 3.2. Elección de géneros

Cómo explicamos en la sección 3.1, contamos con la información del género de los subtítulos de nuestro corpus. Dado que estamos buscando conexiones entre los subtítulos y propiedades asociadas a cada género, intentamos seleccionar géneros y películas que permitan explotar estas diferencias. Es por ello que en este trabajo utilizaremos el siguiente conjunto de géneros: *Acción, Comedia, Drama y Terror*.

Consideramos que estos géneros tienen la propiedad de ser fáciles de reconocer y separar por personas y adicionalmente, coinciden con algunos de los géneros que cuentan con mayor

[1] `http://www.opensubtitles.org/`
[2] `http://www.imdb.com`

cantidad de instancias en nuestro corpus. Por otra parte, la elección está basada en trabajos anteriores como en el artículo de Rasheed, Sheikh y Shah (2005) o el de Zhou y col. (2010) en los que se utilizó el mismo conjunto de géneros.

En nuestro corpus, la mayor parte de los subtítulos están clasificados como multi-género, es decir que contienen etiquetas de más de una categoría simultáneamente. Es el caso, por ejemplo, de la comedia-romántica y las películas de acción-terror, entre muchas otras variantes. A continuación estudiaremos cuánto ocurre esto en los géneros que hemos seleccionado, para luego entender si debemos utilizar este grupo de subtítulos tal como está, o si el conjunto de géneros parece demasiado difuso para que el clasificador pueda distinguirlos correctamente. En la figura 3.1 exhibimos un gráfico de barras con la cantidad de subtítulos por género con los que contamos en nuestro corpus, manteniendo o desechando las intersecciones entre ellos.

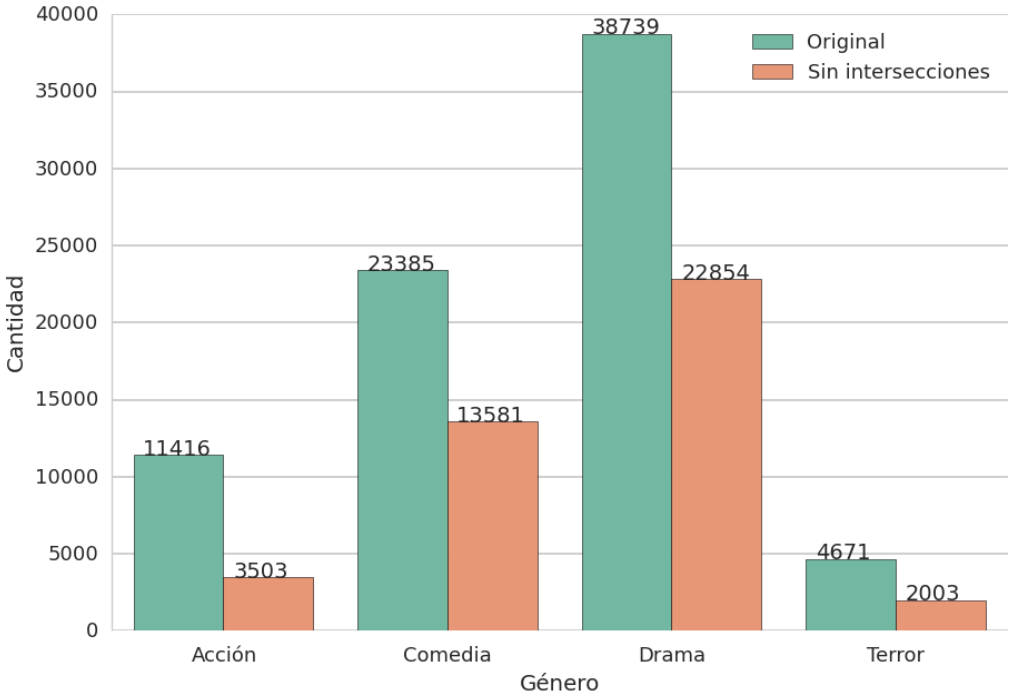


*Fig. 3.1:* Cantidad de subtítulos por género, con y sin intersecciones de géneros. Para cada género, la columna izquierda representa la cantidad original de subtítulos, y la columna derecha la cantidad de subtítulos luego de eliminar todos aquellos que pertenecen a más de un género dentro de las categorías seleccionadas.

Dado que estamos interesados en estudiar las diferencias entre géneros, decidimos trabajar con el conjunto de subtítulos que sólo contiene etiquetas pertenecientes a un único género dentro de los elegidos e ignorar el resto de las películas. Puede verse en el gráfico que la cantidad de instancias disminuye considerablemente pero, cómo mostramos más adelante, el número final será suficiente para que el clasificador funcione correctamente.

## 3.3. Elección de atributos

La siguiente tarea fue la elección de atributos a utilizar. Estos atributos serán extraídos de cada subtítulo para luego ser utilizados en la tarea de clasificación.

Previo a realizar dicha extracción, es necesario realizar un *pre-procesamiento* de los datos originales, tal que facilite el funcionamiento de los métodos de extracción. En este trabajo, fue necesario obtener una versión limpia del texto contenido en los subtítulos, quitando todos los caracteres propios del formato *SubRip*[3] y eliminando símbolos extraños que impedían el correcto funcionamiento del POS tagger, detallado en la sección 2.1.1. También es usual ver que en otros trabajos de este campo, se eliminen las "stopwords", o palabras vacías, del texto. Estas son palabras sin significado como los artículos, las preposiciones y los pronombres. Sin embargo, existen también estudios sobre la importancia que estas pueden tener en determinadas tareas de clasificación cómo en Stamatatos, Fakotakis y Kokkinakis (2000) y Yu (2008). En nuestra tesis hemos optado por mantenerlas ya que a pesar de tener poco significado semántico, pueden ser fundamentales para entender aspectos relacionados a la estructura sintáctica y gramatical del subtítulo.

Una vez realizado nuestro pre-procesamiento, debemos definir un procedimiento de extracción preciso que convierta cada subtítulo en un vector de valores numéricos. En este punto, dividiremos el análisis en dos etapas: una primera etapa en donde nos concentraremos únicamente en atributos gramaticales de los subtítulos mientras que en la segunda, evaluaremos atributos relacionados con emociones.

### 3.3.1. Atributos gramaticales

A continuación el detalle de cada uno de los tipos de atributos seleccionados, separados en grupos según sus características. Entre paréntesis aclaramos la cantidad de atributos que se extraen por cada tipo.

#### G1: Distribución de POS tags (71 atributos)

- **Proporción de cada POS tag (71):** cantidad de cada tipo de etiqueta de POS tag dividido por la cantidad de palabras totales en el subtítulo. Como resultado obtendremos un atributo por cada etiqueta de POS tag. Para más información, puede consultarse el Apéndice Etiquetas de POS tag en español.

#### G2: Complejidad del lenguaje (6 atributos)

- **Densidad de habla (1):** cantidad de palabras dividido por la suma de tiempos de todos los segmentos de habla en el subtítulo.[4]
- **Palabras por oración (1):** cantidad de palabras dividido por la cantidad de oraciones.[5]

[3] `https://es.wikipedia.org/wiki/SubRip#Ejemplo_de_archivo_SubRip_.28.srt.29`
[4] Este atributo no es puramente gramatical, ya que tiene utiliza el tiempo de los segmentos de habla.
[5] Para separar el subtítulo en oraciones se utilizó la librería de python *nltk* (`http://www.nltk.org/`)

- **Longitud media de palabras (1):** promedio de la longitud de las palabras del subtítulo. Las letras con tilde son contabilizadas como la misma letra, sin tilde.

**G4: Distribución de POS tags agrupados (5 atributos)**

- **Proporción de cada POS tag agrupado (5):** análogo a *G1* pero colapsando todas las etiquetas de POS tag en cinco categorías básicas: adjetivo, adverbio, sustantivo, verbo y otros.

En la Tabla 3.1 exponemos un ejemplo de extracción de atributos gramaticales. Para mayor detalle, puede consultarse el Apéndice Traducción de nombres de atributos.

| | |
|---|---|
| **Diálogo** | 00:01:20 –>00:01:22<br>MARTY, ¡tienes que venir conmigo!<br><br>00:01:23 –>00:01:25<br>¿A dónde?<br><br>00:01:27 –>00:01:29<br>¡De regreso al futuro! |

| Grupo | Atributo | Valor | Evidencia |
|---|---|---|---|
| **G1** | tag_percentage(ADV) | 0.066 | *dónde* |
| | tag_percentage(CM) | 0.066 | *“,”* |
| | tag_percentage(CQUE) | 0.066 | *que* |
| | tag_percentage(FS) | 0.2 | *“!”, “?”, “¡”* |
| | tag_percentage(NC) | 0.133 | *regreso, futuro* |
| | tag_percentage(NP) | 0.066 | *MARTY* |
| | tag_percentage(PAL) | 0.066 | *al* |
| | tag_percentage(PREP) | 0.2 | *conmigo, a, de* |
| | tag_percentage(VLfin) | 0.066 | *tienes* |
| | tag_percentage(VLinf) | 0.066 | *venir* |
| **G2** | avg_word_len | 3.533 | |
| | speech_density | 2.5 | |
| | words_per_sentence | 5.0 | |
| | uppercase | 0.133 | *MARTY* |
| | dynamic_verbs | 0.071 | *venir* |
| | stative_verbs | 0.071 | *tienes* |
| **G3** | letter_! | 0.031 | |
| | letter_? | 0.015 | |
| | letter_a | 0.047 | |
| | letter_c | 0.015 | |
| | letter_d | 0.047 | |
| | letter_e | 0.126 | |
| | letter_f | 0.015 | |
| | letter_g | 0.031 | |
| | letter_i | 0.047 | |
| | letter_l | 0.015 | |
| | letter_m | 0.031 | |
| | letter_n | 0.063 | |
| | letter_o | 0.079 | |
| | letter_q | 0.015 | |
| | letter_r | 0.079 | |
| | letter_s | 0.031 | |
| | letter_t | 0.047 | |
| | letter_u | 0.047 | |
| | letter_v | 0.015 | |
| | letter_y | 0.015 | |
| **G4** | grouped_adjectives | 0.0 | |
| | grouped_adverbs | 0.066 | *dónde* |
| | grouped_nouns | 0.2 | *MARTY, regreso, futuro* |
| | grouped_verbs | 0.133 | *tienes, venir* |
| | other | 0.6 | *(todas las restantes)* |

*Tab. 3.1:* Ejemplo de extracción de atributos. Se omiten los atributos de distribución de POS tags y proporción de letras con valor 0.

Cabe aclarar que los atributos que estudian *cantidades* han sido normalizados para independizar su valor con respecto al largo de la película o a la cantidad de diálogo en la misma.

### 3.3.2. Atributos de emociones

A continuación definimos los atributos que utilizaremos en la segunda parte del trabajo con respecto a las emociones presentes en las palabras de los subtítulos.

En este caso, la extracción de atributos se efectuará no sólo en el subtítulo completo, sino también en tres intervalos de tiempo de la película: comienzo, desarrollo y final. Estos intervalos los obtenemos utilizando la duración de los diálogos, siguiendo el esquema indicado en la figura 3.2.

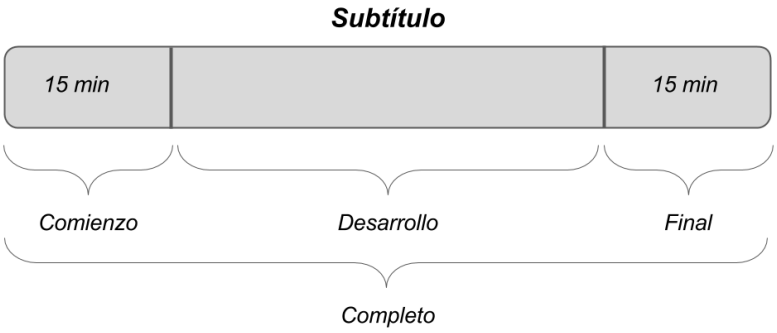


*Fig. 3.2:* Intervalos de los subtítulos obtenidos utilizando la duración de los diálogos.

Veamos, entonces, cuáles son los atributos de emociones que extraemos:

#### Polaridad (8 atributos)

Medida relacionada a la positividad y negatividad de un conjunto de palabras del subtítulo. Se calcula cómo la cantidad de palabras de cada polaridad dividida por la cantidad total de palabras anotadas con algún valor de polaridad.

#### Emociones (32 atributos)

Sentimiento que transmiten las palabras del subtítulo, según las categorías: alegría, anticipación, confianza, disgusto, enojo, miedo, sorpresa y tristeza. Se calcula cómo la cantidad de palabras de cada emoción dividida por la cantidad total de palabras anotadas con alguna de todas las emociones posibles.

Mostramos en la Tabla 3.2 un ejemplo de extracción de atributos de emociones.

| | |
|---|---|
| **Diálogo** | 00:01:20 –>00:01:22<br>Cuando ese vehículo llegue a las 88 millas por hora, te asombrarás.<br><br>00:16:23 –>00:16:25<br>¿Construyó la máquina de tiempo con el DeLorean?<br><br>00:45:23 –>00:47:25<br>Ustedes no están preparados para esta música. ¡Pero a sus hijos les encantará! |

| Grupo | Atributo | Valor | Evidencia |
|---|---|---|---|
| **Polaridad** | complete_negative | 0.0 | |
| | complete_positive | 1.0 | *asombrarás, música, encantará* |
| | beginning_negative | 0.0 | |
| | beginning_positive | 1.0 | *asombrarás* |
| | ending_negative | 0.0 | |
| | ending_positive | 1.0 | *música, encantará* |
| | mid_negative | 0.0 | |
| | mid_positive | 0.0 | |
| **Emociones** | complete_anger | 0.0 | |
| | complete_anticipation | 0.22 | *tiempo, encantará* |
| | complete_disgust | 0.0 | |
| | complete_fear | 0.0 | |
| | complete_joy | 0.33 | *asombrarás, música, encantará* |
| | complete_sadness | 0.11 | *música* |
| | complete_surprise | 0.22 | *asombrarás, encantarás* |
| | complete_trust | 0.11 | *máquina* |
| | beginning_anger | 0.0 | |
| | beginning_anticipation | 0.0 | |
| | beginning_disgust | 0.0 | |
| | beginning_fear | 0.0 | |
| | beginning_joy | 0.5 | |
| | beginning_sadness | 0.0 | |
| | beginning_surprise | 0.5 | |
| | beginning_trust | 0.0 | |
| | ending_anger | 0.0 | |
| | ending_anticipation | 0.2 | *encantará* |
| | ending_disgust | 0.0 | |
| | ending_fear | 0.0 | |
| | ending_joy | 0.4 | *música, encantará* |
| | ending_sadness | 0.2 | *música* |
| | ending_surprise | 0.2 | *asombrarás, encantará* |
| | ending_trust | 0.0 | |
| | mid_anger | 0.0 | |
| | mid_anticipation | 0.5 | *tiempo* |
| | mid_disgust | 0.0 | |
| | mid_fear | 0.0 | |
| | mid_joy | 0.0 | |
| | mid_sadness | 0.0 | |
| | mid_surprise | 0.0 | |
| | mid_trust | 0.5 | *máquina* |

*Tab. 3.2:* Ejemplo de extracción de atributos de emociones.

## 3.4. Construcción del clasificador

Llegado este punto, nos preguntamos si graficando la representación en atributos de los datos, podemos encontrar la manera de separar los géneros fácilmente. Para eso, graficamos histogramas de distribución para cada atributo de manera independiente. En ningún caso se observaron diferencias claras para distinguirlos, por lo que continuamos con la construcción del clasificador automático.

El próximo paso, entonces, consistió en construir el clasificador. Como mencionamos previamente, será utilizando el método de Random Forest. Además, decidimos llevar a cabo la clasificación sobre un género por vez, con el objetivo de discernir entre las instancias de ese género contra los restantes. Definiremos la clase positiva como el género en cuestión, y la negativa como la agrupación de los otros tres géneros.

Como gran parte de los métodos de aprendizaje automático, *Random Forest* posee parámetros que es necesario explorar y ajustar para mejorar el rendimiento en la tarea específica. A continuación intentamos encontrar la combinación de parámetros optimal para el desempeño de nuestro clasificador. Luego, una vez fija la configuración del clasificador, procederemos a realizar pruebas de confianza sobre el modelo construido de manera de tener mayor certeza sobre cuál es la significancia de las conclusiones y resultados que obtendremos.

### 3.4.1. Elección de parámetros del clasificador

Para encontrar la combinación de valores optimal de los parámetros, realizamos la exploración a través de la técnica *Grid Search*. Este método permite recorrer una cantidad finita de combinaciones de valores para distintos parámetros que se intentan fijar. Luego, según una función objetivo, se encuentran los valores que producen el óptimo local.

El funcionamiento de la técnica es el siguiente: se selecciona un conjunto finito de valores para cada parámetro y se generan todas las combinaciones posibles entre cada uno de ellos. De esta forma, todas las combinaciones son exploradas una vez, sin utilizar información sobre corridas anteriores para decidir cuál es la siguiente configuración a evaluar. Para este trabajo, utilizamos el valor por defecto de la mayor parte de los parámetros y evaluamos aquellos que consideramos que podrían tener mayor incidencia sobre los resultados.[6]

Los parámetros a explorar fueron:

- La cantidad de árboles que se utilizarán en el modelo (parámetro "n_estimators" en la implementación). **Valores considerados**: 10, 20, 50, 100, 500, 1000.
- El peso asociado a cada clase que utilizará el clasificador para compensar problemas de desbalanceo (parámetro "class_weight" en la implementación). **Valores considerados**: None, "balanced", "balanced_subsample".

[6] Para una descripción más precisa de los parámetros ver: `http://scikit-learn.org/0.18/modules/generated/sklearn.ensemble.RandomForestClassifier.html` y la sección 2.2.2.

La opción "balanced" utiliza la frecuencia de etiquetas de cada clase para ajustar automáticamente los pesos de manera inversamente proporcional a este número.

La opción "balanced_subsample" es equivalente a la anterior pero en este caso las frecuencias son calculadas directamente sobre las instancias que se utilizan para entrenar cada árbol particular y no las instancias originales.

- El número de atributos a considerar en el momento de buscar el mejor corte para cada nodo (parámetro "max_features" en la implementación). **Valores considerados**: "sqrt" (raíz cuadrada del total), "log2" (logaritmo en base 2 del total), 0.5 (la mitad del total), None (todos los atributos).

Una vez decididos los parámetros a estudiar, el experimento consiste en entrenar y evaluar nuestros modelos utilizando la técnica de validación cruzada para cada una de las distintas configuraciones posibles de los parámetros y luego, comparar el desempeño obtenido en cada combinación.

Para esta prueba utilizaremos los atributos gramaticales G1, G2 y G3; y la métrica F1 cómo medida de comparación entre cada configuración, sobre las películas del género "Terror" que es el que tiene la menor cantidad de subtítulos. Elegimos este género en donde creemos que más puede incidir la configuración elegida debido al desbalance que existe entre esta clase y las demás. El *desbalance de clases* es un problema común a la hora de armar un clasificador, que puede deteriorar drásticamente el desempeño del sistema (Kotsiantis, Kanellopoulos y Pintelas 2006). Ocurre cuando contamos con una cantidad de instancias significativamente inferior de una clase en comparación a otra. En este escenario, los algoritmos de clasificación se entrenarán más en la clase mayoritaria, presentando así una tendencia a favorecer esta clase por sobre las demás.

En la figura 3.3 pueden observarse tres paneles, uno por cada configuración del parámetro "class_weight" posible. Dentro de cada uno de ellos, variamos el número de árboles y mostramos en el eje Y el valor de la métrica F1 obtenido. A su vez, cada curva representa una configuración distinta para la cantidad de atributos a considerar en el corte de cada nodo durante la construcción de los árboles.

Observamos que, en todos los casos, el clasificador consigue resultados con valores de la métrica F1 menores a lo esperado, cercanos a 0,5. Analizando estos resultados, encontramos que la clasificación obtuvo valores relativamente altos para la métrica de *precision*, pero extremadamente bajos (cercanos a cero) para *recall*. Con el objetivo de estudiar en mayor profundidad a qué se deben los bajos niveles de desempeño obtenidos, analizamos en detalle la confianza arrojada por el clasificador para cada instancia evaluada. Para ello, construimos un histograma que puede verse en la figura 3.4 en donde mostramos la probabilidad emitida por el clasificador para cada instancia separado por el género real del subtítulo. En este tipo de figuras se espera ver una amplia separación entre las distribuciones de manera que la clasificación sólo consista en fijar un umbral de distinción adecuado.

Podemos observar que en casi la totalidad de los casos, el clasificador predice en contra del género "Terror". Entendemos que el clasificador tiene un sesgo producto del desbalance de clases en los datos, ya que contamos con aproximadamente 1800 instancias para la clase positiva y aproximadamente 36000 instancias para la clase negativa.

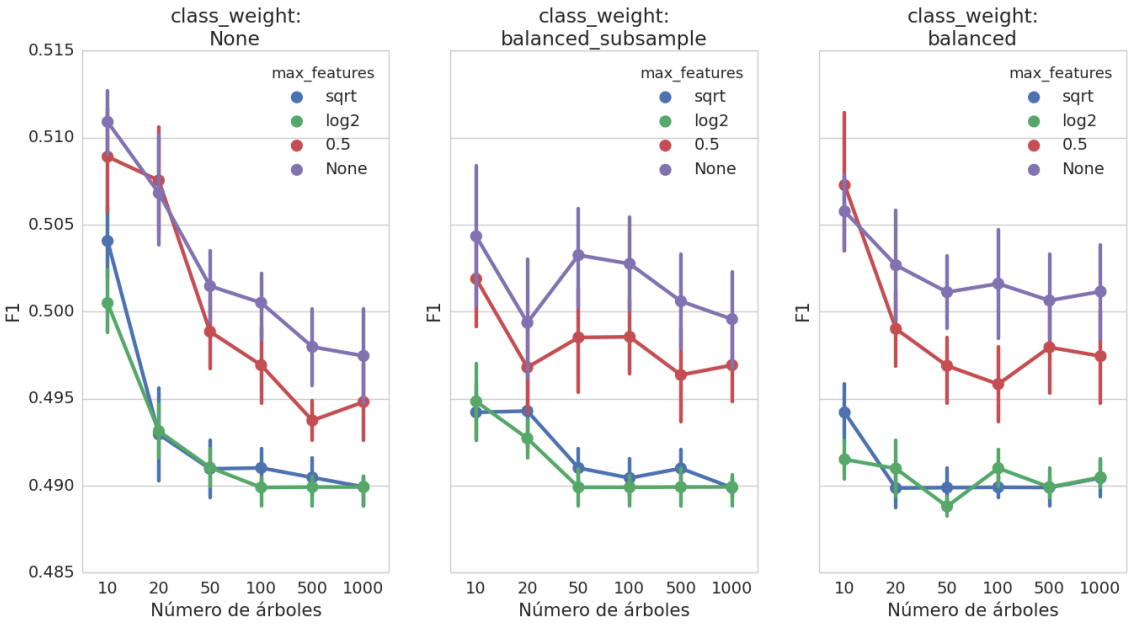


*Fig. 3.3:* Grid Search de parámetros del clasificador Random Forest, para las etiquetas "Terror" y "no-Terror", variando los parámetros *class_weight*, *max_features* y *n_estimators*. Estos resultados fueron obtenidos con un método de validación cruzada de 10 folds. Las barras verticales representan el desvío estandar entre los distintos folds.

Se presentan, entonces, al menos dos caminos de acción posibles para mejorar el desempeño del clasificador. Como primera opción, se podría hacer un análisis del umbral de decisión para buscar un punto más favorable y de esta manera tener mayor cantidad de aciertos. Por otra parte, se podría trabajar en balancear la cantidad de instancias de cada género qué utilizamos para entrenar el clasificador.

En el marco de nuestro trabajo, entendemos que balancear es la mejor decisión ya que estamos interesados en estudiar de qué manera impactan los atributos en la predicción del género, y no en buscar un clasificador que maximice la cantidad de predicciones correctas en una muestra realista de la población en dónde creemos que existe dicho desbalance.

Una técnica posible para balancear las clases es la del *subsampleo*. Consiste en eliminar instancias elegidas al azar de la clase mayoritaria, para quedarnos con la misma cantidad de datos en cada clase, y que el clasificador aprenda ambas en igual medida. La principal desventaja de este enfoque, es que se podría estar descartando información relevante para la clasificación pero veremos más adelante que tenemos información suficiente para realizar la tarea planteada de todas maneras.

Reproducimos, a continuación, el grid search subsampleando la cantidad de instancias de entrenamiento de cada género. También fijamos "class_weight" con valor *None*, es decir, todas las clases tienen el mismo peso relativo, porque estamos utilizando la misma cantidad de instancias para cada clase y este parámetro no tiene relevancia cuando las clases están balanceadas. Para subsamplear las instancias utilizamos la fórmula de la ecuación 3.1. De esta manera, la cantidad de instancias de la clase positiva (Terror) y la clase negativa ($\frac{1}{3}$ Acción, $\frac{1}{3}$ Comedia, $\frac{1}{3}$ Drama) es equivalente, con $\sim$ 1800 instancias de cada clase. Adelantándonos a la sección 3.4.2, veremos que esta cantidad de instancias supera

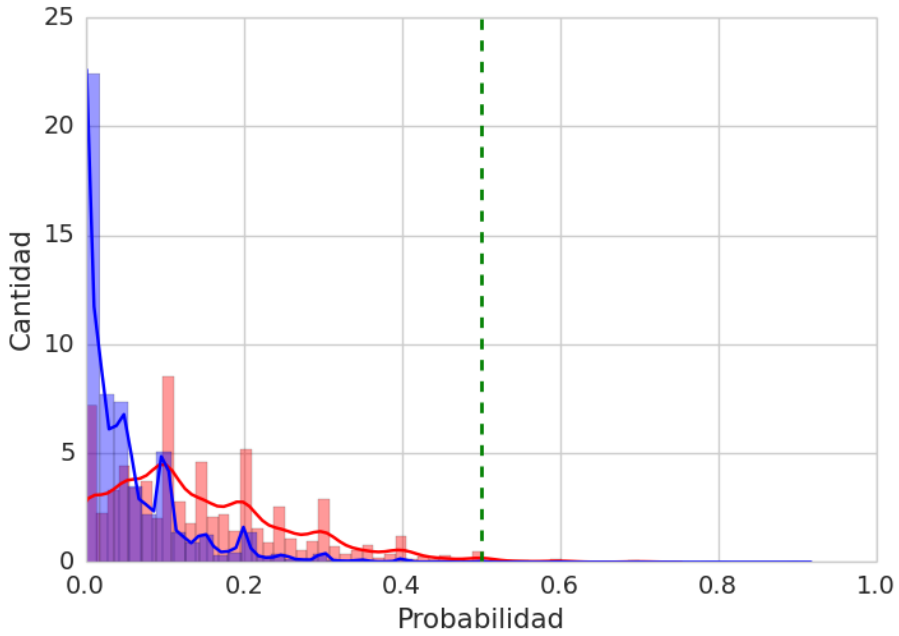


*Fig. 3.4:* Distribuciones de probabilidad de clasificación en la clase "Terror" para las etiquetas "Terror" y "no-Terror". El eje X contiene los valores de dichas probabilidades, y el eje Y la cantidad de casos en los que se obtuvo cada una de ellas. La curva roja es generada por las instancias que verdaderamente pertenecen al género "Terror" y la curva azul por las que no pertenecen a dicho género. Por otro lado, la línea vertical verde punteada representa el umbral por defecto utilizado para decidir la clase de una instancia: si la probabilidad es menor a 0.5, el clasificador decide en contra de la clase "Terror"; en caso contrario decide a favor de ella. Estos resultados fueron obtenidos con un método de validación cruzada de 10 folds, con *max_features=sqrt*, *class_weight=balanced* y *n_estimators=100*

la cantidad mínima obtenida mediante las curvas de aprendizaje, lo que nos indica que podemos utilizar el subsampleo descrito de manera confiable.

$$\begin{aligned} \text{subtítulosSampleados} = {} & \text{tomarAlAzar}(\text{cantidad} * 3, \text{subtítulosGéneroPositivo}) \\ & \cup\ \text{tomarAlAzar}(\text{cantidad}, \text{subtítulosGéneroNegativo1}) \\ & \cup\ \text{tomarAlAzar}(\text{cantidad}, \text{subtítulosGéneroNegativo2}) \\ & \cup\ \text{tomarAlAzar}(\text{cantidad}, \text{subtítulosGéneroNegativo3}) \end{aligned} \tag{3.1}$$

$$\begin{aligned} \text{cantidad} = \mathit{min} \Big\{ & \lfloor \frac{|\text{subtítulosGéneroPositivo}|}{3} \rfloor, \\ & |\text{subtítulosGéneroNegativo1}|, \\ & |\text{subtítulosGéneroNegativo2}|, \\ & |\text{subtítulosGéneroNegativo3}| \Big\} \end{aligned}$$

Podemos observar en la figura 3.5 que los resultados llegan en este caso a un valor de F1 cercano a 0.8. Notamos también que hay poca diferencia entre cada curva, y que la métrica no varía demasiado a partir de los 100 árboles (con excepción de una curva).

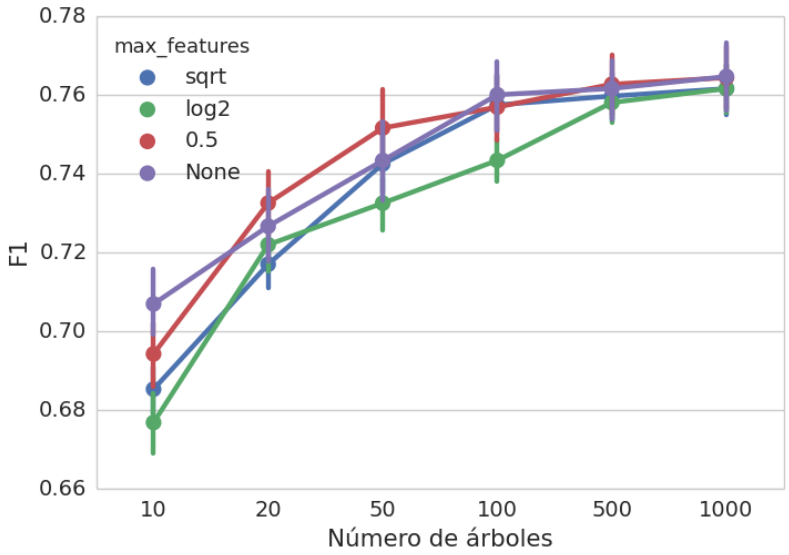


*Fig. 3.5:* Grid Search de parámetros del clasificador Random Forest, para las etiquetas "Terror" y "no-Terror", variando los parámetros max_features y n_estimators, con clases subsampleadas. Estos resultados fueron obtenidos con un método de validación cruzada de 10 folds. Las barras verticales representan el desvío estándar entre los distintos folds.

Generamos, ahora, el gráfico de distribuciones a partir de la clasificación con subsampleo de los géneros en la figura 3.6. En esta figura puede apreciarse que, con los géneros subsampleados, el clasificador decide a favor y en contra de "Terror" casi en la misma proporción, con un alto grado de acierto en la elección de etiqueta para ambas clases.

De aquí en más, optamos por fijar 100 árboles (*n_estimators=100*) porque hasta dicha cantidad, el desempeño del clasificador varía considerablemente, pero después de este punto ya no varía demasiado y la velocidad de procesamiento empeora notoriamente. También utilizaremos *max_features=sqrt* (raíz cuadrada de la cantidad de atributos) porque con la configuración planteada no se observa diferencia significativa entre los distintos valores de este parámetro, y se ejecuta de manera relativamente veloz además de ser el valor por defecto de la implementación utilizada.[7]

Para resumir, elegimos la siguiente configuración de parámetros para utilizar a lo largo del trabajo:

- Subsamplear las clases para lograr un balance en los datos
- Utilizar 100 árboles (n_estimators=*100*)
- No introducir pesos para las distintas clases (weight_class=*None*)
- Utilizar la raíz cuadrada de la cantidad de atributos totales en cada corte del árbol (max_features=*sqrt*)

[7] `http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier`

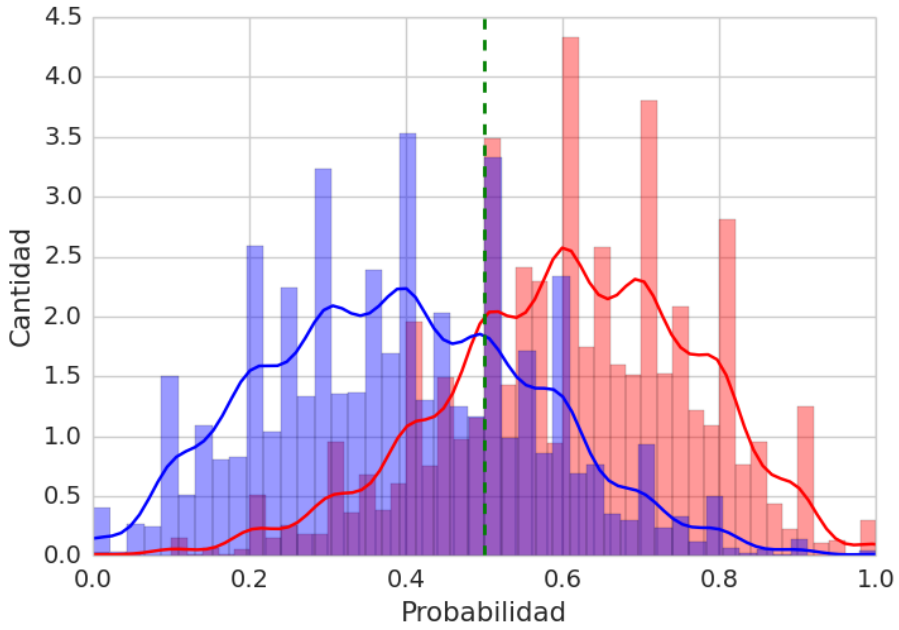


*Fig. 3.6:* Distribuciones de probabilidad de clasificación en la clase "Terror" para las etiquetas "Terror" y "no-Terror". Los subtítulos con probabilidad mayor al umbral (línea verde punteada) han sido clasificados como "Terror" y en el caso contrario como "no-Terror". Estos resultados fueron obtenidos con un método de validación cruzada de 10 folds, *max_feature=sqrt* y *n_estimators=100*, sobre las clases subsampleadas.

### 3.4.2. Confianza sobre el sistema

Una vez que hemos elegido los parámetros de nuestro clasificador, nos proponemos realizar una serie de pruebas para ganar confianza sobre el funcionamiento del sistema.

#### Curvas de aprendizaje

Comenzamos preguntándonos si contamos con instancias suficientes para realizar la clasificación exitosamente. Para responder este interrogante, graficamos las curvas de aprendizaje para cada uno de los géneros. Estas constituyen una forma gráfica de ver cómo evoluciona el desempeño de un clasificador al aumentar la cantidad de datos de entrenamiento. Entendemos que este experimento nos permitirá determinar si tenemos, en todos los casos, la cantidad de datos de entrenamiento suficientes para resolver la tarea de manera exitosa. Además, si el resultado es positivo, podremos observar a partir de qué cantidad se estabiliza el desempeño. En la figura 3.7 puede verse la evolución del resultado de la clasificación a medida que agregamos ejemplos de entrenamiento.

Observamos que el comportamiento de las curvas de aprendizaje de todos los géneros es muy similar, siguiendo una pendiente parecida, y alcanzando una estabilidad cerca de las 200 instancias. A partir de ahí, el valor de F1 no aumenta considerablemente.

A partir de esto, entendemos que conseguir más datos de entrenamiento no tendría

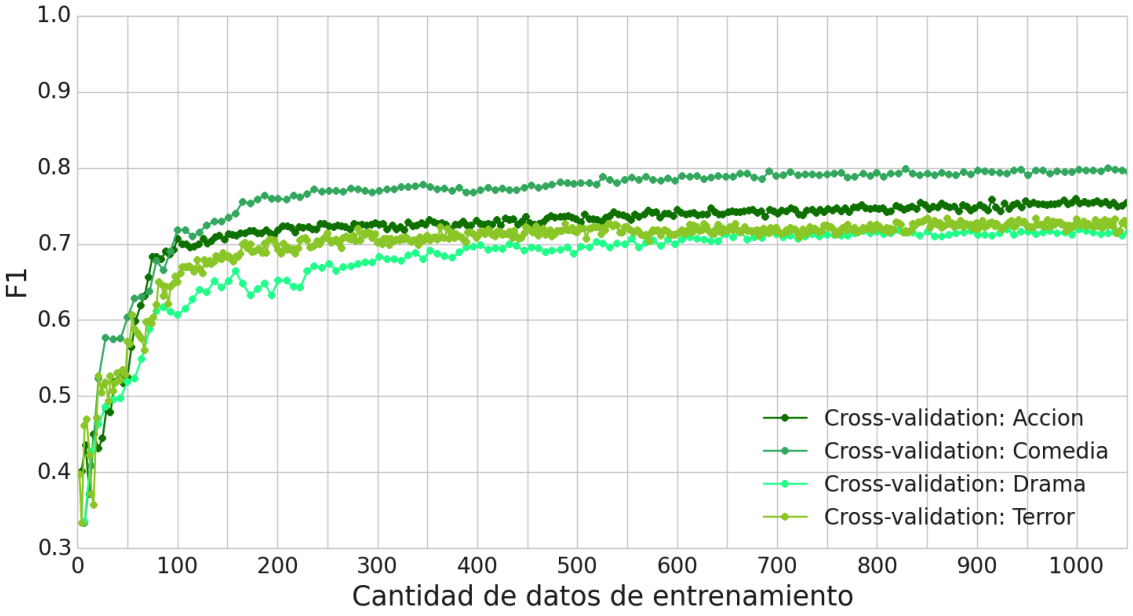


*Fig. 3.7:* Curvas de Aprendizaje del clasificador Random Forest, para las etiquetas "Acción", "Comedia", "Drama" y "Terror". El eje X es la cantidad de datos de entrenamiento. El eje Y el valor de F1 obtenido utilizando dicha cantidad. Cada una de las 4 curvas en distintos tonos de verde representan el resultado de cada evaluar clasificadores para cada uno de los géneros mediante la técnica de validación cruzada de 10 folds.

un impacto significativo en el desempeño del clasificador, por lo que descartamos ese camino. Además, reconocemos la cantidad de 300 instancias como el límite inferior para subsamplear de manera confiable.

### Test de permutación

Nos proponemos ahora, comprobar que el funcionamiento del sistema no es producto del azar sino que realmente los atributos utilizados contienen información relevante. Con este objetivo es que realizamos un *test de permutación* similar al que se explica en el trabajo de Ojala y Garriga (2010). Este método, tiene como objetivo estudiar si el clasificador realmente logra encontrar el vínculo entre los datos de entrenamiento y su clase, o simplemente obtiene sus resultados producto del azar. Para esto, se contrasta el desempeño de la clasificación sobre los datos originales, contra $N$ ejecuciones en donde las etiquetas de clases se desordenan aleatoriamente. Idealmente, veremos que el desempeño del sistema supera significativamente el de las permutaciones aleatorias.

En la figura 3.8 mostramos en verde la curva ROC del clasificador original, una línea punteada roja que representa la curva ROC de un sistema basado únicamente en el azar (sistema "random"), y múltiples (100) curvas azules que representan cada una, una ejecución del clasificador con una permutación aleatoria de las etiquetas originales. Podemos ver que todas las permutaciones se ubican alrededor de la línea de la clasificación random, y muy por debajo de la curva de la clasificación con el orden de etiquetas original. Este resultado indica que los resultados obtenidos son estadísticamente significativos con un

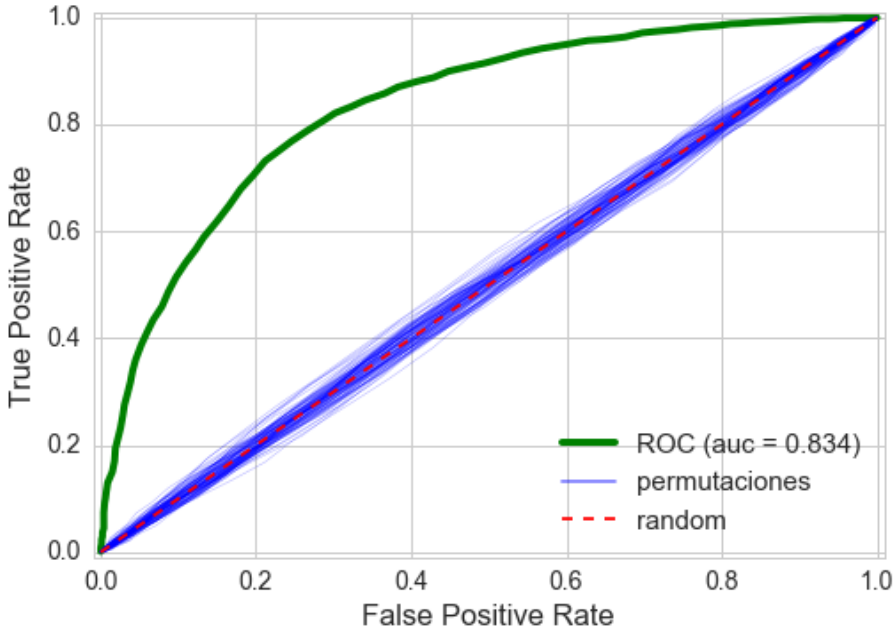


*Fig. 3.8:* Curvas ROC de la clasificación original contra 100 permutaciones aleatorias. Estos resultados fueron obtenidos con un método de validación cruzada de 10 folds.

p-valor menor a 0,01.

Otra forma de visualizar la significancia del resultado es analizando el área bajo la curva (AUC) de la clasificación. El AUC original es de 0,834. Analizaremos cómo es la diferencia en esta medida contra las clasificaciones realizadas sobre las permutaciones aleatorias de las etiquetas. Lo graficamos utilizando el gráfico de histogramas que puede observarse en la figura 3.9. En esta figura observamos cómo el área bajo la curva de las curvas ROC de las permutaciones aleatorias se centran en el valor 0.5 del histograma, muy por debajo del área de la clasificación original. De esta forma reafirmamos que las permutaciones se comportan de la manera esperada y que el desempeño de nuestro clasificador las supera ampliamente para este caso.

## 3.5. Resultados del clasificador utilizando atributos gramaticales

Los resultados y análisis obtenidos en todas las pruebas, nos otorgan cierta confianza sobre el funcionamiento del clasificador implementado. Es importante destacar que, aunque algunas de las pruebas exhibidas se refieren únicamente al género terror, notamos resultados similares en cuanto a la significancia de los resultados para todos los géneros restantes. A sabiendas de esto, exponemos el desempeño alcanzado por clasificadores construidos con las propiedades antes mencionadas para cada uno de los géneros estudiados en la Tabla 3.3.

Observamos que las clasificaciones de todos los géneros alcanzaron valores parecidos para cada una de las métricas, siendo “Comedia” el género que mejor desempeño consiguió,

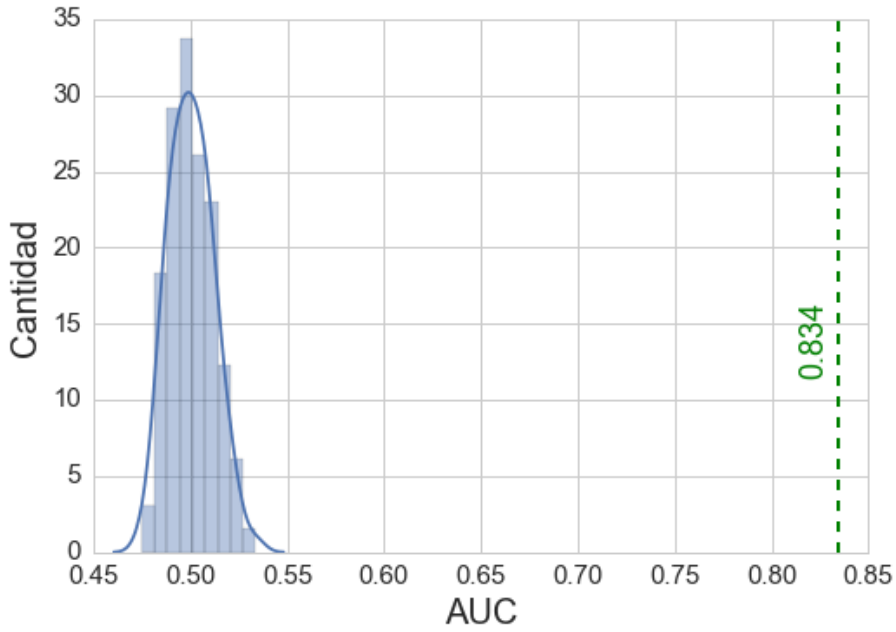


*Fig. 3.9:* Comparación de AUC (área bajo la curva) de la clasificación original contra 100 permutaciones aleatorias. Estos resultados fueron obtenidos con un método de validación cruzada de 10 folds, con *max_feature=sqrt* y *n_estimators=100*, sobre las clases subsampleadas.

| **Género** | **Accuracy** | **AUC** | **F1** |
|---|---|---|---|
| Acción | 0.776667 | 0.855410 | 0.776535 |
| Comedia | 0.813459 | 0.893871 | 0.813448 |
| Drama | 0.739878 | 0.815675 | 0.739789 |
| Terror | 0.733777 | 0.817374 | 0.733703 |

*Tab. 3.3:* Resultados de las diferentes métricas utilizando atributos gramaticales, obtenidas con un método de validación cruzada de 10 folds.

y "Terror" y "Drama" los peores.[8]

En resumen, podemos construir clasificadores que permiten discriminar el género de una película según atributos extraídos de sus subtítulos con un desempeño superior a 0,7 y en donde los resultados son significativamente superiores al azar. En el siguiente capítulo utilizaremos el clasificador construido para cada género de manera de estudiar el impacto de los distintos atributos.

[8] Las diferencias de decimales entre las pruebas de calibración y la tabla final son producto de utilizar distintas semillas para generar números aleatorios.

# 4. RESULTADOS

En esta sección presentamos un análisis completo de los experimentos realizados para determinar la relación que existe entre el género de una película y el aspecto gramatical y el contenido semántico de sus subtítulos.

Partimos de la hipótesis de que esta asociación existe, por lo que diseñamos dichos experimentos para buscar indicios que lo confirmen. Alimentar al clasificador automático con atributos de distintos tipos, permite ver cuáles de los aspectos capturados por cada atributo separan mejor los géneros. Es decir, que estos aspectos capturados por los atributos ganadores son aquellos que presentan mayor relación con el género en cuestión.

Entonces, comenzaremos estudiando qué atributos gramaticales explican de mejor manera las diferencias entre géneros y en qué medida permiten predecirlo. Luego, continuamos con una segunda etapa en dónde analizamos el aspecto semántico a través de las emociones que pueden ser detectadas en las palabras utilizadas en los subtítulos.

## 4.1. Modalidad de los experimentos

Para responder las preguntas que impulsaron los experimentos que presentaremos a continuación, utilizamos el siguiente esquema:

1. Planteo de hipótesis sobre cómo afectarán los distintos atributos a la clasificación en cada género
2. Construcción de un clasificador utilizando la técnica de Random Forest con los parámetros aprendidos anteriormente (ver sección 3.4.1)
3. Análisis de desempeño basado en distintas selecciones de atributos
4. Estudio de la importancia relativa de cada atributo en la clasificación
5. Comparación de atributos particulares y sus diferencias entre géneros

## 4.2. Primer experimento: atributos gramaticales

En este experimento nos proponemos estudiar la incidencia de los distintos atributos gramaticales sobre la clasificación de géneros. Como primer paso ejecutaremos el clasificador utilizando cada grupo de atributos para analizar cuál de ellos consigue un mayor desempeño. Luego, estudiaremos qué atributos en particular resultaron útiles en la clasificación para cada uno de estos grupos e intentaremos explicar los motivos detrás de los resultados obtenidos.

### 4.2.1. Hipótesis sobre los atributos gramaticales

Previo a ejecutar los experimentos, comenzamos desarrollando una serie de hipótesis sobre los atributos que explicamos en la sección 3.3. Estas hipótesis clarifican qué esperamos ver utilizando dichos atributos y de qué manera y en qué medida creemos que van a afectar a la clasificación de los distintos géneros.

1. **Distribución de POS tags**

   *a*) **Proporción de cada POS tag:** Esperamos que sea el atributo más informativo. Suponemos que los géneros "Drama" y "Comedia" presentarán una mayor proporción de adjetivos y nombres propios originado por un lenguaje más descriptivo y con mayor cantidad de personajes.

2. **Complejidad del lenguaje:** suponemos que los géneros de "Comedia" y "Drama" suelen tener diálogos más largos y complejos, por lo que este grupo de atributos podría resultar de mucha utilidad para distinguirlos.

   *a*) **Densidad de habla:** Esperamos ver que "Comedia" y "Drama" tengan mayor cantidad de palabras en el mismo tiempo que "Acción" y "Terror", ya que suponemos que tienen diálogos más largos en tiempos de habla parecido.

   *b*) **Palabras por oración:** Creemos que "Comedia" y "Drama" tendrá mayor cantidad de palabras por oración que "Acción" y "Terror", por un razonamiento análogo al del atributo anterior.

   *c*) **Longitud media de palabras:** Esperamos ver que "Comedia" y "Drama" tengan palabras con mayor longitud que "Acción" y "Terror", debido a la utilización de vocabulario más complejo.

   *d*) **Proporción de palabras en mayúscula:** Creemos que los géneros "Acción" y "Terror" tendrán mayor proporción de palabras en mayúscula, relacionadas a gritos y frases imperativas.

   *e*) **Proporción de verbos dinámicos y estáticos:** Suponemos que en el género "Acción" existirá mayor proporción de verbos dinámicos (de acción) y en "Comedia" y "Drama" de verbos estáticos (abstractos).

3. **Frecuencia de letras**

   *a*) **Frecuencia de símbolos de interrogación y admiración finales ("?" y "!"):** creemos que los géneros "Acción" y "Terror" tendrán mayor proporción de signos de admiración, relacionadas a gritos y frases imperativas; mientras que "Drama", tendrá mayor proporción de signos de pregunta originado por una mayor cantidad de preguntas en sus diálogos.

   *b*) **Frecuencia de letras del abecedario (a,b,...,z):** Lo agregamos a modo exploratorio. No creemos que vayan a hacer un aporte significativo, pero tal vez el algoritmo logre encontrar alguna relación inesperada con el género. Por ejemplo, es posible que aporte información por la aparición de palabras específicas en menor o mayor proporción en algún género en particular que en el resto.

4. **Distribución de POS tags agrupados**

- **Proporción de POS tag agrupados:** Esperamos que también con esta simplificación de las etiquetas sigan cumpliéndose las hipótesis del punto $1a$.

### 4.2.2. Comparación de grupos de atributos

Analizaremos ahora cuáles son los grupos de atributos gramaticales que tienen mayor incidencia en la clasificación de géneros. Dichos grupos (explicados en la sección 3.3) son:

- **G1:** Distribución de POS tags
- **G2:** Complejidad del lenguaje
- **G3:** Frecuencia de letras
- **G4:** Distribución de POS tags agrupados
- **G5:** G1, G2 y G3 en conjunto

Para poder comparar la información que aporta cada grupo de atributos en cuanto al poder predictivo del género de las películas, construimos clasificadores que sólo utilizan un cierto grupo de atributos por vez, para así lograr medir las diferencias en el desempeño obtenido entre uno y otro.

Cada punto de la figura 4.1 muestra el resultado de construir y evaluar un clasificador utilizando el grupo correspondiente de atributos.

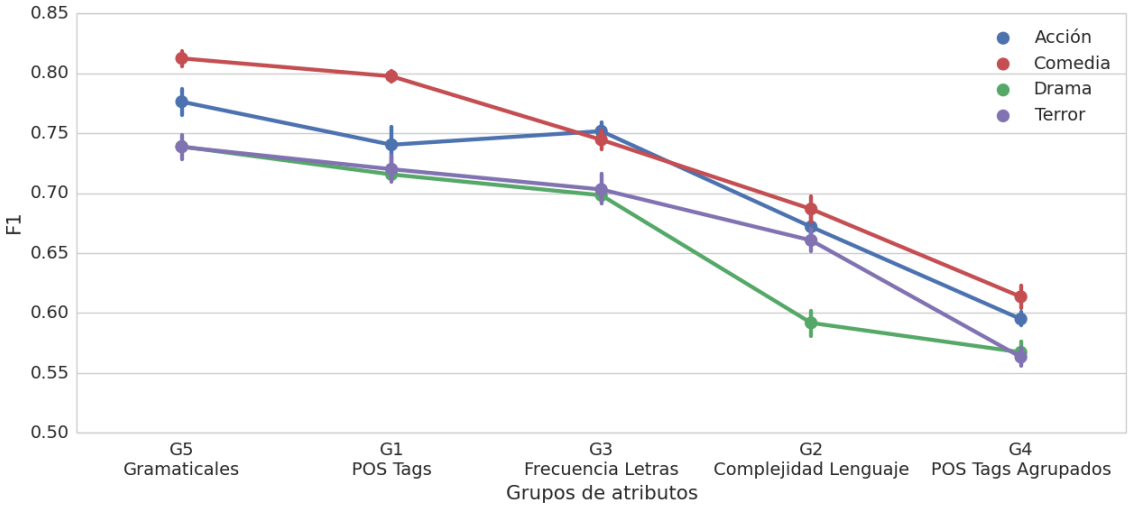


*Fig. 4.1:* Comparación de desempeño de atributos gramaticales para cada género, obtenido con un método de validación cruzada de 10 folds y medido utilizando la métrica F1. Las barras verticales representan el desvío estándar entre los distintos folds.

Puede verse que el grupo de atributos de POS tags resulta ser el de mayor poder predictivo dentro de todos los gramaticales tal como habíamos predicho en la sección 4.2.1. Sin embargo, no esperábamos encontrar que los atributos de frecuencia de letras tuvieran una importancia tan alta en comparación a los de complejidad del lenguaje.

Es importante destacar también la coherencia entre los distintos géneros. Es decir, que el poder discriminativo que aporta cada grupo de atributos se comporta de manera muy similar para todos los géneros.

Como conclusión del experimento vemos que, aunque el grupo de POS tags es el que permite clasificar de mejor manera, todos los grupos contienen información relevante para la clasificación. Además, entendemos que la agrupación de POS tags simplifica demasiado los atributos, lo que resulta en un desempeño mucho más pobre del clasificador. Es por estos motivos que a partir de este punto profundizaremos el estudio sobre los clasificadores construidos utilizando el grupo G5 (todos los atributos gramaticales, salvo los POS tags agrupados) y veremos cómo afecta cada atributo de manera particular.

### 4.2.3. Importancia de atributos

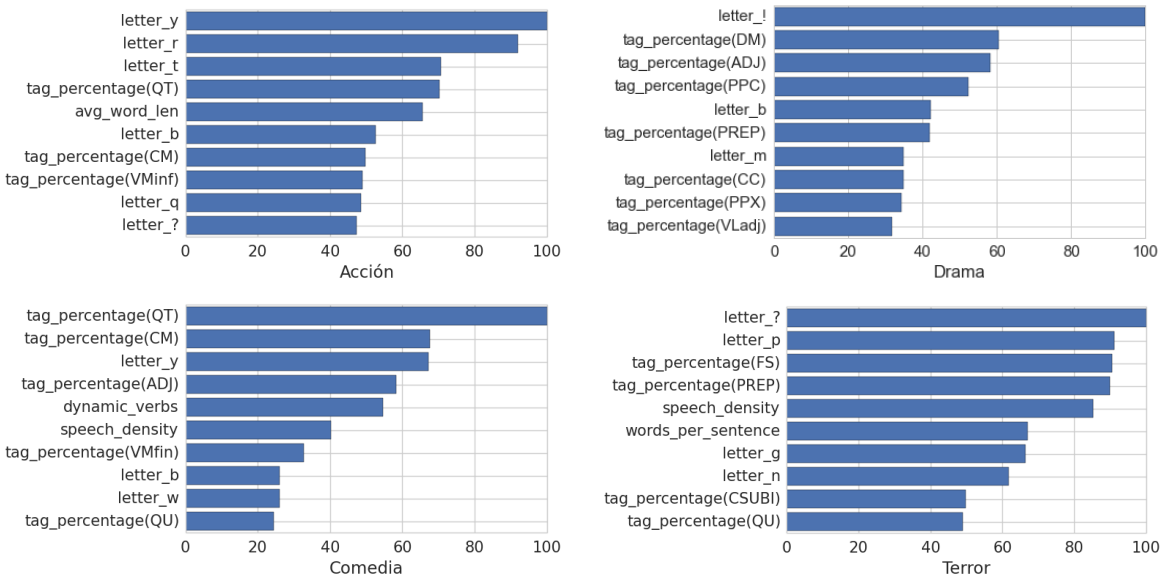


*Fig. 4.2:* Importancia relativa de los 10 atributos gramaticales más importantes para cada género, arrojados por el clasificador Random Forest y obtenidos con un método de validación cruzada de 10 folds. Para mayor información, pueden verse los Apéndices Etiquetas de POS tag en español y Traducción de nombres de atributos.

Siguiendo con el análisis acerca de qué atributos resultaron más importantes para realizar la tarea de clasificación, reportaremos ahora cuáles fueron los atributos individuales con mayor relevancia para cada uno de los géneros. Nos basaremos en la importancia de atributos que provee la implementación de Random Forest utilizada (detallada en la sección 2.2.2), y graficaremos la importancia de los atributos relativa al primero del ranking de cada género, en la figura 4.2.

Observamos cómo el ranking de atributos para cada género coincide fuertemente con el gráfico de comparación de atributos de la figura 4.1: en "Acción" predominan los atributos de distribución de letras mientras que en todos los géneros restantes los atributos de POS tags son los que aparecen con mayor frecuencia. Sólo en el caso de "Comedia" el atributo con mayor importancia pertenece al grupo de distribución de POS tags.

### 4.2.4. Distribución de atributos

Para comprender los motivos que originan la importancia de los atributos, decidimos estudiar la distribución de los valores de cada atributo comparando los subtítulos de cada género contra el resto. Este análisis no necesariamente explica el posicionamiento de los atributos dentro del ranking en su totalidad. Por ejemplo, no estudia las relaciones entre los distintos atributos, lo que puede ser un factor determinante. Sin embargo, provee información valiosa para comprender las tendencias que dan lugar a elecciones del clasificador. A su vez, contrastaremos estos resultados contra las hipótesis planteadas en la sección 4.2.1.

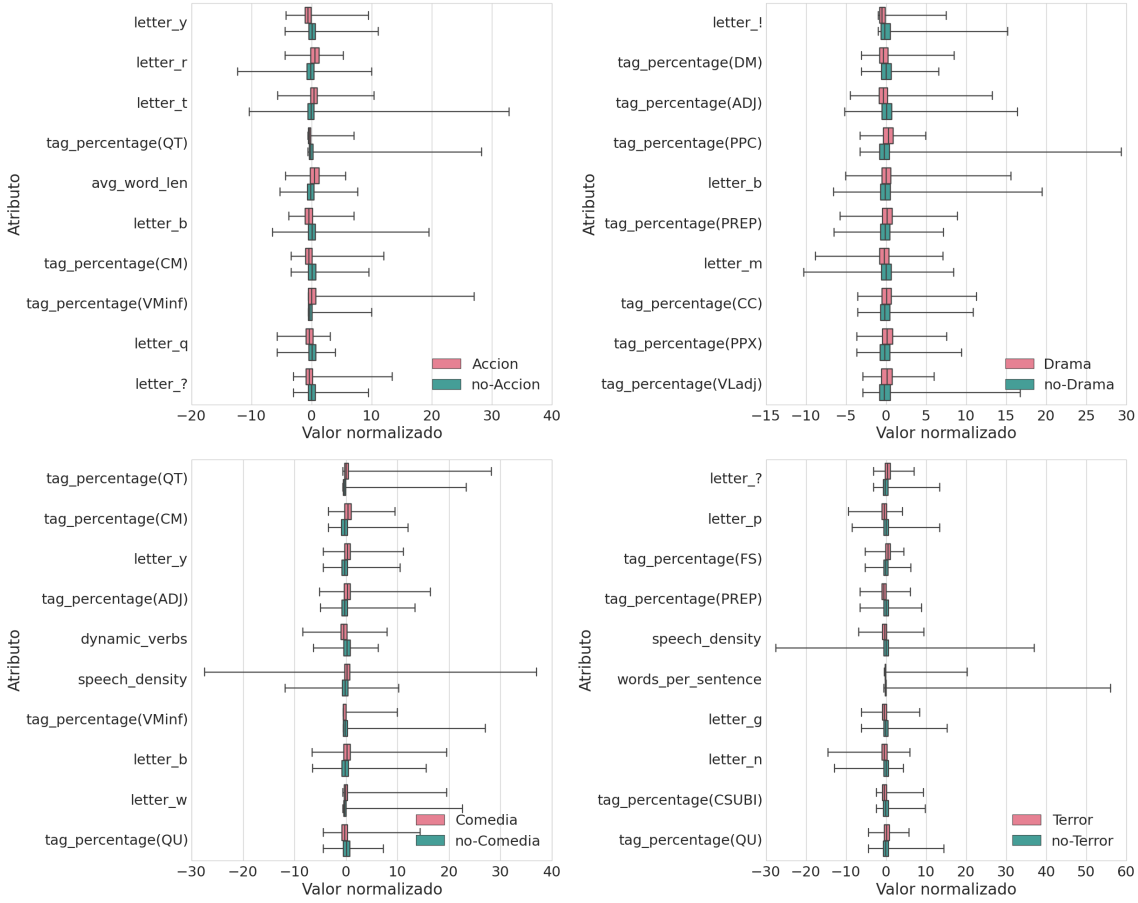


*Fig. 4.3:* Distribución de los 10 atributos gramaticales más importantes para cada género, estandarizados.

En la figura 4.3, mostramos las distribuciones para los primeros diez atributos más importantes en la clasificación a través de gráficos de caja. Para esta visualización, decidimos estandarizar los valores de cada atributo, lo que consiste en restarles la media y dividirlos por el desvío estándar. De esta forma, todos los atributos quedarán en la misma escala y sin unidad; lo que resultará en que sean más fáciles de comparar entre sí.

Acerca de los atributos de POS tags, vemos que algunos de ellos están relacionados a los *signos de puntuación*. Por ejemplo, hay una mayor presencia tanto de comillas (*tag_percentage(QT)*) como de comas (*tag_percentage(CM)*) en las películas de comedia

que en el resto, y menor en el caso de las películas de acción. La aparición de comillas puede explicarse como la cantidad de citas presentes en los subtítulos de cada uno de estos géneros. La cantidad de comas, por otro lado, puede entenderse como parte de la noción de complejidad del lenguaje, para la que se cumplen las hipótesis planteadas acerca de que la complejidad del lenguaje de las comedias supera a la de las películas de acción.

Otro atributo de POS tag interesante para analizar es el de *cantidad de adjetivos* (tag_percentage(ADJ)). Vemos que las comedias presentan mayor cantidad de adjetivos que el resto, como habíamos planteado en las hipótesis. Sin embargo, la cantidad en los dramas es menor a la de los géneros restantes.

Con respecto a la cantidad de *verbos dinámicos*, observamos que no aparece entre los más importantes para clasificar películas de acción, pero sí surge para las comedias dónde presentan una cantidad menor al resto.

Sobre los atributos de complejidad del lenguaje, observamos que el atributo de *densidad de habla* se comporta de la manera que esperábamos. Esto es, las comedias tienen mayor cantidad de palabras por minuto de habla que los géneros restantes, mientras quede los subtítulos de terror tienen menor cantidad. Estos últimos tienen también menores valores de *palabras por oración* que los demás.

En cuanto al atributo de *longitud media de palabras*, los resultados fueron contrarios a nuestras hipótesis, siendo los subtítulos de acción los que presentan valores mayores a los restantes.

Con respecto a los atributos de cantidad de ocurrencias de las letras, vemos que los dramas presentan menor cantidad de *símbolos de admiración* que los demás. Esto podría significar que hay menor presencia de gritos y frases en modo imperativo en ellos.

Los *signos de pregunta*, que dan noción de la cantidad de dudas o preguntas presentes en los subtítulos, ocurren en baja proporción en acción, tal como habíamos predicho. En los de terror, en cambio, presentan una cantidad superior a los demás.

Un fenómeno no esperado en nuestras hipótesis fue el de la gran importancia que asigna el clasificador a la frecuencia de ciertas letras para algunos géneros. A continuación intentamos responder cuál fue el motivo para este suceso.

### 4.2.5. Estudio sobre los atributos de cantidad de ocurrencias de letras

La clasificación de algunos géneros presenta alta importancia en la frecuencia de letras. Partimos de la hipótesis de que las diferencias entre géneros de las frecuencias de las letras se debe a la presencia o ausencia de palabras muy frecuentes asociadas a cada género en particular. Para intentar explicar los casos con estas características, y en particular aquellos en que la cantidad de apariciones de dicha letra en el género es mayor que en los géneros restantes, planteamos el estudio descrito a continuación:

1. Elegimos la letra "*r*" por ser una de las letras útiles para que el clasificador distinga el género *acción*. Dicha letra tiene mayor presencia en este género que en los demás.

2. Elegimos $N = 10$, 50, 100, 200 y 500.

3. Tomamos el conjunto $C_{acción}$: las $N$ palabras con mayor cantidad de apariciones en el género *acción* que contienen la letra "$r$".

4. Tomamos el conjunto $C_{noAcción}$: las $N$ palabras con mayor cantidad de apariciones en todos los géneros distintos a *acción* que contienen la letra "$r$".

5. Contrastamos las diferencias entre las palabras de $C_{acción}$ y $C_{noAcción}$.

En este punto, esperábamos ver que, al elegir las palabras con mayor cantidad de apariciones en el género acción con la letra "r", surjan entre las primeras posiciones algunas muy relacionadas con este tipo de películas, que no aparezcan tan frecuentemente en los demás.

Sin embargo, las palabras que surgen tanto en $C_{acción}$ cómo en $C_{noAcción}$, son a grandes rasgos las mismas. Además, y contrariamente a lo que esperábamos, las palabras que aparecen en $C_{acción}$ tienen *menor* cantidad de ocurrencias en acción que las de $C_{noAcción}$ en su complemento. Incluso dividiendo estas cantidades por la cantidad de palabras totales del género, o géneros, sobre el que se evalúan se mantienen las mismas tendencias.

Repetimos el experimento con varias combinaciones de letras y géneros importantes para la clasificación, obteniendo resultados análogos al anterior. Es decir, descartamos la hipótesis de que la importancia de las letras se deba a presencia o ausencia de palabras asociadas a cada género, sin haber encontrado razones intuitivas por las cuáles la aparición de ciertas letras pueda estar relacionada con el género de la película.

## 4.3. Segundo Experimento: inclusión de emociones

Hasta el momento, hemos desarrollado todos nuestros análisis utilizando únicamente atributos gramaticales (G5). A partir de este punto, contrastaremos los resultados anteriores contra nuevos resultados en donde incorporaremos atributos relacionados a las emociones presentes en cada subtítulo.

### 4.3.1. Hipótesis sobre los atributos de emociones

Naturalmente asociamos el género de las películas con las emociones y la polaridad que ellas nos transmiten. Esperamos verlo reflejado en los atributos explicados en la sección 3.3. Por ejemplo, que para el género "Comedia" la polaridad positiva y la emoción de alegría sean dominantes, mientras que la polaridad negativa y la emoción de miedo lo sean para "Terror". Suponemos que estos tipos de atributos tendrán superior desempeño que los gramaticales debido a que parecen fácilmente asociables a cada género.

### 4.3.2. Comparación con el modelo anterior

Comenzamos analizando los valores de las métricas obtenidas al ejecutar la clasificación utilizando los atributos gramaticales junto con los de emociones. En la Tabla 4.1 podemos ver el impacto que estos últimos tienen en el desempeño del clasificador.

| Atributos | Género | Accuracy | AUC | F1 |
|---|---|---|---|---|
| **Sin emociones** | Acción | 0.776667 | 0.855410 | 0.776535 |
| | Comedia | 0.813459 | 0.893871 | 0.813448 |
| | Drama | 0.739878 | 0.815675 | 0.739789 |
| | Terror | 0.733777 | 0.817374 | 0.733703 |
| **Con emociones** | Acción | 0.798413 | 0.880575 | 0.798362 |
| | Comedia | 0.858292 | 0.933980 | 0.857957 |
| | Drama | 0.789795 | 0.869827 | 0.789648 |
| | Terror | 0.776761 | 0.861266 | 0.776723 |

*Tab. 4.1:* Resultados de las diferentes métricas utilizando atributos gramaticales sin y con emociones, obtenidas con un método de validación cruzada de 10 folds.

Notamos un incremento en los valores de todas las métricas para cada uno de los géneros de entre 0.02 y 0.05 puntos, siendo "Acción" el género cuyo desempeño se ve menos beneficiado tras la adición de los atributos de emociones, y "Drama" el que mayor impacto recibió.

### 4.3.3. Comparación de atributos

Hemos observado que los atributos de emociones aportan información relevante a nuestra tarea de clasificación. Veamos nuevamente la diferencia de desempeño, medido utilizando la métrica F1, que existe para cada grupo de atributos, pero agregando esta vez los atributos de emociones a la comparación 4.4.

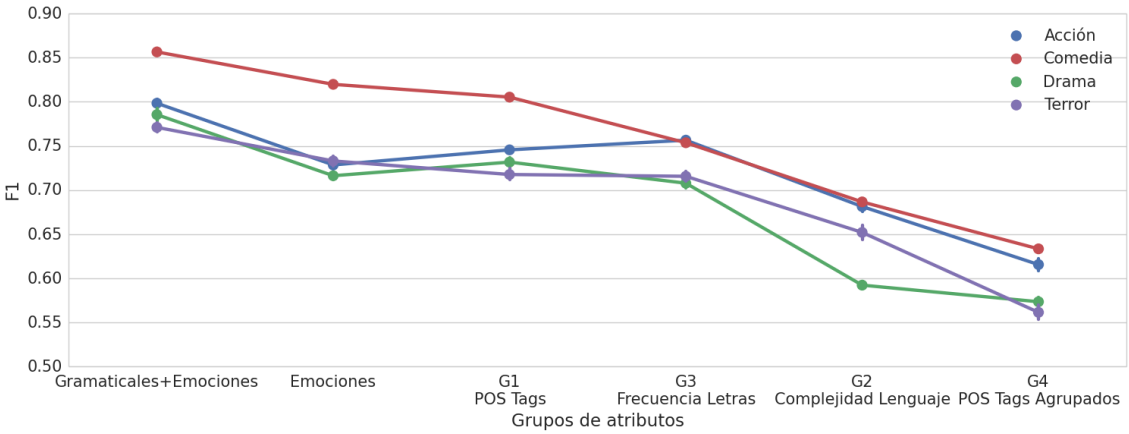


*Fig. 4.4:* Comparación de desempeño de atributos gramaticales y de emociones para cada género, obtenido con un método de validación cruzada de 10 folds y medido con la métrica F1. Las barras verticales representan el desvío estándar entre los distintos folds.

Notamos como el grupo de atributos de emociones resulta el más informativo para la clasificación de los géneros "Comedia" y "Terror", pero no así para los géneros "Acción" y "Drama", dónde la frecuencia de letras y la distribución de POS tags son los grupos de atributos que dominan la métrica.

### 4.3.4. Importancia de atributos

Como paso siguiente, estudiamos cuáles son los atributos particulares que resultan de mayor relevancia para la clasificación incluyendo tanto los atributos de emociones como los gramaticales.

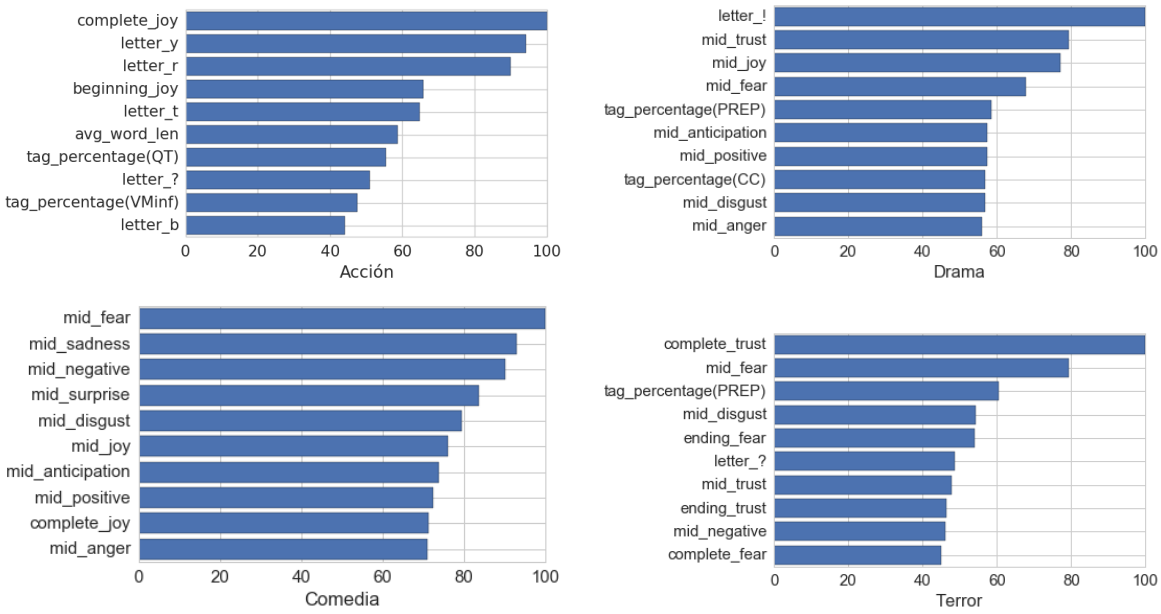


*Fig. 4.5:* Importancia relativa de los 10 atributos gramaticales y de emociones más importantes para cada género, arrojados por el clasificador Random Forest y obtenidos con un método de validación cruzada de 10 folds. Para mayor información, pueden verse los Apéndices Etiquetas de POS tag en español y Traducción de nombres de atributos.

En la figura 4.5 podemos observar el ranking de los diez atributos más importantes para cada género. Vemos que en todos los casos predominan aquellos del grupo de atributos que en la figura 4.4 resultaron de mayor importancia. Por ejemplo, para el género "Comedia" todos los atributos del ranking son de emociones, mientras que sólo 2 lo son para el género "Acción".

Repetiremos ahora el análisis de distribuciones de los atributos de mayor relevancia en cada género, pero esta vez incluyendo los atributos de emociones (ver figura 4.6).

Observamos que todos los *atributos de polaridad* destacados surgen en la sección central de los subtítulos. Vemos en ellos varios aspectos coincidentes con lo que se esperaría de cada género. Por ejemplo, en los subtítulos de terror, hay mayor cantidad de palabras con polaridad negativa que en los demás; así como en las comedias hay menor cantidad de palabras con polaridad negativa.

Sin embargo, algunos otros atributos de este tipo no resultan intuitivos. Es el caso de los dramas, dónde hay más palabras con polaridad positiva; y de las comedias, que tienen menos palabras con polaridad positiva que el resto.

En cuánto a los *atributos de emociones*, hay varios aspectos particulares a analizar sobre cada género. Las películas de acción tienen poca presencia de estos atributos, surgiendo únicamente alegría en menor cantidad que los demás géneros, tanto en la sección inicial

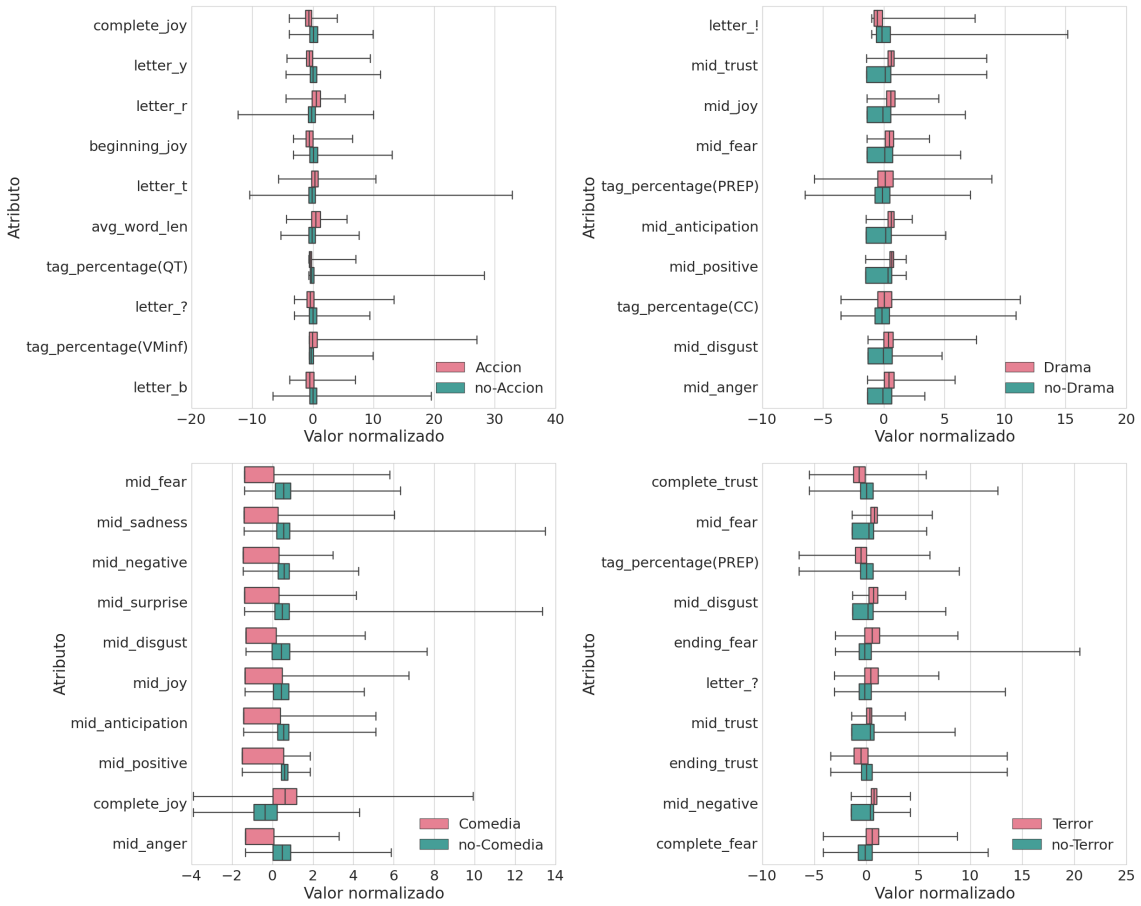


*Fig. 4.6:* Distribución de los 10 atributos gramaticales y de emociones más importantes para cada género estandarizados.

como en el subtítulo completo.

Dentro de las comedias, vemos que la alegría en el subtítulo completo supera ampliamente la del resto de los géneros, como era de esperarse. Todos los atributos restantes para las comedias pertenecen al centro de las películas, y se ve qué todos ellos aparecen en menor cantidad en este género que en el resto. Esto es esperable para las emociones de disgusto, miedo, enojo y tristeza; pero no para las emociones de sorpresa y, sobretodo, alegría.

En los dramas, los atributos provienen en su totalidad de la parte central de las películas. Al contrario de las comedias, aquí todas las emociones superan a las de los géneros restantes. Estas son: alegría, enojo, confianza, disgusto y miedo. Entendemos que los diálogos de los dramas están mucho más cargados de todo tipo de emociones que los demás.

Por último, en terror, los atributos de miedo tanto en el subtítulo completo, como en la sección central y la sección final superan las cantidades de los géneros restantes, tal como era de esperarse. Lo mismo sucede con la emoción de disgusto en la sección central de las películas. La emoción de confianza también tiene un papel importante para este género, en donde la cantidad es menor que en los géneros restantes tanto para las secciones central y final, como para el subtítulo entero.

# 5. CONCLUSIONES

Las películas pueden ser clasificadas según su género cinematográfico, el cual la enmarca tanto en su contenido como en su estructura, permitiéndole al espectador contar con una información mínima de su guión.

El objetivo principal de esta tesis consistió en analizar en qué medida el género de una película se relaciona con la gramática y el contenido emocional de sus diálogos. Para esto, construimos un clasificador automático basado en la técnica de Random Forest, y extrajimos características de los subtítulos para intentar capturar las diferencias entre cada género.

Este análisis fue emprendido a través de dos experimentos. En el primer experimento, estudiamos el desempeño obtenido utilizando atributos únicamente gramaticales, y luego estudiamos cuáles fueron los atributos más relevantes para la clasificación. En el segundo experimento, los resultados del punto anterior fueron contrastados contra la inclusión de atributos de las emociones transmitidas por los subtítulos.

A partir de la realización de este estudio, pudimos concluir que:

1. *El género de una película no sólo se relaciona con su trama, sino también con la estructura gramatical de sus diálogos.* Sin embargo, no tenemos completa certeza de lo anterior debido a que no contábamos con transcripciones del guión de los diálogos sino con una traducción de ellos. Resulta un trabajo a futuro interesante, repetir el análisis de este trabajo con transcripciones de los guiones en su idioma original, para así entender cuánto se vieron afectados nuestros resultados por la traducción.

2. *El género de una película se relaciona sustancialmente y en manera similar tanto con las emociones que transmite en sus diálogos como con su estructura gramatical.* Esto provino de ver cómo los atributos de emociones y de distribución de POS tags resultan ser los más útiles para la clasificación de géneros de las películas. Podemos concluir que los atributos de emociones y gramaticales extraídos con las técnicas utilizadas en nuestro trabajo, aportan un grado de información similar para la clasificación de géneros, ya que en ningún caso se observa una amplia diferencia a favor de uno por sobre el otro.

3. *El género de una película está asociado considerablemente a la distribución de letras de sus diálogos.* Pudimos entenderlo debido a que la proporción de letras constituyó un atributo razonablemente útil para clasificar los géneros de las películas. A pesar de nuestros intentos, no conseguimos dar con los motivos que ocasionaron estos resultados. Entendemos que profundizar en este estudio representa una tarea para un trabajo futuro.

4. Utilizando los atributos estudiados durante el transcurso del trabajo, *el género Comedia es el más fácilmente identificable.*

# BIBLIOGRAFíA


Breiman, Leo (2001). «Random forests». En: *Machine learning* 45.1, págs. 5-32.

Caruana, Rich, Nikos Karampatziakis y Ainur Yessenalina (2008). «An empirical evaluation of supervised learning in high dimensions». En: *Proceedings of the 25th international conference on Machine learning*. ACM, págs. 96-103.

Fischer, Stephan, Rainer Lienhart, Wolfgang Effelsberg y col. (1995). «Automatic recognition of film genres». En: *ACM multimedia*. Vol. 95, págs. 295-304.

Fisher, Sir Ronald Aylmer y col. (1960). «The design of experiments». En:

Helmer, Edmund y Qinghui Ji (2012). «Film Classification by Trailer Features». En:

James, Gareth y col. (2013). *An introduction to statistical learning*. Vol. 6. Springer.

Katsiouli, Polyxeni, Vassileios Tsetsos y Stathes Hadjiefthymiades (2007). «Semantic Video Classification Based on Subtitles and Domain Terminologies.» En: *KAMC*.

Kotsiantis, Sotiris, Dimitris Kanellopoulos, Panayiotis Pintelas y col. (2006). «Handling imbalanced datasets: A review». En: *GESTS International Transactions on Computer Science and Engineering* 30.1, págs. 25-36.

Liaw, Andy y Matthew Wiener (2002). «Classification and regression by randomForest». En: *R news* 2.3, págs. 18-22.

Martin, James H y Daniel Jurafsky (2000). «Speech and language processing». En: *International Edition* 710.

Meyer, David, Friedrich Leisch y Kurt Hornik (2003). «The support vector machine under test». En: *Neurocomputing* 55.1, págs. 169-186.

Mitchell, Thomas M (1997). «Machine learning». En: *New York*.

Mohammad, Saif M. y Peter D. Turney (2013). «Crowdsourcing a Word-Emotion Association Lexicon». En: 29.3, págs. 436-465.

Nitze, I, U Schulthess y H Asche (2012). «Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification». En: *Proc. of the 4th GEOBIA*, págs. 7-9.

Ojala, Markus y Gemma C Garriga (2010). «Permutation tests for studying classifier performance». En: *Journal of Machine Learning Research* 11.Jun, págs. 1833-1863.

Raileanu, Laura Elena y Kilian Stoffel (2004). «Theoretical comparison between the Gini index and information gain criteria». En: *Annals of Mathematics and Artificial Intelligence* 41.1, págs. 77-93.

Rasheed, Zeeshan, Yaser Sheikh y Mubarak Shah (2005). «On the use of computable features for film classification». En: *IEEE Transactions on Circuits and Systems for Video Technology* 15.1, págs. 52-64.

Schmid, Helmut (1995). «Improvements in part-of-speech tagging with an application to German». En: *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer.

— (2013). «Probabilistic part-of-speech tagging using decision trees». En: *New methods in language processing*. Routledge, pág. 154.

Shambharkar, Prashant G y MN Doja (2015). «Automatic classification of movie trailers using data mining techniques: A review». En: *Computing, Communication & Automation (ICCCA), 2015 International Conference on*. IEEE, págs. 88-94.

Stamatatos, Efstathios, Nikos Fakotakis y George Kokkinakis (2000). «Text genre detection using common word frequencies». En: *Proceedings of the 18th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, págs. 808-814.

Yu, Bei (2008). «An evaluation of text classification methods for literary study». En: *Literary and Linguistic Computing* 23.3, págs. 327-343.

Yuan, Xun y col. (2006). «Automatic video genre categorization using hierarchical SVM». En: *2006 International Conference on Image Processing*. IEEE, págs. 2905-2908.

Yuan, Ye, Qin-Bao Song y Jun-Yi Shen (2002). «Automatic video classification using decision tree method». En: *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*. Vol. 3. IEEE, págs. 1153-1157.

Zhou, Howard y col. (2010). «Movie genre classification via scene categorization». En: *Proceedings of the 18th ACM international conference on Multimedia*. ACM, págs. 747-750.

# Apéndice

# A1. ETIQUETAS DE POS TAG EN ESPAÑOL

| POS tag | Descripción |
|---|---|
| ACRNM | Acronym (ISO, CEI) |
| ADJ | Adjectives (mayores, mayor) |
| ADV | Adverbs (muy, demasiado, cómo) |
| ALFP | Plural letter of the alphabet (As/Aes, bes) |
| ALFS | Singular letter of the alphabet (A, b) |
| ART | Articles (un, las, la, unas) |
| BACKSLASH | Backslash ( \ ) |
| CARD | Cardinals |
| CC | Coordinating conjunction (y, o) |
| CCAD | Adversative coordinating conjunction (pero) |
| CCNEG | Negative coordinating conjunction (ni) |
| CM | Comma (,) |
| CODE | Alphanumeric code |
| COLON | Colon (:) |
| CQUE | "Que" (as conjunction) |
| CSUBF | Subordinating conjunction that introduces finite clauses (apenas) |
| CSUBI | Subordinating conjunction that introduces infinite clauses (al) |
| CSUBX | Subordinating conjunction underspecified for subord-type (aunque) |
| DASH | Dash (-) |
| DM | Demonstrative pronouns (ésas, ése, esta) |
| DOTS | POS tag for "..." |
| FO | Formula |
| FS | Full stop punctuation marks |
| INT | Interrogative pronouns (quiénes, cuántas, cuánto) |
| ITJN | Interjection (oh, ja) |
| LP | Left parenthesis ("(", "[") |
| NC | Common nouns (mesas, mesa, libro, ordenador) |
| NEG | Negation |
| NMEA | Measure noun (metros, litros) |
| NMON | Month name |
| NP | Proper nouns |
| ORD | Ordinals (primer, primeras, primera) |
| PAL | Portmanteau word formed by a and el |
| PDEL | Portmanteau word formed by de and el |
| PE | Foreign word |
| PERCT | Percent sign (%) |
| PNC | Unclassified word |
| PPC | Clitic personal pronoun (le, les) |
| PPO | Possessive pronouns (mi, su, sus) |
| PPX | Clitics and personal pronouns (nos, me, nosotras, te, sí) |
| PREP | DEL Complex preposition "después del" |
| PREP | Preposition and Negative preposition (sin) |

| POS tag | Descripción |
|---|---|
| QT | Quotation symbol |
| QU | Quantifiers (sendas, cada) |
| REL | Relative pronouns (cuyas, cuyo) |
| RP | Right parenthesis (")", "]") |
| SE | "Se" (as particle) |
| SEMICOLON | Semicolon (;) |
| SLASH | Slash (/) |
| SYM | Symbols |
| UMMX | Measure unit (MHz, km, mA) |
| VCLIfin | Clitic finite verb |
| VCLIger | Clitic gerund verb |
| VCLIinf | Clitic infinitive verb |
| VEadj | Verb estar. Past participle |
| VEfin | Verb estar. Finite |
| VEger | Verb estar. Gerund |
| VEinf | Verb estar. Infinitive |
| VHadj | Verb haber. Past participle |
| VHfin | Verb haber. Finite |
| VHger | Verb haber. Gerund |
| VHinf | Verb haber. Infinitive |
| VLadj | Lexical verb. Past participle |
| VLfin | Lexical verb. Finite |
| VLger | Lexical verb. Gerund |
| VLinf | Lexical verb. Infinitive |
| VMadj | Modal verb. Past participle |
| VMfin | Modal verb. Finite |
| VMger | Modal verb. Gerund |
| VMinf | Modal verb. Infinitive |
| VSadj | Verb ser. Past participle |
| VSfin | Verb ser. Finite |
| VSger | Verb ser. Gerund |
| VSinf | Verb ser. Infinitive |

## A2. TRADUCCIÓN DE NOMBRES DE ATRIBUTOS

| Tipo | Nombre del atributo | Descripción |
|---|---|---|
| **Proporción de POS tags** | tag_percentage(*posTag*) | proporción del POS tag *posTag* |
| **Complejidad del lenguaje** | speech_density | densidad del habla |
| | words_per_sentence | cantidad de palabras por oración |
| | avg_word_len | longitud media de palabras |
| | uppercase | proporción de palabras en mayúsculas |
| | dynamic_verbs | proporción de verbos dinámicos |
| | stative_verbs | proporción de verbos estáticos |
| **Frecuencia de letras** | letter_*letter* | proporción de la letra *letter* |
| **Proporción de POS tags agrupados** | grouped_adjectives | en la categoría adjetivos |
| | grouped_adverbs | en la categoría adverbios |
| | grouped_nouns | en la categoría sustantivos |
| | grouped_verbs | en la categoría verbos |
| | other | en la categoría otros |
| **Proporción de polaridad** | complete_*polarity* | del subtítulo completo |
| | beginning_*polarity* | del subtítulo en el comienzo de la película |
| | ending_*polarity* | del subtítulo en el final de la película |
| | mid_*polarity* | del subtítulo en el desarrollo de la película |
| **Proporción de emociones** | complete_*emotion* | del subtítulo completo |
| | beginning_*emotion* | del subtítulo en el comienzo de la película |
| | ending_*emotion* | del subtítulo en el final de la película |
| | mid_*emotion* | del subtítulo en el desarrollo de la película |