Universidad de San Andrés

Departamento de Matemática y Ciencias

Maestría en Ciencia de Datos

Inmigración en películas

Un análisis utilizando aprendizaje automático a partir de los subtítulos

Wendy Brau

Directora: Marcela Svarc

2024

Maestría en Ciencia de Datos

Departamento de Matemática y Ciencias

Inmigración en películas

Un análisis utilizando aprendizaje automático a partir de los subtítulos

Wendy Brau **2024**

Directora: Marcela Svarc

Universidad de
SanAndrés

Universidad de
SanAndrés

Resumen

Este trabajo caracteriza el contenido de inmigración de las películas a partir del texto de más de 27 mil subtítulos. Primero, se usa una combinación de Fast K-Medoids, Random Forest y clustering jerárquico para definir temáticas estables e interpretables que estén sistemáticamente asociadas al contenido de inmigración. Segundo, se entrenan varios modelos supervisados con el objetivo de construir un índice continuo del grado de contenido de inmigración en cada película. Finalmente, se explora la asociación entre el consumo en cines de películas con mayor o menor contenido sobre inmigración y las dinámicas reales de inmigración.

Agradecimientos

Muchísimas gracias a Marcela Svarc, quien fue fundamental para este trabajo final. Al jurado por la devolución. Y también a Victoria Luca, Miguel Szejnblum, Florencia Altschuler, Victoria Oubiña, Victoria Carbonari, Paula Luvini, María Prieto Alemandi, Natalia Fonzo, Joaquín Torré, Walter Sosa Escudero, Facundo Carrillo, Daniel Fraiman, Valeria Arza, Kaila Yankelevich, Nico Turjanski, Gabriel Goren, Raquel, Nacho, Oli, Lauti, mamá, papá, Lidia, Dalila...

Índice general

Universidad de
SanAndrés

CAPÍTULO 1

Introducción

Las películas reflejan e influyen en dinámicas sociales del contexto como la violencia, los roles de género y la representación de minorías étnicas [18, 19, 12, 46]. Hay trabajos que han analizado el tratamiento de la temática de inmigración en otros corpus [53, 48, 30, 3], o el contenido de las películas a partir de los subtítulos [18, 63, 48, 42, 13, 10, 26, 41, 28]: este trabajo se ubica en la intersección entre ambos, estudia la relación entre las películas y la inmigración. Para hacerlo, caracteriza el contenido de inmigración en las películas a partir de un corpus de más de 27 mil subtítulos, y explora su relación con las dinámicas de inmigración reales.

Se desarrollan tres tipos de análisis con diferentes objetivos. El objetivo principal es determinar desde qué diferentes temáticas las películas pueden tratar sobre inmigración. El segundo, cuantificar el grado de contenido de inmigración de cada película. Finalmente, explorar la relación entre consumo de películas con distinto contenido de inmigración y las oleadas inmigratorias.

Como punto de partida, se considera a una película como "de inmigración" si aparece etiquetada por IMDb bajo las palabras clave "immigration", "immigrant", "migration", "migrant". Primero, se analiza si hay temáticas que aparecen sistemáticamente en este grupo de películas en comparación con el resto. Se define "temática" como un subconjunto del vocabulario con una semántica común. Se propone la siguiente metodología para identificar temáticas de inmigración: (i) agrupar los lemas únicos presentes en los subtítulos en clústers usando Fast K-Medoids a partir de la distancia coseno entre las representaciones vectoriales de los lemas según GloVe; (ii) asignar un valor a cada película en cada clúster mediante la matriz *term frequency - inverse document frequency* ($TFIDF$); (iii) tomar los clústers que tienen mayor poder predictor de que una película sea de inmigración según un modelo *Random Forest*. Repetir este procedimiento bajo distintas iniciaciones aleatorias del método con el objetivo de obtener temáticas de inmigración estables y robustas. Usar clustering jerárquico para reagrupar entre sí los clústers obtenidos en las distintas repeticiones. Una vez definidas las temáticas, se asigna un valor a cada película en cada temática.

Segundo, para tratar de cuantificar el grado de contenido de inmigración en las películas, se entrenan modelos supervisados que predicen la etiqueta binaria de inmigración de IMDb, y se toman las probabilidades predichas por un modelo como índice de contenido de inmigración. Esto permite obtener una medida continua del contenido de inmigración de una película, que aporta información más granular que la etiqueta binaria, dado que una película puede

tratar la temática de inmigración en menor o mayor medida. Se exploran dos alternativas. Por un lado, entrenar modelos supervisados (Naive Bayes, Análisis Discriminante, Regresión Logística, Random Forest, K vecinos más cercanos) usando el producto interno entre la matriz $TFIDF$ y la matriz de representaciones vectoriales de los lemas como matriz de *features*. Por el otro, un *finetuning* de RoBERTa, dado que los modelos de *transformers* suelen tener muy buen desempeño cuando se adaptan a una tarea específica de interés como la clasificación.

Finalmente, se combinan los valores de las películas en las temáticas de inmigración previamente definidas y del índice de contenido de inmigración general con datos de recaudación de cada película en cines en distintos países y años para definir una métrica de consumo de contenido de inmigración en películas en cines. Se explora la asociación con las dinámicas de inmigración real en distintas geografías y épocas.

En lo que sigue, el Capítulo 2 sitúa el aporte de este trabajo en relación a trabajos previos. El Capítulo 3 detalla los datos y la metodología. El Capítulo 4 presenta los resultados para cada uno de los análisis realizados. El Capítulo 5 lista las limitaciones, alternativas o posibles mejoras para las distintas etapas del análisis, y plantea posibles líneas de investigación a futuro. Las conclusiones, al final.

CAPÍTULO 2

Marco teórico

Este trabajo se inscribe en el marco de las contribuciones del aprendizaje automático y del procesamiento de lenguaje natural (PLN) al estudio de fenómenos sociales y culturales como la inmigración. Una amplia literatura ha estudiado las actitudes públicas hacia los inmigrantes, pero la mayoría lo hizo a partir de encuestas [7, 55, 24]. La disponibilidad de grandes corpus de texto de consumos culturales -como los subtítulos de películas- constituye una fuente de datos novedosa donde analizar la representación de la inmigración aprovechando las técnicas de PLN.

En particular, este trabajo se inspira en aquellos que estudian la interacción entre consumos de películas o programas de televisión y fenómenos sociales más amplios. Estudiar esta interacción es relevante porque los consumos no sólo reflejan, si no que son una fuente de adquisición de valores y comportamientos sociales aceptables. Por lo tanto, sus contenidos, sesgos y estereotipos pueden favorecer la adquisición de cierto tipo de valores en detrimento de otros [18, 19, 12, 46]. A grandes rasgos, este conjunto de trabajos puede dividirse en dos tipos:

1. Aquellos que miden los efectos del consumo de películas o programas sobre ciertas temáticas en las conductas sociales asociadas. Por ejemplo, encuentran que el crimen violento disminuye en los días con mayor audiencia en películas violentas, explicado por una sustitución entre estas actividades por la autoselección de individuos más violentos a este tipo de películas; que las familias cercanas a las zonas donde se mostró un documental de Al Gore aumentan su compra voluntaria de *carbon offsets* los meses posteriores; o que las personas que vieron el programa de televisión educativa *MTV Shuga* mejoran sus actitudes respecto a la violencia doméstica, el HIV y comportamientos sexuales de riesgo [15, 31, 5, 6]. Todos ellos parten de una definición a priori sobre el conjunto de películas o programas referentes a la temática de interés.

2. Aquellos que analizan la expresión de fenómenos sociales en el contenido de las películas, es decir, buscan cuantificar distintos tipos de contenido, y la presencia de sesgos y estereotipos. Entre los hallazgos, las mujeres y las minorías étnicas están sub-representadas en cuanto a tiempo en pantalla y uso de la palabra; las mujeres aparecen menos asociadas a palabras referidas a la inteligencia que los hombres; existen patrones de asociación entre las características sociodemográficas de los personajes, las relaciones entre ellos y las líneas que tienen (por ejemplo, películas con personajes

latinos o de orígenes mixtos tienen más líneas sexuales, y personajes de mayor edad usan más palabras asociadas a logros y religión); es posible identificar contenido de odio en las películas [20, 46, 18, 63, 40]. Estos trabajos buscan cuantificar distintos tipos de contenido, muchas veces usando técnicas de PLN a partir del guión o los subtítulos.

El presente trabajo se inscribe en la segunda línea, y, partiendo de una mayor cantidad de subtítulos que en los trabajos revisados previamente, usa PLN para analizar una temática menos explorada en las películas hasta el momento: la inmigración. Esta temática ha sido analizada en otros tipos de corpus. Por ejemplo, en *tweets* de legisladores de los Estados Unidos, para clasificar su posición respecto de la inmigración; en grandes bases de datos administrativas, para identificar organizaciones sin fines de lucro orientadas a la inmigración; en artículos de la prensa británica, para caracterizar el tratamiento la inmigración en ese medio [53, 48, 30, 3]. Incluso hay trabajos que se dedicaron específicamente a crear un corpus de textos sobre inmigración, como el *Corpus Multilíngue sobre Migração e Refúgio* (COMMIRE), que agrupa documentos y materiales lingüísticos que circulan entre personas migrantes y quienes trabajan con ellas [17]. Por otra parte, los textos que reflejan el contenido de las películas (ya sea sinopsis, guión o subtítulos) han sido muy usados con otros objetivos de análisis, principalmente para predecir el género (tema) de las películas y usarlo en el desarrollo de algoritmos de recomendación [42]. Entre estos últimos, algunos aplican técnicas no supervisadas y modelado de tópicos, y otros aplican métodos supervisados, como redes recurrentes o K vecinos mas cercanos [42, 13, 10, 26, 41, 28].

Este trabajo combina métodos no supervisados y supervisados con el objetivo principal de identificar distintas temáticas asociadas específicamente al contenido de inmigración, en vez de a distintos tipos de contenido en general. Entre los métodos no supervisados, usa clustering jerárquico y Fast K-Medoids, similar a [26], aunque partiendo de 40 veces más cantidad de subtítulos y de las representaciones vectoriales de las palabras según modelos pre-entrenados, en vez de desde representaciones de tipo bolsa de palabras. Entre los métodos supervisados, usa K vecinos más cercanos como en [28] y modelos de *transformers* adaptados como en [63], en este caso con el objetivo secundario de construir el índice de contenido de inmigración.

Para la definición de las temáticas de inmigración, se usa una metodología que no requiere de anotación previa de los tópicos, en contraste con varios métodos recientes sobre análisis de tópicos que requieren de corpus anotados, ya sea con clases de contenido predefinidas o con anotaciones de a pares de ejemplos indicando si pertenecen a un tópico similar o no [64, 21]. Entre los métodos que no requieren de ningún tipo de anotación previa, los más tradicionales se basan en modelos probabilísticos como Asignación Latente de Dirichlet (ALD). Ahora bien, una limitación de estos modelos es que parten de representaciones de bolsa de palabras (BOW), que no considera la relación semántica con palabras del contexto. Por lo tanto, la literatura más reciente aprovecha la información provista por representaciones vectoriales de las palabras (modelos de *embeddings*), sobre las cuales aplica métodos de clustering [52, 11, 23, 1, 59]. Este es el camino que también toma el presente trabajo.

La bondad de la metodología para el armado de temáticas se evalúa en base a dos características: la estabilidad y la alineación con categorías que tengan

un sentido semántico para un humano. Las técnicas tradicionales suelen tener un mejor desempeño en ese sentido que otras como los modelos neuronales de tópicos. Para asegurar la estabilidad, la metodología propuesta involucra un método de conjunto y reagrupación a partir de varias repeticiones de la estimación del modelo, lo que también suele generar mejores resultados [2]. Finalmente, cabe mencionar que varios trabajos recientes exploran el uso de grandes modelos de lenguaje para el aprendizaje *few-shot* o *zero-shot* (es decir, dando pocos o ningún ejemplo de etiquetado correcto) de tópicos en textos, o para la evaluación de modelos de tópicos [22, 54, 49, 37]. A futuro podría explorarse las ventajas y desventajas de su uso para analizar las temáticas de inmigración en películas.

Hipótesis

Se considera que una película es de inmigración si en la trama haya una historia o conflicto principal o secundario relacionado a la inmigración, o si la condición de inmigrante o descendiente de inmigrante es relevante para la construcción de alguno de los personajes.

Si se piensa en películas arquetipo como Pandillas de Nueva York (2002), Brooklyn (2015), Tori y Lokita (2022) o Las Nadadoras (2022), probablemente se encuentren distintas temáticas en las películas de inmigración relacionadas a los países de orígenes y destino de los inmigrantes y al conflicto principal de la trama. A lo largo de los años, el foco de las películas ha cambiado, desde las pandillas urbanas de descendientes chinos, irlandeses o italianos para alcanzar riqueza y poder, al optimismo y el trabajo duro de los inmigrantes para integrarse en la sociedad y progresar materialmente, la convivencia e integración de personas de distintos orígenes en los ejércitos de la Segunda Guerra Mundial, los problemas de pobreza, prejuicios y discriminación, la nostalgia por las raíces, los problemas de la ilegalidad, las brechas generacionales entre los inmigrantes y sus hijos [43].

En cuanto a la relación entre las dinámicas de inmigración reales y el consumo de contenido de inmigración en películas, hay motivos para pensar que la asociación puede ser tanto positiva como negativa. Por un lado, será positiva si las nuevas sociedades con inmigrantes tienen interés por las historias propias o de los nuevos vecinos. Hay quienes sostienen que hubo una proliferación de películas sobre el tema tras los aumentos de la inmigración en Europa, lo que muestra el creciente interés al respecto por parte de los hacedores de políticas [4]. Ahora bien, esto no significa que el público esté crecientemente interesado en ese tipo de consumos. La cobertura muchas veces negativa sobre los inmigrantes por parte de los medios de comunicación (asociándolos a la ilegalidad o la delincuencia), puede generar rechazo o hartazgo por parte del público para con el tema inmigratorio. Además, así como las actitudes públicas hacia los inmigrantes varían dependiendo de distintas características sociodemográficas de los inmigrantes, puede ser que el sentido de la asociación sea distintos para películas que traten sobre inmigración desde distintos ejes temáticos o perspectivas [7].

CAPÍTULO 3

Metodología

A continuación se detallan, en orden, los datos usados y los pasos seguidos en el análisis, resumidos en el diagrama 3.1. El código principal puede encontrarse en https://github.com/wbrau-udesa/tesis-MCD.

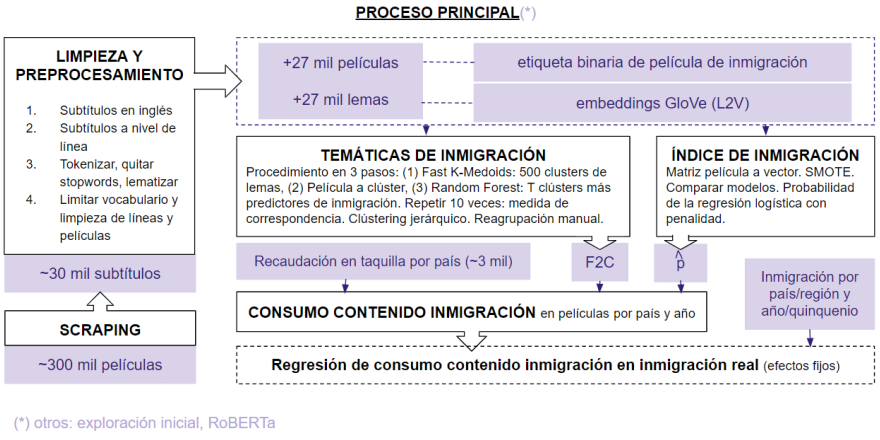
3.1. Datos

Películas

El conjunto de películas a usar se tomó de las bases de datos no comerciales de IMDb, en particular de la base *titlebasics*, que consta de casi 10 millones de películas, y entre otras variables tiene un identificador de cada título (`tconst`), el nombre, el año de estreno, el tipo, el género [29]. Se filtró por tipo, quedándose solamente con las películas; por género, excluyendo documentales, cortos, shows, películas sólo para adultos, y aquellas sin un género imputado; por año de estreno, excluyendo aquellas con datos faltantes de año de estreno y aquellas con fecha de estreno posterior a 2023, dado que la base incluye películas sin estrenar. Luego del filtro, quedaron 364054 películas.

Adicionalmente, se listaron los títulos asociados a las palabras "immigration",

Figura 3.1: Diagrama de datos y métodos



"immigrant", "migration" y "migrant" en la búsqueda por palabra clave[1], de los cuales aproximadamente la mitad está entre los títulos filtrados en el paso anterior. El resto (de aquí en más, títulos "extra-inmigración") también fue incluido en una primera instancia para poder construir el corpus más grande posible de subtítulos de películas de inmigración, que representan una cantidad muy pequeña respecto del total de películas. De aquí en más, "películas de inmigración" referirá a las películas etiquetadas por IMDb como de inmigración, es decir, aquellas que aparecen listadas cuando se buscan películas asociadas a las palabras clave "immigration", "immigrant", "migration" y "migrant". "Películas de no-inmigración" referirá a todo el resto de las películas.

Para construir el corpus de subtítulos de películas se descargaron los archivos de subtítulos en inglés disponibles en https://yts-subs.com/ vía *scraping*. Es posible buscar el subtítulo correspondiente a cada título utilizando el identificador `tconst`. Sólo se encontraron subtítulos para un 8 % de las películas, quitando 42 duplicados (casos donde dos películas distintas tienen nombres muy parecidos y por error de YIFY ambas tienen cargado el mismo archivo de subtítulos). Esto es, hay subtítulos para 28557 películas, de las cuales 456 son de inmigración (el 1.6 %).

Finalmente, Box Office Mojo tiene datos de recaudación de las películas en cines de distintos países. Entre las películas con subtítulos, hay datos de recaudación por país para un 9 % de ellas (2444 películas), de las cuales el 3 % (82 películas) son de inmigración.

Para la construcción de las temáticas e índice de inmigración, se usó todo el corpus de subtítulos de películas, descartando algunas tras la limpieza que se detallará en la Sección 3.2. Para el índice de inmigración, también se descartaron 34 títulos que eran solamente de "extra-inmigración". El corpus de subtítulos contiene muchas más películas nuevas, sobre todo desde los 2000 (Figura 3.2.a.).

Para la aplicación a la medición de consumo de contenido inmigratorio por país y año, se usó el subconjunto de las películas con datos de taquilla (i) que tuvieron un sólo estreno, porque como los datos no distinguen la recaudación a nivel país y estreno a la vez, si una película fue estrenada más de una vez no es posible determinar qué fracción de la recaudación asignar a cada año; (ii) en países que tienen datos para al menos 200 películas en toda su historia. Esto resulta en un conjunto de 2160 películas con datos de recaudación en 60 países. También hay más datos desde los 2000 (Figura 3.2.b.). A diferencia del patrón general de las películas con subtítulos, hay una caída abrupta durante la pandemia debido al cierre de los cines, y una baja en torno a 2015, que podría reflejar la irrupción de las plataformas de streaming.

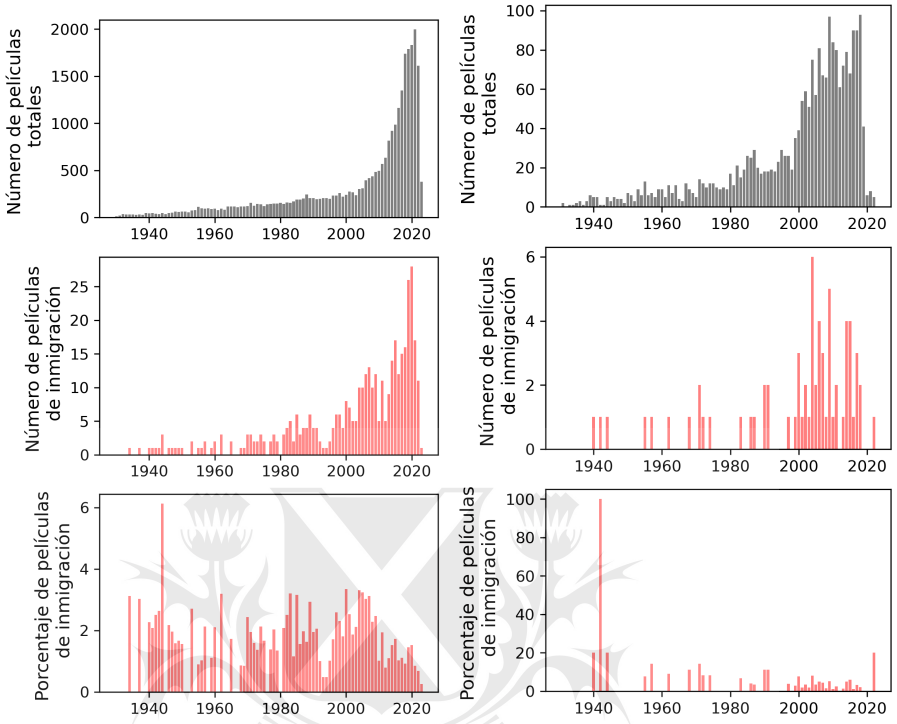
Inmigración

Se usaron dos fuentes principales de datos de inmigración:

- Los datos de inmigración por país de destino y quinquenio del Departamento de Asuntos Económicos y Sociales de las Naciones Unidas, disponibles entre 1990 y 2020 [60]. Provee datos del total de inmigrantes como porcentaje de la población total en el país de destino y la variación anual promedio en el total de inmigrantes.

[1] https://www.imdb.com/search/keyword/

Figura 3.2: Películas totales y de inmigración estrenadas cada año



(a) Con subtítulos analizados (b) Con subtítulos y datos de recaudación

- Los datos de refugiados y otra población migrante vulnerable por país de asilo y año del Alto Comisionado de las Naciones Unidas para las personas Refugiadas (UNHCR), disponibles entre el 2000 y el 2023 [61]. Provee el stock total a final del año del número de refugiados bajo el mandato de la UNHCR, personas que buscan asilo, personas internamente desplazadas de preocupación para UNHCR, otras personas bajo necesidad de protección internacional, personas sin estado, y otros grupos de preocupación a los cuales UNHCR les ha brindado protección o asistencia[2].

3.2. Limpieza y preprocesamiento

Según el objetivo específico del análisis, puede haberse usado un dataset con distintos niveles de limpieza y preprocesamiento. Se parte de una base de datos a nivel de película donde el texto del subtítulo está guardado como texto en una variable, y se aplican los pasos listados en orden:

1. **Subtítulos en inglés.** Predecir el idioma para cada subtítulo usando la librería `langdetect`[3] y quedarse con los subtítulos que son en inglés con mayor probabilidad.

[2]Ver más detalles en https://www.unhcr.org/refugee-statistics/methodology/definition/.
[3]Detalles en https://pypi.org/project/langdetect/.

Cuadro 3.1: Dataset de subtítulos a nivel de línea

line_number	time	line	tconst
1408	01:26:31,108 –>01:26:35,246	to get your sleep.	tt2084952
164	00:06:15,588 –>00:06:17,647	i'll call you back.	tt0988595
2122	01:16:19,242 –>01:16:21,878	thank you.	tt0145503

2. **Nivel de línea.** Los subtítulos descargados son archivos que contienen texto con los diálogos de la película y otra información específica como las marcas temporales donde cada línea aparece o el nombre del creador de los subtítulos. Se armó un dataset a nivel de línea del subtítulo, guardando en una variable el texto del subtítulo y en otra las marcas temporales, y eliminando las líneas referidas al autor de los subtítulos o a los creadores de la película. Además, se eliminó de cada línea las marcas html de negritas o itálicas, y la parte donde se anuncia qué personaje está hablando para los subtítulos que lo indican. Se aplicó esta limpieza a los subtítulos en archivos de tipo ".srt" que son el 99 % de los subtítulos de la muestra, y el resto fue descartado. El Cuadro 3.1 muestra 3 filas aleatorias del dataset resultante (que tiene más de 47 millones de filas), donde las variables `tconst` y `line_number` identifican a la película y la línea respectivamente.

3. **Tokenizar, quitar *stopwords* y lematizar.** Se usó el modelo de lenguaje pre entrenado para inglés `en_core_web_sm` de Spacy[4], con algunas pequeñas modificaciones. En el tokenizador se agregó "-" a la lista de sufijos y prefijos; se agregaron algunas palabras a la lista de *stopwords*, principalmente onomatopeyas "oh", "whoa", "yikes". Para la lematización se utilizó el método *lookup*, que busca cada token en tablas de búsqueda precargadas para lematizarlo, en vez de usar un modelo de lenguaje que prediga la función sintáctica de cada palabra en cada línea para lematizar. En este caso resulta más eficiente el método de búsqueda porque la predicción de la función sintáctica de cada palabra en líneas tan cortas no es buena (en general predice sustantivos), mientras que las tablas de búsqueda son bastante abarcativas. Se eliminaron las líneas que quedan vacías tras este proceso (más de 6 millones).

   Este paso del preprocesamiento será útil principalmente para el armado de temáticas de inmigración (Sección 3.4), donde querremos resumir la información de las palabras y capturar su campo semántico. Para otros tipos de análisis, lematizar puede llevar a la pérdida de información relevante.

4. **Limitar el vocabulario y limpiar líneas y películas**. Luego del proceso anterior se tienen 478 mil lemas únicos aproximadamente. A continuación, se quitaron lemas, líneas e incluso películas, obteniendo un dataset limpio 27743 películas finales y un vocabulario definido de 27016 lemas únicos. Se eliminaron:

   - Lemas muy infrecuentes, aquellos con menos de 50 apariciones. Probablemente no serán útiles para extraer características de los subtítulos que de forma general -para más de una película- estén

[4] https://spacy.io/models/en#en_core_web_sm.

asociadas a contenido de inmigración. El 82 % de los lemas, de hecho, aparece una única vez.

- Los 20 lemas más frecuentes, que pueden pensarse casi como stopwords, son palabras muy frecuentes asociadas a una gran variedad de contenidos, como "know", "come", "get".
- Lemas que no tienen representación vectorial en el modelo preentrenado glove-wiki-gigaword-300 (GloVe) disponible en la librería Gensim [45]. Por lo general son palabras con errores de tipeo o insultos, o poco comunes.
- Las líneas con valores extremos en cantidad de lemas. Se tomaron las 50 líneas con más cantidad de lemas y se filtró de forma manual cuáles eliminar. Por lo general contenían algún error.
- 11 películas con muy pocos lemas alfabéticos o menos de 50 líneas.
- Lemas que aparecen en casi todas las películas o en casi ninguna, usando la frecuencia inversa entre los documentos (de aquí en más IDF, por la *inverse document frequency*). La IDF del lema l se computa como:

$$IDF(l) = \log\left(\frac{F}{df(l)}\right) + 1, \tag{3.1}$$

donde F es el número de películas y $df(l)$ la cantidad de películas en las que aparece el lema l. Se quitaron los 25 lemas con mayor IDF, que son en su mayoría verbos, y los lemas con IDF mayor o igual a 8 (5399 lemas), que son en su mayoría nombres propios.

3.3. Exploración inicial

Para explorar si los subtítulos pueden tener contenido informativo sobre si una película trata sobre inmigración, se definió un "vocabulario de inmigración" a partir de corpus de textos distintos a los subtítulos de películas, y se computaron y compararon dos medidas de la cantidad de vocabulario de inmigración en películas de inmigración versus en películas de no-inmigración.

Para construir el vocabulario de inmigración se tomaron hasta las 1000 palabras más similares a las palabras "immigration", "immigrant", "migration", "migrant" según los *thesaurus* de Sketch Engine de distintos corpus en inglés. La elección de 1000 se debe a que corresponde al límite de descarga de Sketch Engine. Se usaron los siguientes corpus:

1. *enteten21*: textos de internet, descargados entre octubre y diciembre de 2021 y enero de 2022, y que contiene 52 mil millones de palabras, recolectadas filtrando textos con "mayor valor lingüístico", quitando oraciones incompletas, anuncios publicitarios, contenido repetitivo [56, 32].

2. *bnc2*: el British National Corpus (BNC) tiene 100 millones de palabras de muestras de inglés británico escrito (el 90 %) y hablado de la última parte del siglo XX [58].

3. *unpc_en*: el United Nations Parallel Corpus en inglés consiste de registros oficiales y otros documentos parlamentarios de las Naciones Unidas entre 1990 y 2014 [38, 57].

4. *parlamint21_gb*: debates del parlamento británico [16].

El *thesaurus* es una lista palabras que pertenecen al mismo campo semántico que una palabra clave inicial. La pertenencia al mismo campo semántico se define a partir del supuesto de que las palabras que aparecen en el mismo contexto tienen un significado similar. El puntaje de similaridad se calcula a partir del porcentaje de colocaciones compartidas entre la palabra clave y las otras palabras del corpus con la misma función sintáctica[5].

Una vez obtenidas las palabras similares a las cuatro palabras clave, se aplicó el Porter Stemmer de la librería NLTK para quedarse con las raíces de las palabras. Para cada *stem* único, se tomó el máximo valor del puntaje de similaridad.

Finalmente, para cada película se construyeron dos puntajes que miden el contenido de *stems* asociados a inmigración, y se comparó la distribución de cada puntaje en las películas de inmigración y de no-inmigración. Para la comparación se usaron 402 películas de inmigración y una muestra aleatoria de 402 películas de no-inmigración entre el conjunto de películas finales (luego del paso 4 de la limpieza descrita en la Sección 3.2), pero tomando todo el texto del subtítulo luego de haber quitado las marcas propias de los subtítulos (es decir, luego del paso 2 de la Sección 3.2).

Siendo S el total *stems* únicos y $similarity_s$ la similaridad de cada *stem* s con las palabras clave de inmigración, los puntajes de contenido de inmigración se definen de la siguiente manera:

1. **Puntaje 1**: la suma de las similaridades de los *stems* que aparecen en el texto de los subtítulos de la película, sin importar cuántas veces aparecen. Si d_s es una variable indicadora igual a 1 si el *stem* s aparece en el subtítulo de la película f y 0 si no aparece, entonces:

$$puntaje1_f = \sum_{s=1}^{S} similarity_s \, d_s$$

2. **Puntaje 2**: la suma de las similaridades de los *stems* que aparecen en el texto de los subtítulos de la película, multiplicada por la cantidad de veces que aparece cada *stem*. Sea c_s la cantidad de veces que el *stem* s aparece en el subtítulo de la película f, entonces:

$$puntaje2_f = \sum_{s=1}^{S} similarity_s \, c_s$$

Esta exploración inicial parte de definir el contenido de inmigración de forma externa a los subtítulos, con una regla rígida por la cual las palabras similares a las cuatro palabras clave ("immigration", "immigrant", "migration", "migrant") son las que constituyen el vocabulario de inmigración. Sin embargo, puede ser que las películas traten la inmigración de una forma distinta a la que se habla del tema en otros ámbitos (como el parlamento británico, las Naciones Unidas o internet), y que lo hagan a través de temáticas diversas. Por lo tanto, el análisis que sigue tratará de encontrar distintas temáticas asociadas al contenido de inmigración presentes en las películas a partir de sus subtítulos.

[5]Ver más en https://www.sketchengine.eu/guide/thesaurus-synonyms-antonyms-similar-words

3.4. Temáticas de inmigración usando clústers

De aquí en más, "temática de inmigración" referirá a un conjunto de lemas con una semántica común que distingue a las películas de inmigración de las películas de no-inmigración. Se asignará a cada película un valor en cada temática. Para hacerlo, se comienza por el siguiente procedimiento de 3 pasos:

1. Armar de K clústers del conjunto de lemas de todas las películas:

   a) Tomar la representación vectorial de cada lema usando el modelo preentrenado GloVe. Como los lemas de los subtítulos se filtraron previamente para incluir solamente aquellos presentes entre los 6 mil millones de tokens del modelo GloVe, no hay palabras fuera de vocabulario (ver la sección 3.2). Se obtiene la matriz $L2V_{27016 \times 300}$, donde las filas son los lemas únicos y las columnas los 300 elementos de la representación vectorial de cada lema.

   b) Calcular la matriz de distancias coseno entre cada par de lemas en base a sus representaciones vectoriales y normalizarla entre 0 y 1: matriz $D_{27016 \times 27016}$.

   c) En base a D, agrupar los lemas en K clústers vía Fast K-Medoids. Este método busca minimizar la suma de las disimilaridades entre los puntos de los clústers a definir y los K puntos (lemas) en los datos que se tomen como centroides de cada clúster. La disimilaridad puede basarse en distintas medidas de distancia, en este caso, será la distancia coseno calculada en D. El algoritmo que se suele usar para encontrar los clústers es Partitioning Around Medoids (PAM) [36]. En este caso, dada la cantidad de datos, se utilizó la versión acelerada FasterPAM [51, 50]. Como el algoritmo tiene inicialización aleatoria, usar una semilla s para poder reproducir los resultados.

2. Una vez obtenidos los clústers de lemas, asignar un valor a cada película en cada clúster, definiendo la matriz $F2C_{F \times K}$ ("película a clúster") como:

$$F2C_{F \times K} = TFIDF_{F \times L} \cdot C_{L \times K} \tag{3.2}$$

   Donde:

   - F es la cantidad de películas, L la cantidad de lemas únicos en el vocabulario, y K la cantidad de clusters.
   - $TFIDF_{F \times L}$ es la matriz *term frequency - inverse document frequency*, que indica la importancia de cada lema en cada película tomando en cuenta la frecuencia de aparición del lema en la película de interés respecto de la frecuencia de aparición en todo el conjunto de películas. Si $IDF(l)$ está dado por la ecuación (3.1) y $TF(l, f)$ es la frecuencia de aparición del lema l en el subtítulo de la película f:

$$TFIDF^{T}_{F \times L} = TF(l, f)\ IDF(l)$$

   - $C_{L \times K}$ es una matriz donde cada lema es una fila y cada clúster una columna, y se tiene un valor de 1 en la posición $c_{l,k}$ si el lema l pertenece al clúster k.

Es decir, para asignar el valor de la película f en el clúster k, sumar los valores en la matriz $TFIDF$ de todos los lemas del clúster k para la película f.

3. Para identificar cuáles de las temáticas están asociadas a la inmigración, usar la matriz $F2C_{F\times K}$ para predecir si una película es de inmigración o no, es decir, los K clústers son las *features*. Usar Random Forest como método de clasificación (crossvalidando los hiperparámetros de máxima profundidad y mínima caída de impureza para minimizar la métrica de área bajo la curva ROC) y ordenar los clústers según la *feature importance* (también llamada *mean impurity decrease*, MDI): el promedio entre todos los árboles de la caída acumulada de impureza asociada a ese clúster-*feature*. Quedarse con los T clústers más predictores (con mayor MDI) de que la película sea de inmigración.

Resumen del procedimiento:

1. Obtener K clústers de lemas a partir de su representación vectorial usando Fast K-Medoids. Semilla s para la iniciación aleatoria.
2. Calcular el valor de cada película en cada clúster (matriz $F2C$) mediante la matriz $TFIDF$.
3. Predecir si una película es de inmigración en base a $F2C$ y tomar los T clústers más predictores de que lo sea.

Deben definirse los valores de K -cuántos clústers de lemas armar en el paso inicial- y los valores de T -cuántos de los clústers más predictores de que una película sea de inmigración tomar como temáticas de inmigración-.

Para definir el valor de K, se realizó el procedimiento anterior para distintos valores de K y se compararon los resultados (Cuadro 3.2). Lógicamente hay una relación de proporcionalidad inversa entre la cantidad de clústers y el número promedio de lemas en cada uno de ellos. Al aumentar K, los clústers quedan más diferenciados entre sí, los lemas en cada uno de ellos son más homogéneos, y sube el Silhouette promedio. Aunque para todos los valores de K los clústers más predictores de películas de inmigración resultaban ser conjuntos de lemas con un sentido semántico claro y las temáticas que aparecían usando distintos K eran parecidas (entre ellas, ley inmigratoria, historia, lenguajes), para valores más altos de K resulta más fácil asignar una etiqueta con un sentido semántico específico a cada clúster, dado que los lemas dentro del clúster son más homogéneos. Finalmente, el Cuadro 3.2 muestra que al aumentar K, el porcentaje de clústers que son relevantes a la hora de predecir contenido de inmigración disminuye. Esto significa que cuando se construye una menor cantidad de clústers (por ejemplo, cuando $K = 50$) y quedan incluida una mayor cantidad de lemas en cada uno de ellos, muchos de esos lemas no son muy predictores de inmigración, aunque sean cercanos semánticamente a lemas que sí lo son. Se fija $K = 500$, dado que permite obtener temáticas homogéneas, fácilmente etiquetables, sin muchos lemas "ruidosos" a la hora de predecir inmigración, pero conservando un número razonable de lemas por clúster.

Se definió el valor de T con el objetivo de poder obtener temáticas de inmigración robustas, clústers de lemas que sean predictores de que una película

Cuadro 3.2: Valores de K probados para el armado de clústers

K	**50**	**250**	**500**	**1000**
Silhouette	0.027	0.033	0.036	0.038
N promedio de lemas por clúster	540	108	54	27
N clústers con $MDI > 0$ (%)	35 (70 %)	111 (44 %)	132 (26 %)	180 (18 %)

es de inmigración independientemente de la aleatoriedad en su armado. Para eso, una vez fijado $K = 500$, se repitió el procedimiento anterior 10 veces con una iniciación aleatoria s distinta en cada repetición, y se midió la correspondencia entre los T clústers más predictores de inmigración obtenidos en las diferentes repeticiones. Para medir la correspondencia bajo distintos valores de T se definió $mean\_prop\_intersec_T$, la proporción promedio de lemas compartidos entre cada par de clúster de distintas repeticiones. Sea $c_{i,1}$ el conjunto de lemas del clúster $k = i$ de la repetición $r = 1$ y $c_{j,2}$ el conjunto de lemas del clúster $k = j$ de la repetición $r = 2$, se llama Coeficiente de Jaccard ($jaccard(c_{i,1}, c_{j,2})$) a la proporción de lemas compartidos entre ambos clústers:

$$jaccard(c_{i,1}, c_{j,2}) := \frac{|c_{i,1} \cap c_{j,2}|}{|c_{i,1} \cup c_{j,2}|} \tag{3.3}$$

El promedio de la proporción de intersección entre todos los clústers de la repetición 1 y todos los clústers de la repetición 2 es :

$$mean\_prop\_intersec(R_1, R_2) = \sum_{i=1}^{T} \sum_{j=1}^{T} \frac{jaccard(c_{i,1}, c_{j,2})}{T^2}$$

Si se comparan todas las repeticiones entre sí, para cada valor de T:

$$mean\_prop\_intersec_T = \frac{1}{\binom{10}{2}} \sum_{i=1}^{10} \sum_{j=i+1}^{10} mean\_prop\_intersec(R_i, R_j)$$

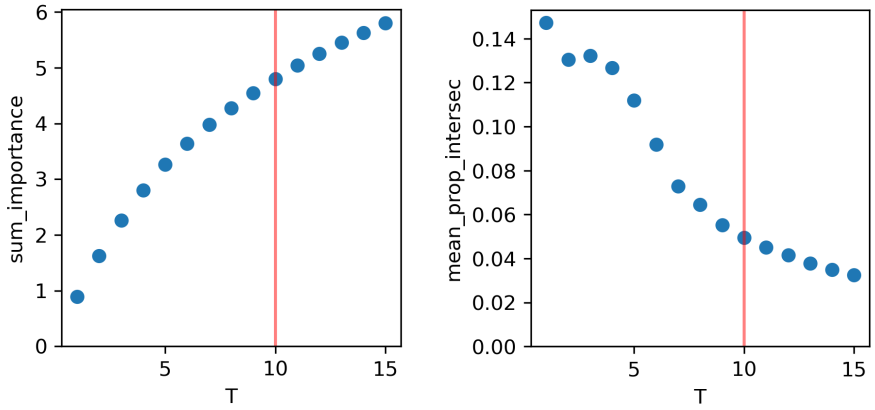
Otro objetivo deseable a la hora de definir T es que queden capturadas varias de las temáticas que efectivamente están asociadas a que una película sea de inmigración. Para medir si esto es así, se calculó la suma de las importancias (MDI) de todos los clústers en todas las repeticiones para cada valor de T:

$$sum\_importance_T = \sum_{i=1}^{T} \sum_{j=1}^{10} MDI(c_{i,j})$$

En la elección del valor de T hay un *trade-off* entre capturar más temáticas relevantes para predecir migración ($sum\_importance_T$) y obtener temáticas más estables ($mean\_prop\_intersec_T$), como se ve en la Figura 3.3.

- A mayor T, aumenta $sum\_importance_T$, aunque cada vez menos. Es decir, aparecen nuevas temáticas asociadas a la inmigración, pero cada nueva temática es menos predictora en el margen de que la película sea de inmigración. En promedio entre las 10 repeticiones, sólo 129 de los 500 clústers son relevantes para predecir inmigración.

Figura 3.3: Definición de T cuando $K = 500$



(a) $sum\_importance_T$ (b) $mean\_prop\_intersec_T$

- A mayor T, cae $mean\_prop\_intersec_T$, la intersección promedio entre los lemas de los clústers de distintas repeticiones, lo que significa que se vuelve menos estable u "obvio" decidir qué temáticas adicionales son las más asociadas a contenido de inmigración.

Elegir un T muy pequeño conlleva el riesgo de dejar por fuera temáticas relevantes para la inmigración, pero un T muy grande puede significar que hay un sobreajuste, esto es, que se están incluyendo temáticas que por casualidad están relacionadas con inmigración en esta muestra de películas. El objetivo es tener una buena cantidad de temas que expliquen inmigración, sin que caiga tanto la estabilidad de los clústers. Hasta los 4 o 5 clústers más importantes, la importancia de los clústers aumenta en mayor proporción de lo que cae la intersección entre ellos. Esto significa que suelen ser los mismos clústers los que están en el top 5 de cada repetición, sólo cambia el orden de importancia asignado a cada uno de ellas. Ahora bien, se decidió incluir el doble de clústers ($T = 10$) para abarcar una mayor cantidad de temáticas, lo que corresponde al 8 % de los clústers con importancia mayor a 0.

Definidos $K = 500$ y $T = 10$, dado que muchos de los clústers obtenidos en las 10 distintas iteraciones intersecan entre sí, se los reagrupó entre sí usando clústering jerárquico aglomerativo. Este método va agrupando los puntos (en este caso, los clústers iniciales) entre sí de forma sucesiva hasta que queden todos agrupados en un gran clúster único. En cada paso, se decide qué par de puntos unir a partir de una medida de distancia distancia entre ellos. Además, a partir del segundo paso, debe definirse un criterio para decidir entre qué puntos de cada conjunto tomar la distancia: si agrupar aquellos que tengan la menor distancia entre los pares de puntos más cercanos (*single linkage*) o entre los pares de puntos más lejanos (*complete linkage*). Se probaron ambas opciones y se usaron dos medidas de distancia entre clústers: el Coeficiente de Jaccard de la ecuación (3.3) y el promedio de la similaridad coseno entre los lemas de los dos clústers, normalizada entre 0 y 1, que se llamará *mean_cos_sim_*01. Sean $c_{i,1}$ y $c_{i,2}$ dos clústers (conjuntos de lemas) de distintas repeticiones y $d(l_u, l_v)$ la distancia coseno entre los lemas u y v, entonces:

Cuadro 3.3: Correspondencia entre distintos criterios y medidas para el clustering jerárquico aglomerativo

complete *mean_cos_sim_01*	***single*** *jaccard*	***complete*** *jaccard*	***single*** *mean_cos_sim_01*
1	[7]	[1]	[2]
2	[24 19]	[16 15]	[21 16]
3	[5]	[2]	[6]
4	[13]	[7]	[10]
5	[21 29]	[22 30]	[20 23]
6	[8]	[3]	[9]
7	[27]	[25]	[25]
8	[10]	[4]	[19]
9	[15]	[5]	[11]
10	[3 16 17]	[17 19 20]	[18 13 14]
11	[20]	[6]	[17]
12	[18]	[8]	[15]
13	[23]	[28]	[24]
14	[11]	[9]	[4]
15	[28]	[27]	[26]
16	[6 12]	[21 29]	[3 5]
17	[9]	[10]	[8]
18	[2]	[24]	[1]
19	[14 26 25]	[11 26 12]	[11 22 12]
20	[22]	[13 14]	[16]
21	[1]	[23]	[7]
22	[4]	[18]	[27]

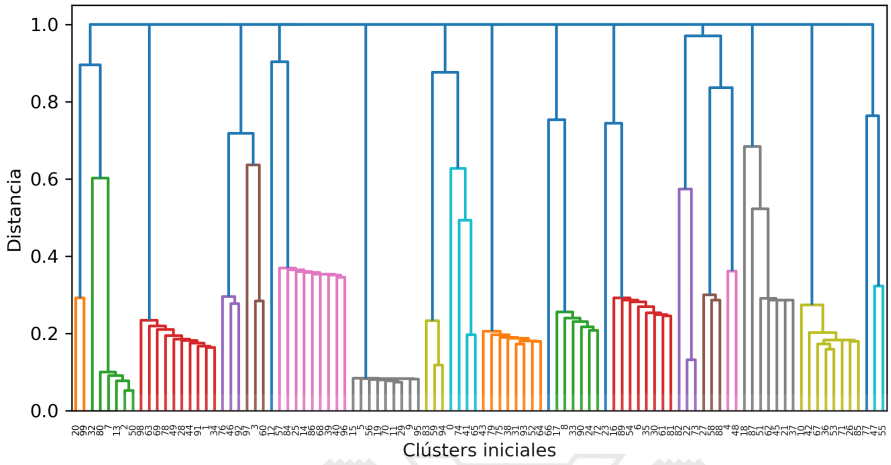
$$mean\_cos\_sim\_01(c_{i,1}, c_{i,2}) = \frac{1}{|c_{i,1}| \cdot |c_{i,2}|} \sum_{l_u \in c_{i,1}} \sum_{l_v \in c_{i,2}} d(l_u, l_v).$$

La reagrupación de clústers obtenida usando las 4 posibles combinaciones de criterios y medidas de distancia es similar. Para la definición de los clústers finales se tomó el criterio *complete* con la medida *mean_cos_sim_01* y se estableció el punto de corte en 0.7 (la Figura 3.4 muestra el dendograma). En el Cuadro 3.3 se puede ver la correspondencia entre los números de clúster-de-clústers asignados a cada clúster inicial usando dicho criterio y medida con los números asignados por el resto de los criterios y medidas.

De 100 clústers de lemas iniciales (10 repeticiones donde seleccionan $T = 10$ clústers), se obtuvieron 22 clústers reagrupados utilizando el clustering jerárquico. A continuación, a cada uno se le asignó manualmente un nombre o etiqueta que hiciera referencia o resumiera el significado de los lemas que incluye, y se hizo una última reagrupación manual en 12 clústers, juntando aquellos clústers con nombres que hacían referencia a un contenido similar.

Una vez obtenidos de esa manera los clústers finales, se calculó el valor de cada película en cada uno de los clústers finales, es decir, la matriz $F2C_{F\times 12}$ de la ecuación 3.2.

Figura 3.4: Dendograma - clustering jerárquico aglomerativo, criterio *complete linkage*, distancia *mean_cos_sim_*01



Modelos de regresión

Una vez elegidas las temáticas de inmigración, el valor de cada película en cada clúster pasa a ser una etiqueta que se puede predecir a partir de un modelo de regresión. Esta sección explora cómo predecir el valor de una nueva película en cada clúster sin conocer el contenido de las películas con las que se entrenó el modelo o se construyeron los clústers, ni tener necesariamente el mismo vocabulario.

Se entrenó un modelo relativamente sencillo para cada temática, donde la variable de respuesta Y_c es el vector de valores de las películas en el clúster c (es decir, es un vector columna de la matriz $F2C_{F\times 12}$).

Para la matriz de regresores a nivel de película se tomó un promedio simple de las representaciones vectoriales de los lemas en la matriz $L2V_{L\times 300}$ que están presentes en cada película. Es decir, si F_1 indica los índices de fila de los lemas de la matriz $L2V$ presentes en la película 1, para esa película se calcula el vector fila $\bar{L2V}^{(1)}_{1\times 300}$ como:

$$\bar{L2V}^{(1)}_{1\times 300} = \frac{1}{|F_1|}\sum_{i\in F_1} L2V_{i\times 300}$$

donde $L2V_{i\times 300}$ representa la fila i de la matriz $L2V$ (la representación vectorial del lema L_i) y $|F_1|$ el número de lemas únicos en la película 1.

Entonces obtenemos la matriz de regresores $X_{F\times 300}$:

$$X_{F\times 300} = \begin{bmatrix} \bar{L2V}^{(1)}_{1\times 300} \\ \bar{L2V}^{(2)}_{1\times 300} \\ \vdots \\ \bar{L2V}^{(F)}_{1\times 300} \end{bmatrix}$$

Luego se quitaron algunas variables entre las 300 que son altamente multicolineales entre sí. Para cada dimensión se calculó la correlación con

todas las demás, y con cuántas la correlación era mayor a 0.5 en valor absoluto. Se quitó el 5 % de las dimensiones (16 dimensiones) que correlacionan con mayor cantidad de otras dimensiones.

A continuación se listan los modelos de regresión entrenados con el objetivo de minimizar el el Error Cuadrático Medio (ECM) entre las predicciones de los modelos y el valor real en cada clúster. Cuando corresponde, se listan los hiperparámetros crossvalidados. Llamando X a la matriz de variables explicativas, F a la cantidad de películas del conjunto de entrenamiento y V al número de vectores de la matriz X[6]:

- Regresión Lineal. La predicción está dada por $\hat{Y_c} = X\hat{\beta}$, donde los coeficientes son estimados como:

$$\hat{\beta} = \arg\min_{\beta} \Big\{ \sum_{i=1}^{F} \left(Y_{ci} - \beta_0 - \sum_{j=1}^{V} x_{ij}\beta_j \right)^2 \Big\}$$

- Regressión Ridge: introduce una penalización L_2 a la regresión lineal, penalizando valores muy grandes en los coeficientes (regulariza):

$$\hat{\beta} = \arg\min_{\beta} \Big\{ \sum_{i=1}^{F} \left(Y_{ci} - \beta_0 - \sum_{j=1}^{V} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{V} \beta_j^2 \Big\}$$

  La penalidad λ es un hiperparámetro a crosvalidar.

- Regresión Lasso: introduce una penalización L_1 a la regresión lineal, premiando un modelo esparso donde los coeficientes correspondientes a alguna de las dimensiones de $L2V$ sean iguales a 0 (selecciona *features*):

$$\hat{\beta} = \arg\min_{\beta} \Big\{ \frac{1}{2} \sum_{i=1}^{F} (Y_{ci} - \beta_0 - \sum_{j=1}^{V} x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{V} |\beta_j| \Big\}$$

  La penalidad α es un hiperparámetro a crossvalidar.

- Regresión Elastic Net: combina las penalidades L_1 y L_2 de los modelos anteriores. La ponderación de cada penalidad es un hiperparámetro a crossvalidar.

- Regresión Random Forest: es una agregación de árboles de regresión. El árbol de regresión es un modelo no lineal que particiona el espacio de *features* en regiones y estima un modelo simple o asigna un valor constante a cada una de ellas. El número de árboles estimadores a promediar y la máxima profundidad del árbol son parámetros a crosvalidar.

Como referencia para evaluar el desempeño de los modelos, se comparó con el resultado de simplemente imputar la media o la mediana de los valores de otras películas en el clúster para calcular el valor de una nueva película en ese clúster. Como métrica para elegir entre modelos se calculó el error cuadrático

[6] Más detalles de los modelos en [27, 33].

medio estandarizado (ECM_std). Sea F el número de películas del conjunto de datos considerados, pueden ser entrenamiento o test:

$$ECM\_std_c = \frac{1}{F}\frac{\sum_{i=1}^{F}(Y_{ci} - \hat{Y_{ci}})^2}{\sum_{i=1}^{F} Y_{ci}^2}$$

Una aclaración importante: como los clústers se construyeron usando todas las películas posibles del conjunto de datos para maximizar el uso de información, necesariamente hubo que usar esos mismos datos para entrenar y evaluar estos modelos. Esto significa que se sobreestimará la capacidad predictiva con respecto al desempeño real a la hora de clasificar una película nueva con la que no se hayan construido los clústers. Por lo tanto, los resultados de estos modelos de regresión deberán tomarse como una exploración inicial a mejorar.

3.5. Índice de inmigración usando clasificación

IMDb asigna la etiqueta de inmigración solamente a algunas películas, pero una película puede tratar la temática de inmigración en menor o mayor medida. Para poder obtener una medida continua del contenido de inmigración de una película, una opción que se explora en esta sección es entrenar un modelo que prediga la etiqueta binaria de inmigración provista por IMDb, no con el objetivo de obtener una buena predicción *per se*, si no para tomar las probabilidades de que la película sea de inmigración predichas por el modelo como índice de contenido de inmigración de esa película.

Se exploran dos alternativas para la clasificación que parten de un nivel distinto de preprocesamiento de los datos:

1. Entrenar y elegir el clasificador con mejor desempeño entre Naive Bayes gaussiano, Análisis Discriminante Lineal y Cuadrático, Regresión Logística con penalidad, Random Forest, K Vecinos Más Cercanos.

2. *Finetuning* del modelo RoBERTa [39], incluido en el Anexo B.

A continuación se detalla la primera estrategia.

F2V

El primer grupo de clasificadores se entrenó definiendo la matriz de *features* $F2V_{F\times 300}$ ("película a vector") como:

$$F2V_{F\times 300} = TFIDF_{F\times L} \cdot L2V_{L\times 300} \tag{3.4}$$

Donde $L2V$ es la representación vectorial de los lemas únicos de todas las películas usando el modelo GloVe (ver Sección 3.4). Es decir, para obtener una representación vectorial de las películas a partir de la representación vectorial de sus lemas, agrega los vectores de todos los lemas que tiene cada película ponderados por la importancia de cada lema en la película según la matriz $TFIDF$. Dicho de otra manera, si cada vector de lemas representa una dimensión semántica abstracta, cada película se acercará al sector del "espacio semántico" de las palabras que tengan más importancia en la película. Entonces, el problema es delimitar regiones de ese espacio semántico abstracto

Cuadro 3.4: Resultados de predecir inmigración con Logit-Lasso con SMOTE versus sin SMOTE

	Con SMOTE	**Sin SMOTE**
% películas de inmigración	1.4 %	1.4 %
% películas de inmigración predichas	20 %	0 %
balanced accuracy	0.65	0.5
roc auc	0.65	0.5
accuracy	0.80	0.98

que correspondan a películas de inmigración versus a películas de no-inmigración. Una vez obtenida $F2V$ se quitaron algunas *features* con alta colinealidad con otras, obteniendo una matriz $F2V_{27709\times285}$[7].

Desbalance

Para resolver el problema del desbalance en el conjunto de datos, dado que hay solamente 1.4 % películas de inmigración entre las 27709:

- Las métricas relevantes para evaluar el desempeño de los modelos son aquellas que prioricen los verdaderos positivos y penalicen falsos negativos, como el área bajo la curva ROC (*roc_auc*) o el *balanced accuracy*. Guiarse por el *accuracy* llevaría a clasificar a todas las películas como de no-inmigración, lo que permitiría lograr un 98.6 % de *accuracy*.

- Se usa la Técnica de Sobresmuestreo Sintético de la Minoría (SMOTE) que genera nuevas observaciones sintéticas de la clase minoritaria cercanas en el espacio de *features* a los ejemplos existentes de la clase minoritaria. Crea ejemplos sintéticos a lo largo del segmento que una a cada observación de dicha clase con algunos de los k vecinos más cercanos también de la clase minoritaria. Es decir, las muestras sintéticas se generan sumando al vector de *features* de una de las observaciones de la clase minoritaria la diferencia entre ese vector y el de su vecino, multiplicada por un número aleatorio entre 0 y 1 [14].

Para entrenar el modelo, se usan los datos de entrenamiento con sobremuestra. Para evaluar el modelo, se toman los datos reales de testeo, sin sobremuestra. El Cuadro 3.4 muestra los cambios en el desempeño cuando se usa SMOTE, tomando como ejemplo para un modelo Logit con penalidad con validación cruzada de hiperparámetros[8]. Aunque predice películas de inmigración por demás, la mejora en las métricas relevantes *roc_auc* y *balanced accuracy* es significativa, porque permite identificar alguna película de inmigración. En cambio, si no se usa SMOTE, el modelo no predice ninguna película de inmigración.

[7] Se partió de un conjunto de 27709 películas luego del cuarto paso del preprocesamiento de la Sección 3.2, quitando 34 títulos de "extra-inmigración".

[8] Cuando se usa SMOTE, los hiperparámetros elegidos son C igual a 1 y penalidad L_2; cuando no se usa SMOTE, los hiperparámetros elegidos son C igual a 0,5 y penalidad L_1.

Modelos entrenados

A continuación, se entrenaron los sieguientes modelos eligiendo los hiperparámetros con validación cruzada y tomando el área bajo la curva ROC como métrica a maximizar[9]:

- Regresión logística con penalidad. La regresión logística aplica una transformación para, a partir de un modelo lineal con penalidad, poder estimar la probabilidad $p(X)$ de que la etiqueta sea 1 en función de la matriz de *features* X (en este caso, la probabilidad de que una película sea de inmigración en función de la matriz $F2V$):

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X + \gamma||\beta||$$

  Donde $||\beta||$ puede ser una penalidad de tipo L_1 o L_2. Los hiperparámetros elegidos fueron: el máximo número de iteraciones, 20000; la penalidad, L_2; el *solver*, saga.

- Random Forest. Es una agregación de las predicciones de múltiples árboles de decisión. Cada árbol de decisión particiona al espacio de covariables en regiones de forma iterativa (definiendo un valor umbral para una variable del espacio de *features* en cada iteración) con el objetivo de maximizar alguna métrica de separación entre clases, y asigna un valor a cada partición final. Los hiperparámetros elegidos fueron: el criterio, minimizar la entropía en cada nodo; la máxima profundidad, 10; la mínima caída de impureza, 0,001; el número de árboles, 200.

- Análisis Discriminante Lineal. Modela la distribución de la matriz de *features* de forma separada para cada una de las clases. Luego usa el teorema de Bayes para obtener estimaciones para $p(X)$, identificando una combinación lineal de *features* que maximice la separación entre clases. Si la distribución de las *features* dentro de cada clase es normal, el modelo es muy similar a la regresión logística [33]. El hiperparámetro tolerancia se fijó en 0,0001.

- Análisis Discriminante Cuadrático. Extiende el análisis discriminante lineal para permitir estructuras de covarianza distintas entre las clases. El hiperparámetro de regularización se fijó en 0,05.

- Naive Bayes Gaussiano. Naive Bayes también usa el teorema de Bayes y asume independencia entre las *features* condicional en cada clase para aproximar $p(y|X)$. Esta versión asume una distribución gaussiana para $p(x_i|y)$.

- K vecinos más cercanos. Clasifica cada punto en los datos basándose en la clase mayoritaria (ponderada o no por la distancia) de sus k vecinos más cercanos. Los hiperparámetros elegidos fueron: la métrica, la distancia coseno; el número de vecinos, 21; ponderar los puntos de acuerdo a la inversa de la distancia entre ellos.

[9] Detalles de los modelos en [27, 33].

Finalmente, se comparan los resultados en el conjunto de datos de testeo para los distintos clasificadores, se elige el de mejor desempeño, y se predicen las probabilidades de que cada película sea clasificada como de inmigración según el modelo elegido. Dichas probabilidades conforman el valor del índice de contenido de inmigración para cada película.

3.6. Aplicación: métricas de consumo de contenido inmigratorio

Combinando las temáticas de inmigración (Sección 3.4) y el índice de contenido de inmigración (Sección 3.5) con datos de recaudación en taquilla es posible construir una medida del consumo de contenido de inmigración en películas, general y por temática, y estudiar su evolución. Es importante notar que esta medida se limita al consumo de contenido de inmigración en las películas vistas en cines. A futuro sería interesante incorporar medidas del consumo de estas películas en plataformas de *streaming*.

Para medir el consumo de contenido de inmigración, se define una medida de recaudación del contenido de inmigración en general y de cada temática por país y año. Primero se calcula el porcentaje de recaudación de cada película respecto del total recaudado en ese país y año, lo que controla por el hecho de que hay países y años donde en general se recauda más. Luego, se multiplica el porcentaje recaudado por la película por el índice de inmigración o por el valor asignado a la película en cada temática. Si $USD_{f,p,a}$ es el total recaudado por la película f en el país p y año a e Y_f es el grado de contenido de inmigración en general o de alguna temática de inmigración de la película f, el consumo del contenido Y en el país p y año a se calcula como:

$$consumo_{Y,p,a} = \sum_f \frac{USD_{f,p,a}}{\sum_j USD_{j,p,a}} Y_{f,p,a} \tag{3.5}$$

Luego, basándose en la cantidad de películas con datos de recaudación disponibles en cada país y año de la Figura 3.5, se hacen dos tipos de análisis:

- Para el período 2002 - 2019 y los países que tienen más de 5 películas todos los años: 1300 películas en 32 países.
- Para el período 1960 - 2019 y cuatro países que tienen datos desde entonces: Estados Unidos, Francia, Nueva Zelanda, Reino Unido.

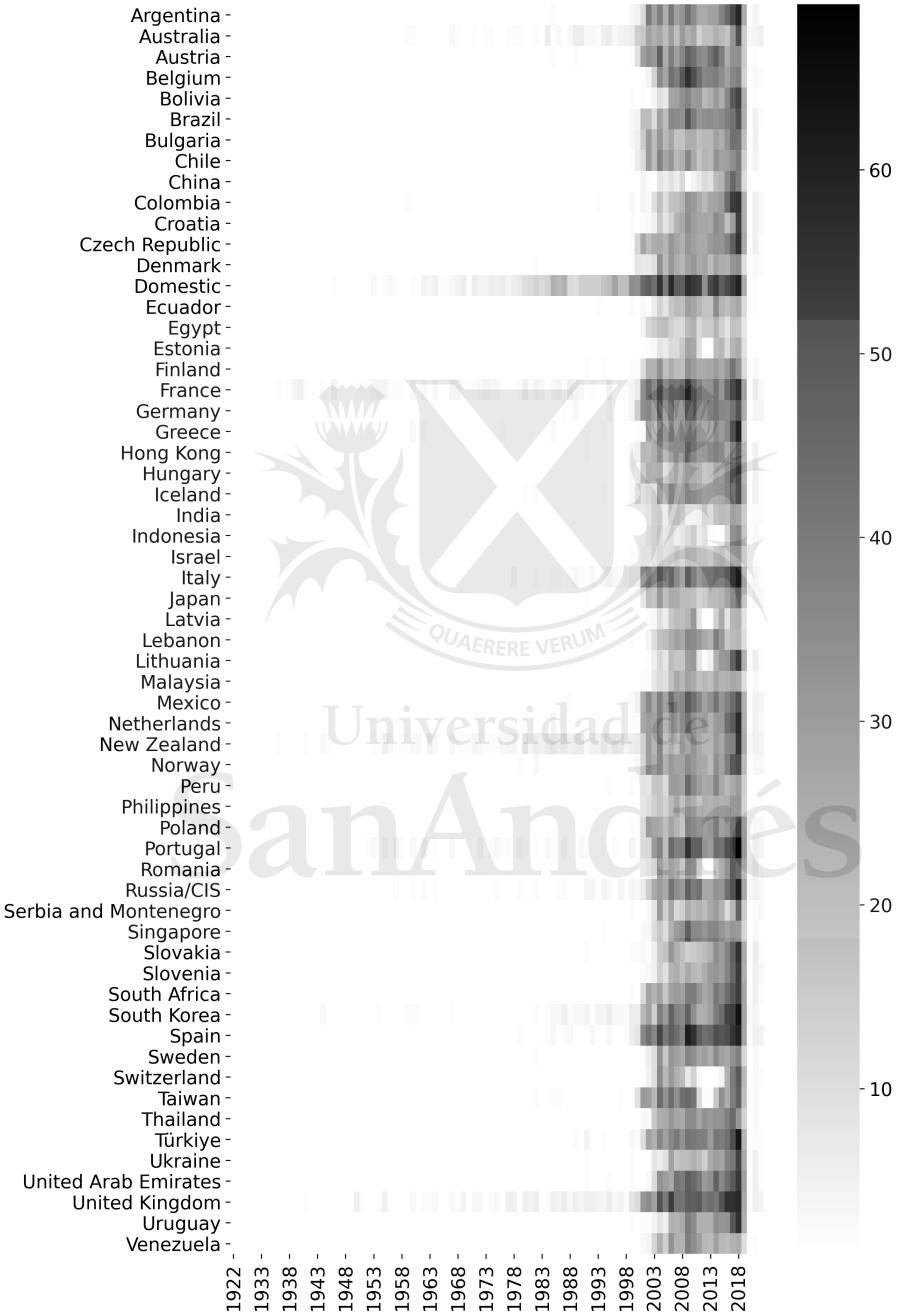
Se calcula $consumo_{Y,p,a}$ para esos años y países y se hacen algunos ejercicios de asociación entre la evolución de dicho consumo y la evolución de la recepción de inmigración en esos años y países, como se detalla a continuación.

Análisis período 2002 -2019

Primero, se calcula la evolución del consumo de contenido de inmigración en cada país a lo largo del tiempo. Luego, se lo calcula agrupando a los países en cuatro grupos:

- Europa: Austria, Bulgaria, República Checa, Dinamarca, Finlandia, Francia, Alemania, Grecia, Hungría, Islandia, Italia, Países Bajos, Noruega, Polonia, Portugal, España, Suecia, Reino Unido.

Figura 3.5: Número de películas por país y año con datos de recaudación



- Estados Unidos.
- América Latina: Argentina, Brasil, Chile y México.
- Corea del Sur, Japón y Hong Kong.

Se realizó un ejercicio de asociación entre el consumo de contenido de inmigración en esas regiones y la recepción de inmigración de refugiados y otra población migrante vulnerable en cada una de ellas según datos de UNHCR (descripción detallada en la Sección 3.1). Las tres primeras regiones tuvieron aumentos movimientos inmigratorios notables de este tipo en los últimos años en números absolutos, mientras que en Corea del Sur, Japón y Hong Kong el número se mantuvo relativamente estable, por lo cual se la toma como región de comparación.

Como exploración inicial, se estimó la asociación entre inmigración (en nivel y en variación interanual) y consumo de inmigración por mínimos cuadrados ordinarios (MCO) a partir del siguiente modelo:

$$consumo_{Y,r,a} = \beta_0 + \beta_1 X_{r,a} + \delta_r + \alpha_a + \mu_{r,a}. \tag{3.6}$$

Donde r refiere a la región y a al año; $consumo_{Y,r,a}$ está dado por la ecuación (3.5) calculado a nivel de región en vez de a nivel de país; $X_{r,a}$ es el nivel o la variación interanual en la inmigración de refugiados y otra población migrante vulnerable en la región r y año a (se estiman dos regresiones, una para el nivel y otra para la variación interanual); δ_r es una variable binaria indicadora de la región, igual a 1 si se trata de Europa, Estados Unidos o América Latina (se estiman tres regresiones separadas, una por cada una de estas regiones) e igual a 0 si se trata de Corea del Sur, Japón y Hong Kong; α_a son variables indicadoras del año; $\mu_{r,a}$ es el término de error. Es decir, aprovechando que se tienen datos en panel que varían tanto entre regiones como en el tiempo, se usa un modelo de efectos fijos por región y año que controla por todos los factores que puedan afectar tanto a la inmigración real como al consumo de contenido inmigratorio que varíen sólo entre años o sólo entre regiones. $\hat{\beta}_1$ es la estimación de la asociación entre inmigración y consumo de inmigración en las películas.

Análisis Estados Unidos, Francia, Nueva Zelanda, Reino Unido

Como hasta los 2000 hay muchas menos películas por año, se agruparon los datos por quinquenio, lo que además permite comparar la evolución del consumo quinquenal de contenido de inmigración en las películas con la llegada de inmigrantes en ese quinquenio, usando los datos de inmigración de las Naciones Unidas disponibles desde 1990 (descripción detallada en la Sección 3.1). Para hacerlo, se estimó por MCO el siguiente modelo:

$$consumo_{Y,p,q} = \gamma_0 + \gamma_1 Z_{p,q} + \kappa_p + \theta_q + u_{p,q}. \tag{3.7}$$

Donde p refiere al país y q al quinquenio; $consumo_{Y,p,q}$ está dado por la ecuación (3.5) calculado a nivel quinquenal en vez de anual; $Z_{p,q}$ es el porcentaje de inmigrantes sobre la población total o la variación interanual promedio por quinquenio en el stock de inmigrantes del país p y el quinquenio q (se estiman regresiones separadas para cada variable); κ_p son variables indicadoras del país; θ_q son variables indicadoras del año; $u_{p,q}$ es el término de error. Una vez más,

se aprovechan los datos en panel para estimar un modelo de efectos fijos, esta vez por país y quinquenio, que controla por todos los factores que puedan afectar tanto a la inmigración real como al consumo de contenido inmigratorio que varíen sólo entre quinquenios o sólo entre países. $\hat{\gamma}_1$ es la estimación de la asociación entre inmigración y consumo de contenido de inmigración en los cines.

CAPÍTULO 4

Resultados

Es posible detectar contenido de inmigración en las películas a partir de los subtítulos y definir temáticas que consistentemente aparecen asociadas a las películas de inmigración.

4.1. Exploración inicial

La exploración inicial da indicios de que el contenido de los subtítulos puede ser útil a la hora de distinguir películas de inmigración versus de no-inmigración.

En la Figura 4.1 se compara la distribución de los puntajes que miden la presencia de palabras similares a "immigration", "migration", "immigrant", "migrant" en los subtítulos de películas de inmigración versus películas de no-inmigración. El Puntaje 1 contabiliza si las palabras de inmigración aparecen al menos una vez, mientras que el Puntaje 2 contabiliza la cantidad de veces que aparecen (ver descripción detallada de los puntajes en la sección 3.3).

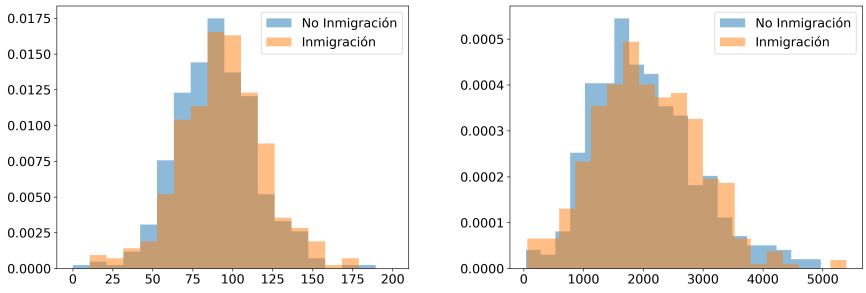
Los histogramas presentan una distribución unimodal y sus soportes se encuentran solapados, es decir, algunas películas de no-inmigración tienen más vocabulario de inmigración que algunas películas de inmigración. Sin embargo, los histogramas para las películas de inmigración están desplazados hacia la derecha, lo que significa que las palabras asociadas a la inmigración suelen aparecer más en las películas de inmigración que en las de no-inmigración. La diferencia es más clara en el Puntaje 1. En cambio, en el Puntaje 2 hay películas de no-inmigración con un puntaje muy alto y películas de inmigración con puntaje muy bajo.

4.2. Temáticas de inmigración

Una película tiene una mayor probabilidad de ser de inmigración si trata alguno de estos temas: cuestiones legales, historia, política, conflictos sociales, economía, lenguaje, cosmovisión y lugares del mundo. El Cuadro 4.1 detalla las distintas temáticas asociadas al contenido de inmigración: el nombre asignado y el número de lemas de cada clúster final, y el Cuadro 4.2 presenta medidas resumen de la cantidad de lemas por clúster. En el Anexo A se detallan los lemas incluidos en cada clúster.

A cada película se le asignó un valor en cada temática. En promedio, las películas de inmigración tratan más (tienen valores más altos) las temáticas de inmigración que las películas de no-inmigración (Cuadro 4.3), con excepción

Figura 4.1: Histogramas de puntajes de contenido de *stems* asociados a inmigración en películas de inmigración vs. de no-inmigración



Puntaje 1 Puntaje 2

Cuadro 4.1: Temáticas de inmigración en las películas

Nombre	**Cantidad de lemas**
Británico	23
Conflictos medio oriente	102
Economía y empleo	112
Europa	141
Gentilicios y lugares del mundo	110
Latino	268
Lenguaje	44
Ley inmigratoria	97
Nazismo	34
Nueva York y Estados Unidos	166
Religión, ideología, cosmovisión	156
Tecnología	58

Cuadro 4.2: Resumen de estadísticas de los clústers finales

Número de clústers finales	12
Promedio nro. lemas por clúster	109
Max. nro. lemas por clúster	268
Min. nro. lemas por clúster	23

Cuadro 4.3: Valores promedio en cada temática para las películas de inmigración versus de no-inmigración

Clúster	(1) No-inmigración	(2) Inmigración	(2) / (1) Ratio
Británico	0.008	0.018	2.25
Conflictos medio oriente	0.045	0.069	1.53
Economía y empleo	0.077	0.100	1.30
Europa	0.046	0.105	2.28
Gentilicios y lugares del mundo	0.054	0.096	1.78
Latino	0.084	0.191	2.27
Lenguaje	0.016	0.039	2.44
Ley inmigratoria	0.047	0.073	1.55
Nazismo	0.022	0.031	1.41
Nueva York y Estados Unidos	0.096	0.134	1.40
Religión, ideología, cosmovisión	0.041	0.052	1.27
Tecnología	0.020	0.012	0.6

de la temática tecnología. El porcentaje de películas de inmigración entre las películas con valores positivos en todos los clústers es del 2.4 %, casi el doble que el porcentaje de películas de inmigración en la muestra, y no hay ninguna película de inmigración con valor 0 en todos los clústers. Ahora bien, las películas de no-inmigración también tocan temáticas asociadas a la inmigración, e incluso puede ocurrir que una película de no-inmigración toque más esas temáticas que una película de inmigración. Esto puede deberse a dos motivos: a que las temáticas no solamente están asociadas a la inmigración, y a que la etiqueta de película de inmigración provista por IMDb sea imperfecta.

Algunos ejemplos ilustran que los clústers sirven para capturar las temáticas centrales de las películas de inmigración. El Cuadro 4.4 muestra la película de inmigración y de no-inmigración con mayor valor en cada temática. En la temática "Británico", Paddington (2014) trata de un oso peruano que viaja a Londres en busca de un hogar. En "Conflictos medio oriente", Exodus (1960) narra la fundación del Estado de Israel. Revenge of the Green Dragons (2014) trata de dos inmigrantes chinos en Nueva York, lo que es coherente con su lugar central en la temática de "Lenguaje". En "Ley inmigratoria" está Crossing Over (2009), que justamente trata de inmigrantes ilegales de diferentes nacionalidades que quieren legalizar su estatus en Los Angeles. La clásica Casablanca (1942) es la principal película de inmigración en la temática "Nazismo": el dueño de un café ayuda a su amada y a su esposo a escapar de los nazis.

Hay temáticas de inmigración que suelen aparecer juntas en las mismas películas. Por ejemplo, la Figura 4.2 muestra que hay una correlación alta entre la temática de Europa y de nazismo, lo que es esperable, y también entre Europa y lugares del mundo, ya que muchos de dichos lugares son europeos.

La evolución de las temáticas de inmigración en las películas se mantiene

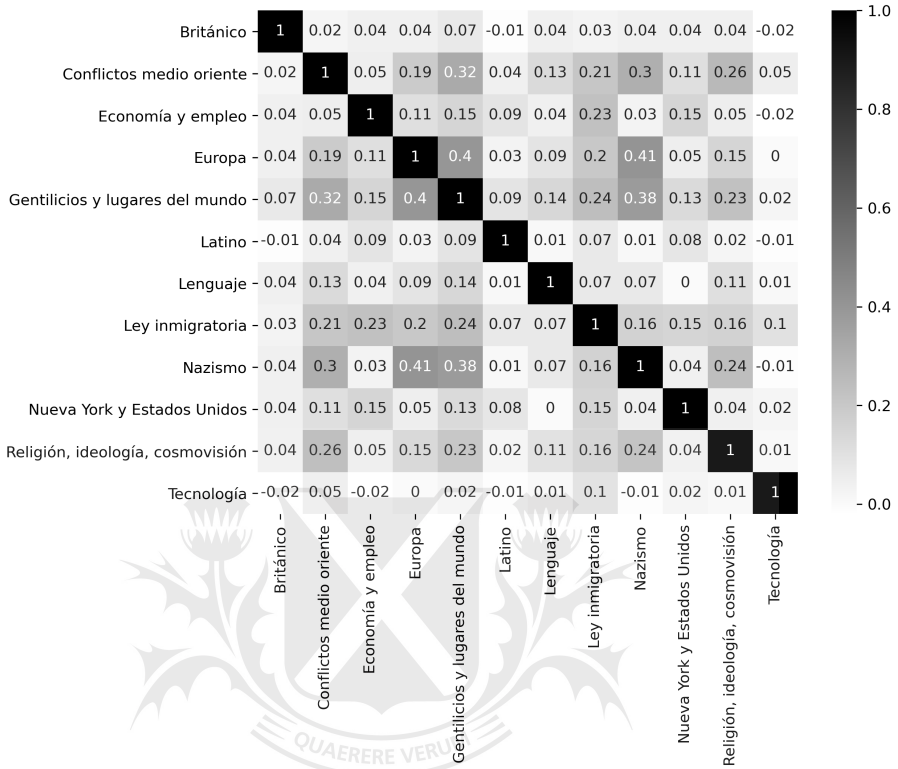
Cuadro 4.4: Películas con mayor valor en cada temática

clúster	No-inmigración	Inmigración
Británico	Paddington (2014)	My Not So Irish Bride (2014)
Conflictos medio oriente	Exodus (1960)	Shock and Awe (2017)
Economía y empleo	Mediterranea (2015)	One Last Deal (2018)
Europa	The Swissmakers (1978)	Man on a String (1960)
Gentilicios y lugares del mundo	Tolo Tolo (2020)	In the Land of Blood and Honey (2011)
Latino	The Motorcycle Diaries (2004)	Che: Part Two (2008)
Lenguaje	Revenge of the Green Dragons (2014)	Christmas in Miami (2021)
Ley inmigratoria	Crossing Over (2009)	A Human Position (2022)
Nazismo	Casablanca (1942)	Orders from Above (2021)
Nueva York y Estados Unidos	E.14 (2020)	Domain (2016)
Religión, ideología, cosmovisión	Viceroy's House (2017)	AmericanEast (2008)
Tecnología	The Bourne Legacy (2012)	Desire Path (2020)

mayormente estable a lo largo del tiempo, aunque se observan algunas tendencias (Figura 4.3). Como era de esperar, vemos que hay un pico en las películas sobre Nazismo en los años '40. Además, hay una caída de la temática de Nueva York y Estados Unidos y de la temática economía y empleo desde los 2000, y un ascenso de la temática de Tecnología. También hay un paulatino ascenso de la stemáticas Lenguaje y Latino.

Finalmente, se presentan los resultados de los modelos de regresión para predecir el valor de cada película en cada clúster (Figura 4.4). Es importante aclarar que, como los valores reales clústers se construyeron usando todas las películas posibles del conjunto de datos para maximizar el uso de información, necesariamente hubo que usar esos mismos datos para entrenar y evaluar estos modelos, y por lo tanto los resultados sobrestimarán su desempeño. En promedio los modelos suelen funcionar algo mejor que los *benchmarks* (imputar la media o la mediana a todas las películas). En los datos de entrenamiento hay sobreajuste y Random Forest parece ser el modelo con mejor desempeño, pero en los datos de testeo sólo es el mejor modelo para predecir los valores de tres clústers (Economía y Empleo, Ley inmigratoria y Nueva York y Estados Unidos). Según los resultados en datos de testeo, para cada temática funciona mejor un modelo distino. Ridge parece ser la mejor opción para Británico, Europa, Latino y Lenguaje. El modelo lineal sin penalización, para los conflictos de medio oriente, Nazismo y Religión, ideología, cosmovisión. Lasso, para Gentilicios y lugares del mundo y Tecnología, o alternativamente el modelo lineal sin penalización.

Figura 4.2: Correlación entre las temáticas de inmigración en las películas



4.3. Índice de contenido de inmigración

El porcentaje de películas etiquetadas por IMDb como de inmigración es muy bajo, y para muchos países y años no hay ninguna película de inmigración a pesar de que se hayan estrenado películas en ese país y año. Este problema se acentúa en los datos de recaudación por película. Aunque en general hay muchas más películas con datos de taquilla para los últimos años (Figura 3.5), aún en los años en los que hay varias películas son muy pocas las clasificadas por IMDb como de inmigración (Figura 4.5.a.). Además, la etiqueta provista por IMDb es discreta, y no captura el grado en que una película trata la temática de inmigración, por lo que muchas películas pueden tener cierto grado de contenido asociado a la inmigración aun cuando IMDb no las clasifica como tales.

Usar el índice de contenido de inmigración permite capturar diferencias en el grado de contenido de inmigración de las películas, y tener más variabilidad en una medida de consumo de ese contenido entre años y países. Como mide el contenido de inmigración de un conjunto mucho más amplio de películas, sirve para medir el consumo de contenido de inmigración en más países y años (4.5.b.).

A continuación se presentan los resultados de entrenar modelos de aprendizaje supervisado para predecir si una película es de inmigración a partir de la matriz $F2V$ (ecuación 3.4). La Figura 4.6 muestra las métricas de desempeño de los modelos de clasificación. Dado el alto desbalance en los

Figura 4.3: Evolución de las temáticas de inmigración a lo largo del tiempo desde 1930 - media móvil 5 años, desde 1930

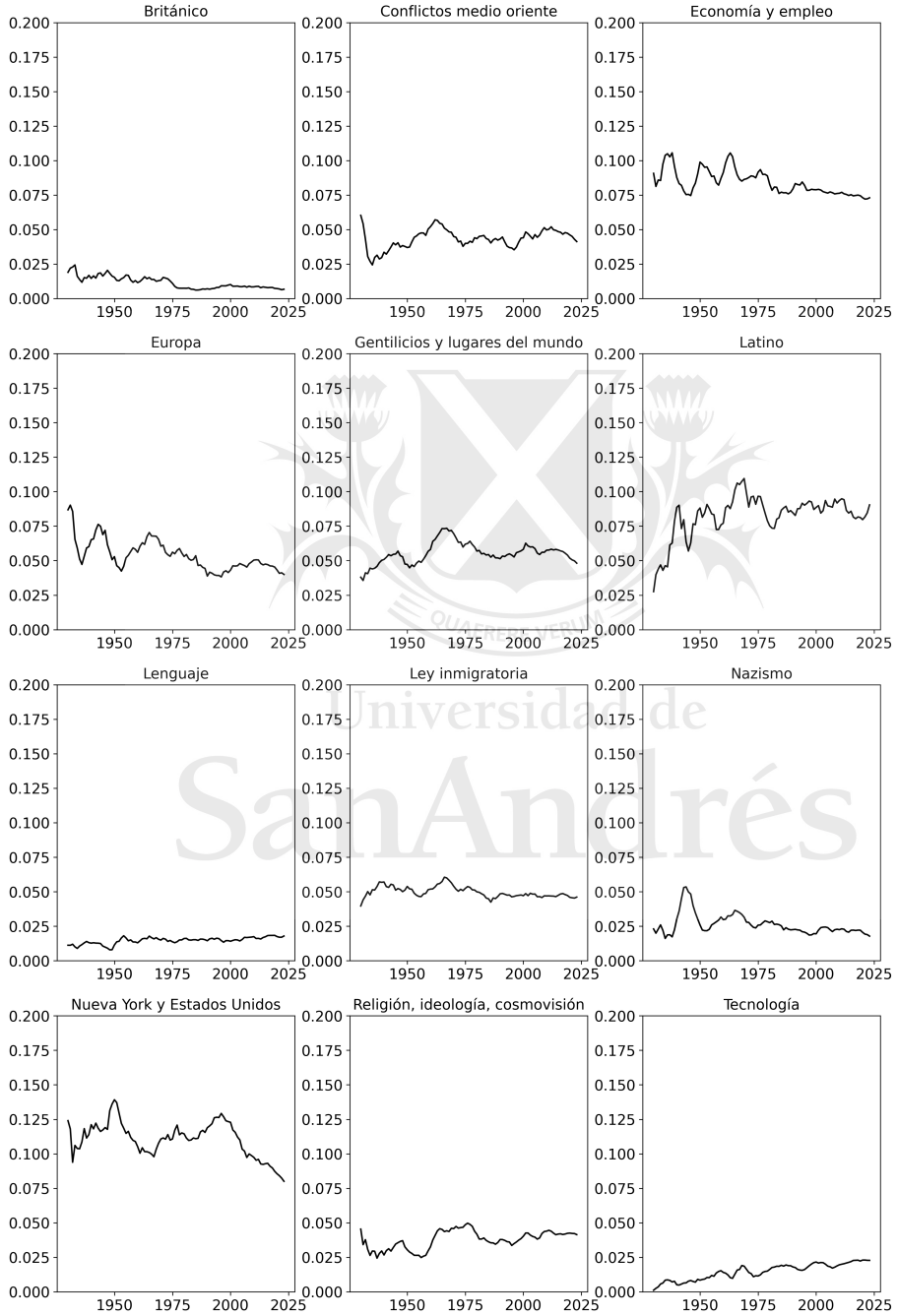


Figura 4.4: Desempeño modelos de regresión para predecir el valor de una película en cada temática de inmigración

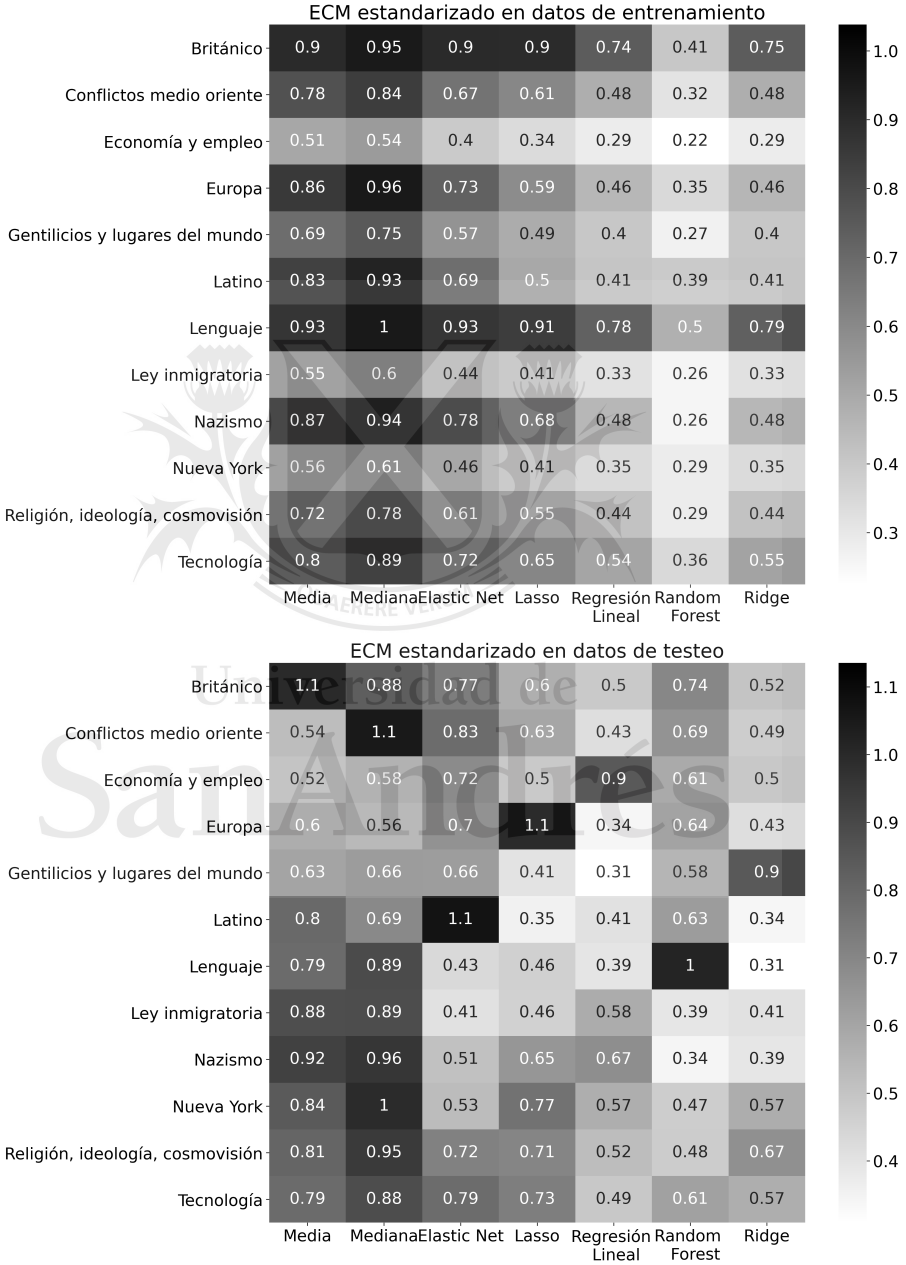
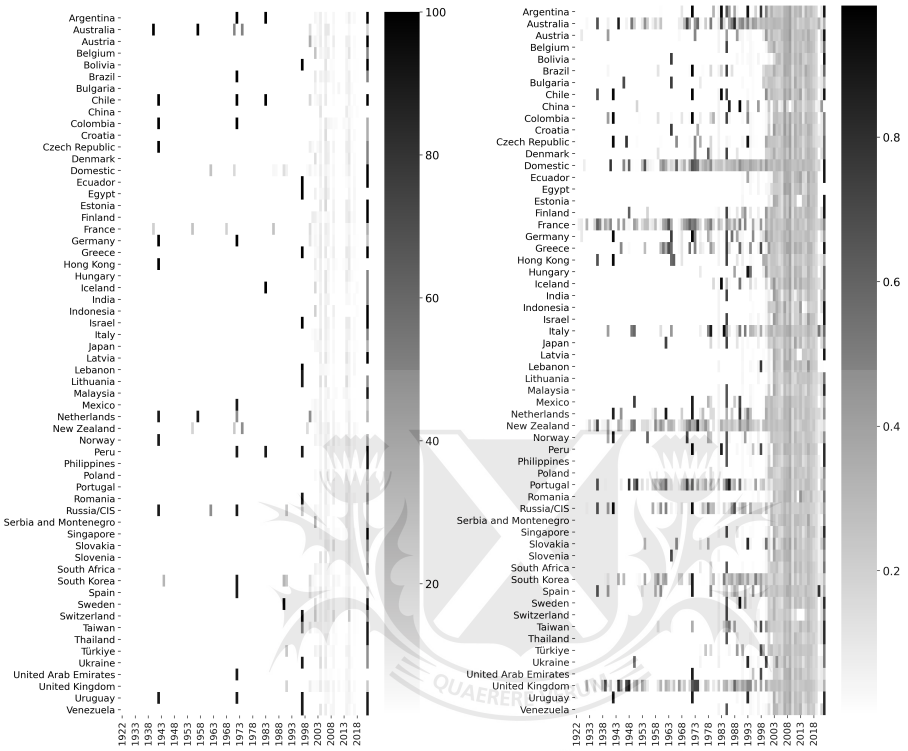


Figura 4.5: Películas de inmigración con datos de recaudación por país y año: porcentaje etiquetado por IMDb versus la probabilidad asignada por la regresión logística con penalidad



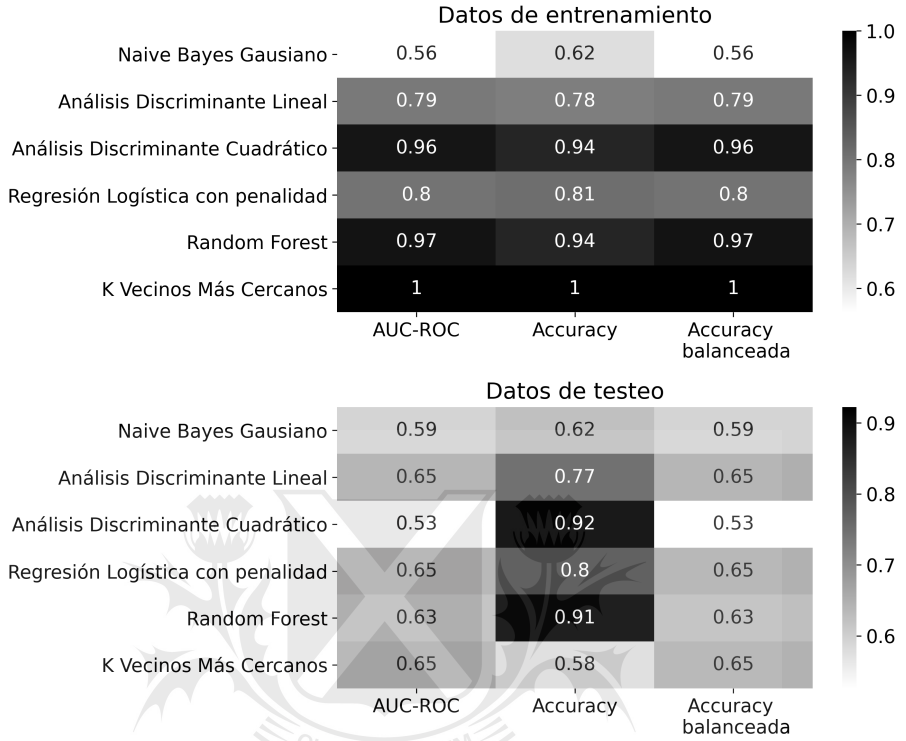
(a) Porcentaje (b) Probabilidad

datos, las métricas relevantes a tener en cuenta son el área bajo la curva ROC o el *accuracy* balanceado. Hay sobreajuste a los datos de entrenamiento. En los datos de testeo, los modelos tienen un desempeño similar, siendo los mejores LDA, Regresión Logística con penalidad y KNN.

Tomando la regresión logística con penalidad, la figura 4.7 muestra su desempeño predictivo[1]. En datos de testeo, usando el umbral *default* de 0.5 para predecir que una película es de inmigración, identifica correctamente a aproximadamente la mitad de las películas de inmigración, y al 80 % de las películas de no-inmigración. Por un lado, esto significa que hay lugar para mejorar la identificación de las películas de migración. Por otro, que la regresión logística predice muchas más películas de inmigración que las que reales: el 20 % de las películas de no-inmigración son clasificadas erróneamente como de inmigración, y como las películas de inmigración son el 98.6 % de las películas, el modelo termina prediciendo 20 % de películas de inmigración, cuando en realidad hay solo 1.4 %. Aunque hay espacio para mejorar, la curva ROC muestra que la regresión logística es superior a una predicción aleatoria. Existe un *trade-off*

[1]La diferencia en el cálculo de AUC-ROC en la figuras 4.6 y 4.7 se debe a que esta última se calculó usando las probabilidades de clasificar a cada película como de inmigración, mientras que para la primera se usaron las predicciones con un umbral de 0.5.

Figura 4.6: Métricas modelos de clasificación

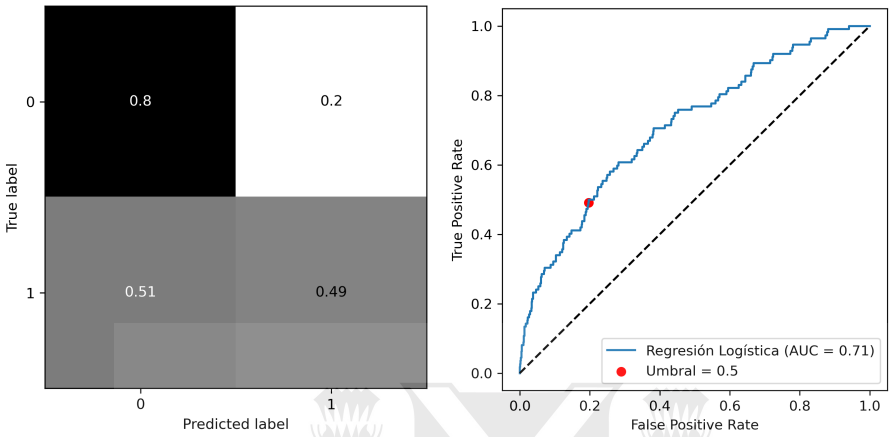


entre tratar de identificar más películas de migración y la tasa de falsos positivos, y no hay un codo evidente en la curva ROC que favorezca la elección de un umbral alternativo a 0.5.

De cualquier modo, el interés principal no es la predicción de la etiqueta binaria de inmigración en sí, si no las probabilidades de que cada película sea de inmigración asignadas por el modelo para usar como índice de contenido de inmigración. La Figura 4.8 muestra la distribución de probabilidades predichas para todas las películas (incluyendo conjuntos de entrenamiento y test) según si IMDb las etiqueta o no como de inmigración. El modelo tiende a asignar probabilidades más altas de ser de inmigración al grueso de películas que efectivamente lo son. Comparar las probabilidades asignadas por este modelo con los puntajes construidos de forma *ad-hoc* en la exploración inicial (Figura 4.1) da cuenta de la enorme mejora en la capacidad de discernir entre películas de inmigración y no-inmigración cuando usamos modelos de aprendizaje automático a partir de los subtítulos.

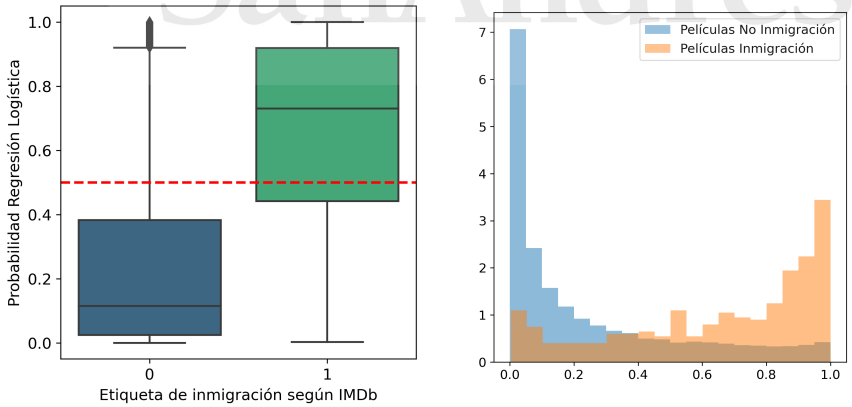
Las dos películas con la probabilidad más alta de ser de inmigración según la regresión logística no están etiquetadas como de inmigración por IMDb. La primera de ellas es 300 Miles to Heaven (1989), cuyo argumento resume Google: «Dos chicos polacos, un adolescente y su hermano pequeño, escapan de la Polonia comunista, se ocultan debajo de un camión y terminan en un campo de refugiados en Dinamarca.» La segunda es R.M.N. (2022): «Matthias regresa al pueblo de montaña de sus padres en Transilvania, deseando volver a ver a su ex Csilla. Al llegar al lugar, se da cuenta del malestar causado por

Figura 4.7: Desempeño predictivo Regresión Logística en datos de testeo



(a) Matriz de confusión (b) Curva ROC

Figura 4.8: Índice de contenido de inmigración en base a la Regresión Logística, según etiqueta real



(a) Boxplots (b) Histogramas

Csilla al ofrecer a dos extranjeros trabajo en su panadería». Ambas películas tienen clara relación con la inmigración, lo que da cuenta de que el etiquetado de IMDb es imperfecto. Entonces, otra de las ventajas del uso de un modelo predictivo es que resume información de los subtítulos asociada a películas de inmigración que es útil para predecir contenido inmigratorio en películas que no fueron etiquetadas como tales.

De manera similar, la película con menor probabilidad predicha de ser de inmigración que es etiquetada por IMDb como de inmigración es Passengers (2016). Guiándose por el resumen, tampoco es evidentemente de inmigración:

> Dos pasajeros que viajan en hibernación a un planeta lejano despiertan, por un error técnico, 90 años antes de llegar a destino. Solos y rodeados de lujos, entre ellos surge el amor; sin embargo, descubren que hay una avería en la nave y tendrán que repararla a tiempo para salvar a las 5000 personas que permanecen hibernando, y a sí mismos.

Ahora bien, la segunda película con menor probabilidad predicha de ser de inmigración pero que es etiquetada por IMDb como de inmigración es Go (2001), que «Go es una película sobre la mayoría de edad (...) que cuenta la historia de un adolescente norcoreano nacido en Japón, Sugihara, y una chica japonesa con prejuicios, Tsubaki Sakurai, de quien se enamora», es decir, sí trata de inmigración: el modelo también comete errores.

4.4. Aplicación a la medición del consumo de contenido de inmigración

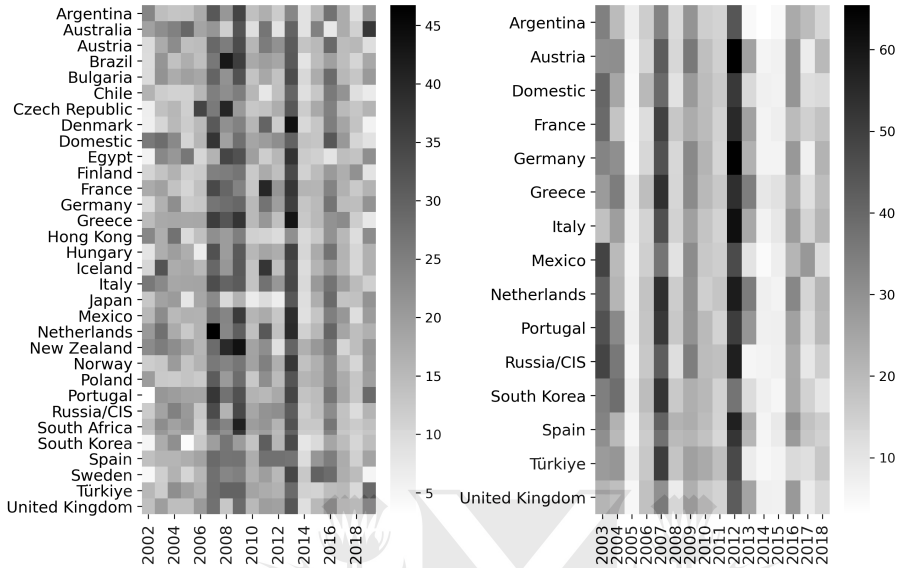
A continuación se realiza una exploración inicial de la evolución del consumo de contenido inmigratorio en películas, y su asociación con movimientos inmigratorios reales en distintos países y años.

La Figura 4.9 (a) muestra la variación en el consumo de contenido de inmigración para los países con más de 5 películas en cada año entre 2002 y 2019. La Figura 4.9 (b) muestra la variación para un "soporte común" de 132 películas estrenadas en 15 países entre 2003 y 2018, esto es, para un subconjunto de películas que se estrenaron en todos los países, eliminando las películas que sólo se estrenaron en algunos de ellos. Sirve para ilustrar el sentido de las métricas de consumo construidas (detalles en la Sección 3.6): para un mismo grupo de películas que está en cartelera en todos los países, el porcentaje de recaudación -el consumo- de las películas con más contenido de inmigración varía entre ellos.[2]

Las Figuras 4.10 y 4.11 agrupan a los países con más de 5 películas en cada año entre 2002 y 2019 en regiones, y exploran la asociación entre el consumo de contenido de inmigración en las películas y la recepción de refugiados y otra población migrante vulnerable en cada una de ellas. Se toman 3 regiones que tuvieron aumentos movimientos inmigratorios notables de este tipo en los últimos años y se las compara con Corea del Sur, Japón y Hong Kong ("región de comparación"), donde la tendencia se mantuvo relativamente estable

[2]El soporte común permite ver si algunos países consumieron menos contenido de inmigración que otros, pero no si dejaron de consumir por completo ese tipo de películas.

Figura 4.9: Variación del consumo de contenido de inmigración por país y año



(a) Países con más de 5 películas todos los años 2002-2019

(b) Soporte común

a lo largo del tiempo. En Europa hubo una tendencia descendiente hasta 2013, cuando se produjo un quiebre y aumentó notablemente de la llegada de refugiados; en Estados Unidos se produjo un cambio de tendencia del mismo signo en 2012, aunque el aumento es menos pronunciado que en Europa. En los países de América Latina (Argentina, Brasil, Chile y México) la tendencia era relativamente estable hasta que se produce un fuerte aumento a partir de 2016.

Se observa que, en las tres regiones, el consumo de contenido inmigratorio promedio disminuyó tras los aumentos de inmigración, mientras que en la región de comparación hubo un aumento. Es decir, hay una asociación negativa entre la recepción de refugiados y otra población migrante y el consumo de contenido de inmigración en general. El consumo de la temática Europa también tiene una asociación negativa con la recepción de inmigración en Europa y Estados Unidos (donde cayó, mientras que en la región de comparación se mantuvo relativamente estable). Ahora bien, hay fuertes oscilaciones en el consumo de contenido inmigratorio año a año, y no hay quiebres evidentes en las tendencias.

Las regresiones de las medidas de consumo de contenido inmigratorio en la cantidad de refugiados y otros migrantes, controlando por año y por región (ver ecuación 3.6), muestran que el consumo de contenido de inmigración suele estar negativamente asociado a la llegada de refugiados y otros migrantes (Figuras 4.12 y 4.13). En Europa hay una asociación negativa y significativa al 5 % entre el número de refugiados y el consumo de películas con más contenido de inmigración en general y de la temática Economía y Empleo, y entre la variación interanual en la cantidad de refugiados y el consumo de la temática de Religión, ideología, cosmovisión.

La Figura 4.14 muestra la evolución en el consumo de contenido de inmigración en películas por quinquenio (indicado por el año de comienzo)

Figura 4.10: Asociación entre el consumo de contenido de inmigración en las películas y la recepción de refugiados y otra población migrante UNHCR, por región y año [PARTE I]

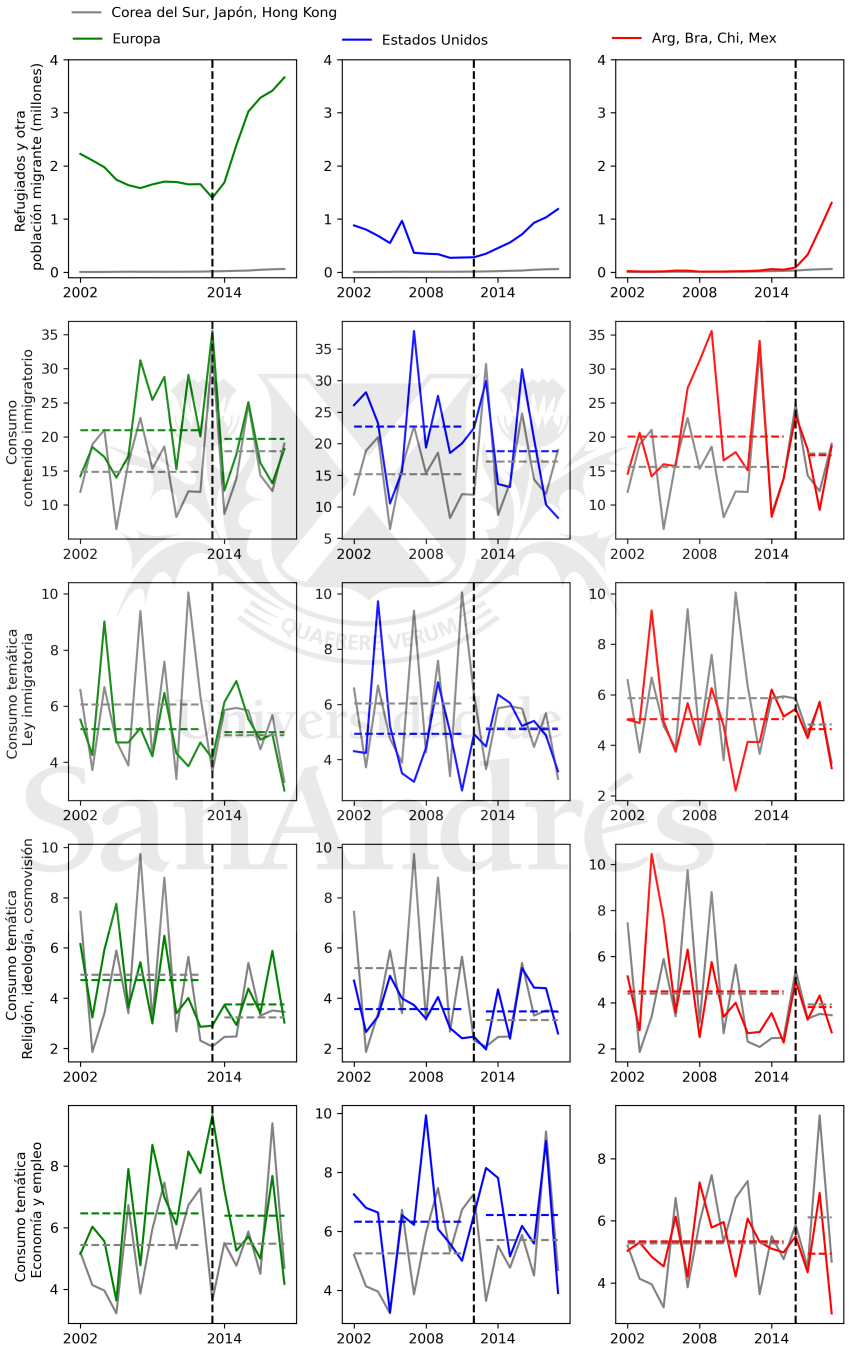


Figura 4.11: Asociación entre el consumo de contenido de inmigración en las películas y la recepción de refugiados y otra población migrante UNHCR, por región y año [PARTE II]

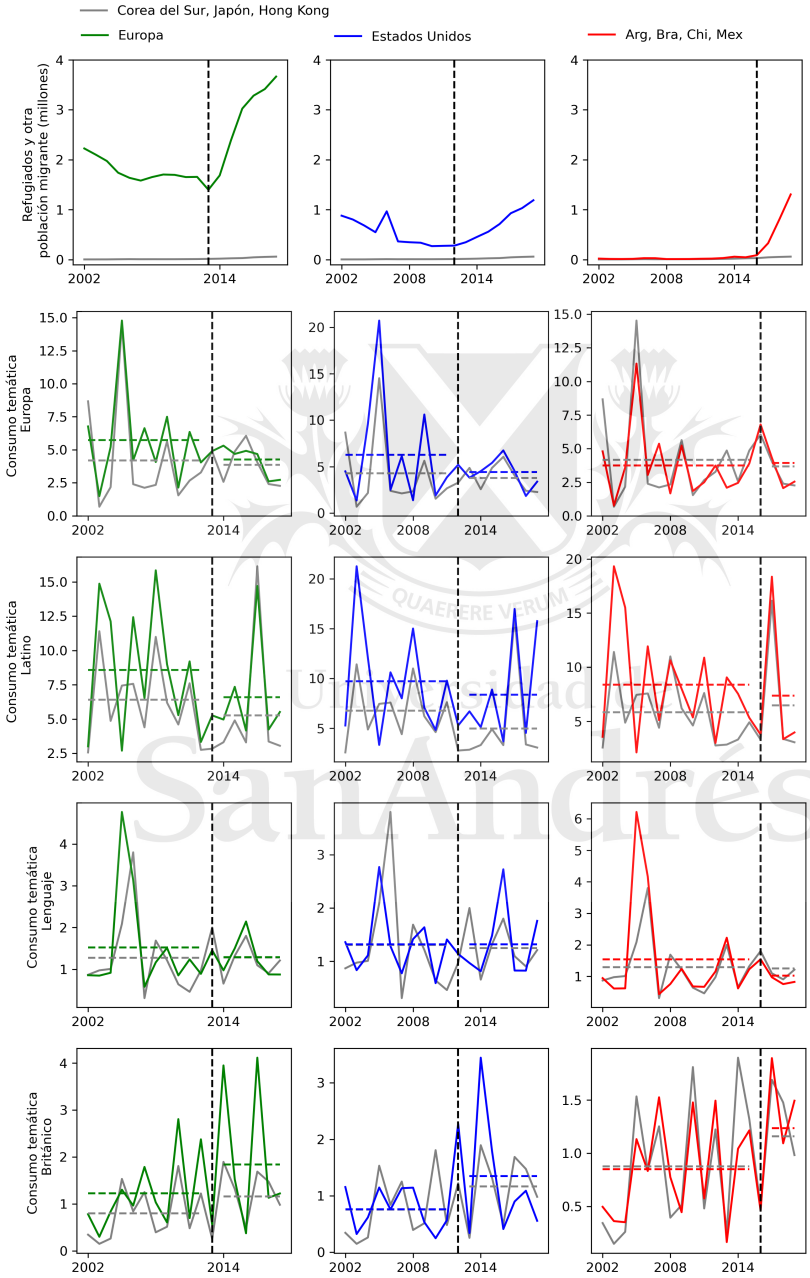


Figura 4.12: Regresión de consumo de contenido inmigratorio en cantidad de refugiados y otros migrantes, controlando por año y región.

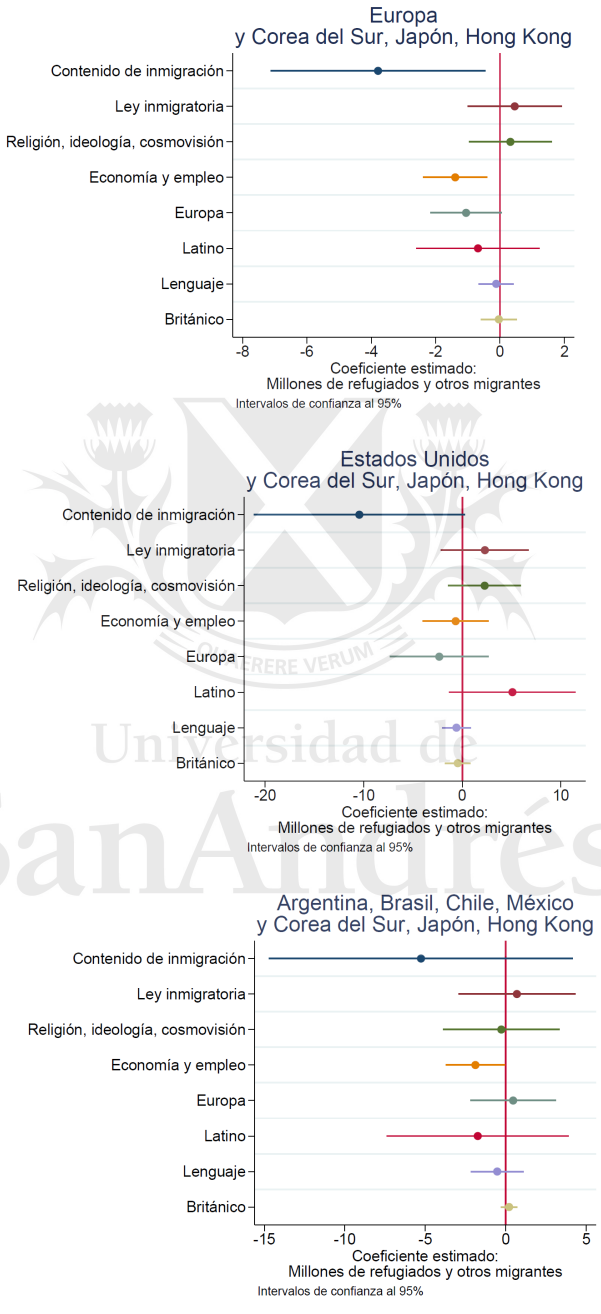


Figura 4.13: Regresión de consumo de contenido inmigratorio en la tasa de variación interanual de refugiados y otros migrantes, controlando por año y región.

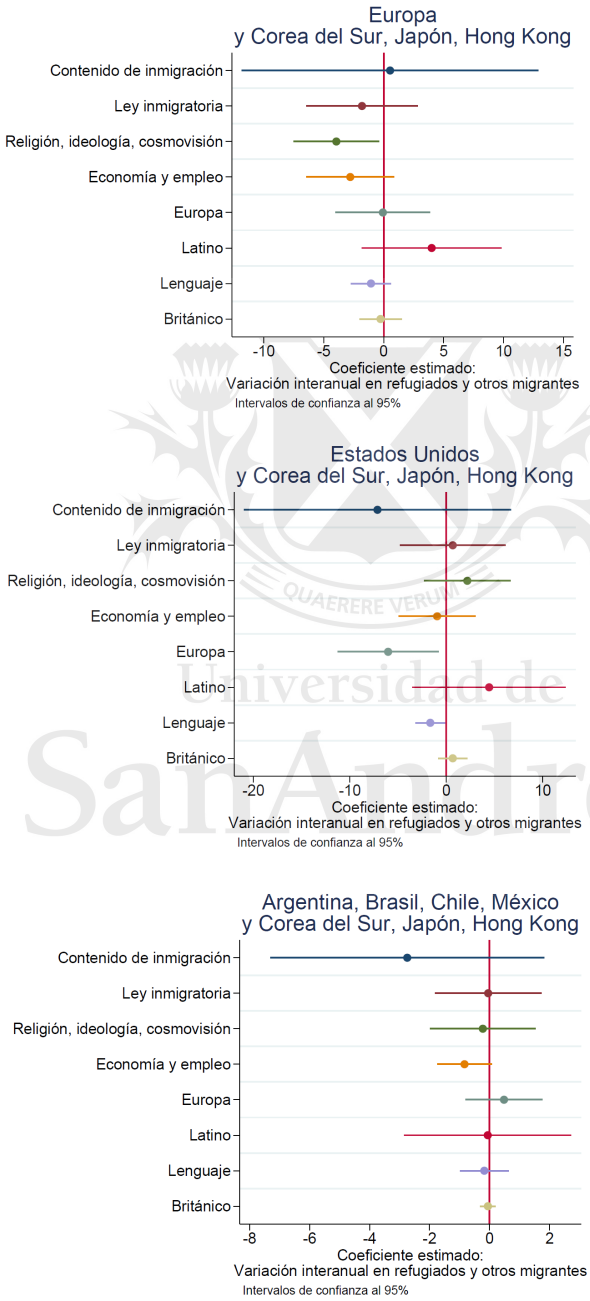


Figura 4.14: Variación del consumo de contenido de inmigración por país y quinquenio.

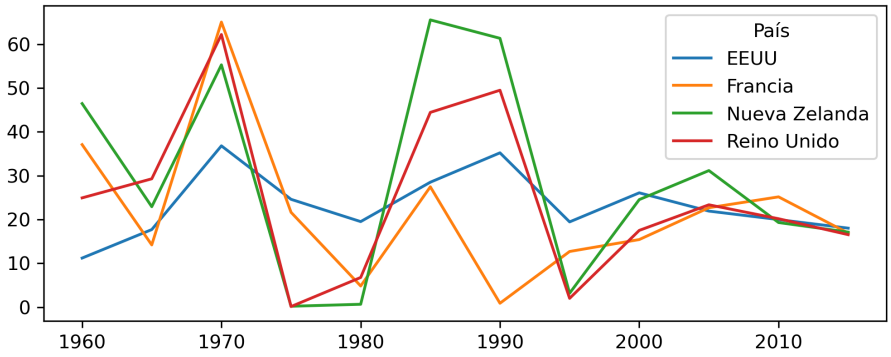
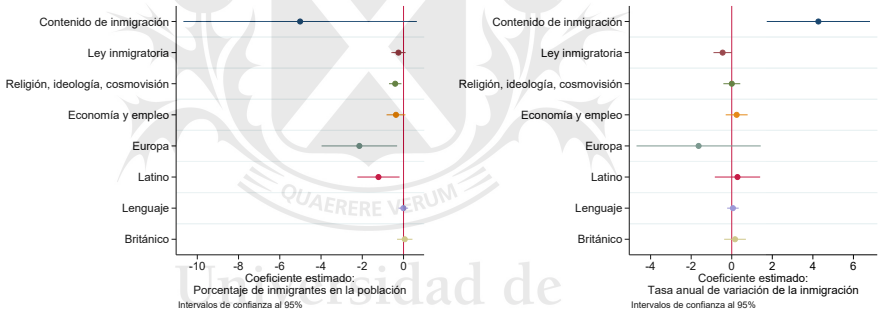


Figura 4.15: Regresión de consumo de contenido inmigratorio en la tasa de variación interanual de inmigración y en el porcentaje de inmigrantes respecto de la población, controlando por quinquenio y país.



(a) Porcentaje de inmigrantes

(b) Variación porcentual interanual en la llegada de inmigrantes

para cuatro países que tienen datos de taquilla para un período más largo de tiempo: Estados Unidos, Reino Unido, Nueva Zelanda, Francia. La evolución es similar entre los países. A partir de 1990, también hay datos de inmigración por quinquenio, y la Figura 4.15 muestra el resultado de regresar las medidas de consumo de contenido inmigratorio en el porcentaje de inmigrantes sobre la población total (a) y en la variación interanual en el stock de inmigrantes (b), controlando por país y quinquenio (ecuación 3.7). El porcentaje de inmigrantes se asocia de manera negativa y significativa (al 5 %) al consumo de ciertas temáticas de inmigración como Europa, Latino y Religión, ideología, cosmovisión en esos cuatro países. En cambio, una mayor tasa de variación interanual de la inmigración se asocia positiva y significativamente con el consumo de películas de contenido inmigratorio.

CAPÍTULO 5

A futuro

A continuación, se listan algunas limitaciones, alternativas o posibles mejoras para las distintas etapas del análisis. El capítulo primero presenta las mejoras posibles en la metodología (recorre las distintas secciones del Capítulo 3), y agrega dos secciones finales: una referida al uso de la etiqueta de IMDb como referencia para distinguir entre películas de inmigración y de no-inmigración, y otra con nuevas preguntas que podrían estudiarse partiendo del trabajo realizado.

5.1. Limitaciones, alternativas y mejoras metodológicas varias

Limpieza y preprocesamiento

En el Paso 2 de la limpieza (Sección 3.2), para identificar líneas que refieran al creador de los subtítulos o de las películas podrían usarse modelos predictivos; por el momento se utilizaron reglas predeterminadas para identificarlas. Además, las funciones de esta parte de la limpieza se armaron solamente para los subtítulos en archivos de tipo ".srt", que representan el 99 % de los subtítulos, y el resto fue descartado por tratarse de un porcentaje muy chico. Ahora bien, como mejora a futuro, puede adaptarse a otros tipos de archivos.

Exploración inicial

Una alternativa a usar la medida de similaridad provista por Sketch Engine para obtener un conjunto de términos similares a "immigrant", "migrant", "migration", "immigration" sería basar la similaridad en la distancia coseno usando algún modelo *word embeddings*. Se podría calcular la distancia coseno de todos los tokens del subtítulo.

Temáticas de inmigración usando clústers

Los clústers de inmigración se armaron partiendo del conjunto de lemas únicos de los subtítulos. Partir del conjunto de lemas en vez del texto completo es útil para resumir la información de las palabras y capturar su campo semántico. Ahora bien, a futuro sería interesante pensar una estrategia que no parta de los lemas y pueda capturar otra información de las palabras, como el género, el plural o singular, o la conjugación verbal; es posible que las películas de

inmigración y no-inmigración también tengan diferencias en esos aspectos. Incluso podrían tenerse en cuenta otras características, como el momento en que se dice cada línea dentro de la película.

Se usó GloVe para la representación vectorial de los lemas, pues se filtró el vocabulario para incluir solamente aquellos lemas presentes entre los 6 mil millones de tokens de GloVe como estrategia de limpieza, con lo cual no quedaron lemas fuera de vocabulario (ver la Sección 3.2). Si no hubiera sido así, se podría haber usado fasttext [9], que calcula la representación vectorial sobre n-gramas de los caracteres de las palabras y de esa manera elimina el problema de contar con palabras fuera de vocabulario.

Además, la agrupación en clústers partiendo de la representación vectorial de GloVe tiene sesgos importantes, producto del conjunto de datos en el que se ha entrenado ese modelo de *embeddings*. Por ejemplo, "trafficker" o "undocumented" queda en el clúster de nombres y gentilicios latinos. De hecho, está estudiado que la dinámica temporal de los *embeddings* -qué palabras quedan más cercanas entre sí cuando se calculan las representaciones vectoriales usando textos de distintos momentos históricos- guarda una correspondencia con los cambios sociales, y algunos sesgos disminuyen mientras que otros aumentan a lo largo del tiempo [19]. En este trabajo se construyó una única definición de temáticas "atemporal" usando los textos de los subtítulos de todas las épocas y la cercanía semántica está influenciada por la época de los textos sobre los que se entrenó GloVe. Podría ocurrir que la propia definición de las temáticas varíe en distintas épocas históricas. Para analizar si esto es así, deberían entrenarse modelos de *embeddings* a partir de corpus o de los propios subtítulos en distintos momentos del tiempo.

Se usó Fast K-Medoids en vez de otras técnicas comúnmente usadas para análisis de tópicos como Análisis Semántico Latente (ASL) o Asignación Latente de Dirichlet (ALD) porque escala mejor para muchos datos y para aprovechar la información de las representaciones vectoriales de las palabras o lemas ya entrenadas a partir de corpus de muchísimos textos y que reflejan el contexto donde las palabras aparecen, por como están entrenados modelos de *embeddings*. En cambio, ASL y ALD representan a las palabras como una colección donde el orden y el contexto no importa (representación BOW). Ahora bien, como mejora a futuro, se podría tratar de combinar la técnica usada en este trabajo con algunas ventajas que sí tienen estos métodos:

- Una mejora sería poder asignar una probabilidad de pertenencia de cada palabra a cada clúster usando alguna técnica *fuzzy* a nivel de palabra (los clústers armados sí son "*fuzzy* a nivel de película"). Esto permitiría ordenar la importancia de los lemas en cada clúster. Una alternativa para hacerlo que se puede probar con el método actual es tomar la palabra central del clúster como la más importante, y las distancias a ese centro para definir la centralidad de cada lema en el cluster. Además, permitiría capturar mejor la polisemia a nivel de palabra. Vale aclarar que, dado que los clústers actuales se arman a partir de 10 repeticiones con distinta iniciación aleatoria, sí hay algunos lemas que pertenecen a más de un clúster, pero son muy pocos.

- Otra mejora sería usar la aparición de dos palabras en un mismo documento (en este caso, película) como medida de cercanía o similaridad semántica

entre ellas. En un extremo, podría entrenarse un modelo de *embeddings* exclusivamente a partir de los textos de las películas, o tomar los clústers obtenidos usando la matriz $TFIDF$ como representación vectorial de cada lema, y comparar los resultados obtenidos. Una opción intermedia para aprovechar la mayor cantidad de información posible sería calcular una nueva representación vectorial que pondere las representaciones vectoriales de cada lema obtenidas con un modelo de *embeddings* externo por una medida de similaridad basada en la ocurrencia conjunta de los lemas en las películas, por ejemplo:

$$TFIDF^{T}_{F\times L} \cdot TFIDF_{F\times L} \cdot L2V_{L\times 300}$$

De hecho, estas representaciones podrían calcularse por separado para las películas de diferentes épocas históricas, y de esa manera capturar temáticas cuya misma definición varíe a lo largo del tiempo.

La asignación manual final de un nombre a cada clúster podría hacerse usando más anotadores.

Para asignar un valor en cada clúster (es decir, construir la matriz $F2C$) se sumó los valores en la matriz $TFIDF$ de todos los lemas en esa película y clúster (ver ecuación 3.2) . Otra opción sería usar el promedio en vez de suma de valores de la matriz $TFIDF$, lo que resolvería el hecho de que los clústers con más palabras tendrán valores más altos en promedio.

Finalmente podrían entrenarse modelo predictivos más complejos para predecir el valor de cada película en cada clúster, como RoBERTA. Una opción alternativa, sería *promptear* grandes modelos de lenguaje como GPT-4 para clasificar el resto de las películas en clústers a partir de los títulos de los clústers y un par de ejemplos de películas que pertenecen a cada clúster (*few-shot learning*). Otra opción sería desarrollar alguna asignación de forma mecánica a partir de la cercanía de los lemas de las nuevas películas a los lemas de los clústers ya definidos, aunque siempre teniendo en cuenta las dos restricciones propuestas: (i) no alcanza con calcular la pertenencia a un clúster considerando exclusivamente los lemas que el clúster contiene según se definió cuando se construyó inicialmente, porque podría suceder que una nueva película no tenga ninguno o tenga pocos de los lemas de un clúster pero sea semánticamente cercana, (ii) no debe asignarse el valor a partir de una matriz $TFIDF$ que requiera usar el contenido del resto de las películas con las que se definieron los clústers iniciales. Una mejora importante a futuro sería poder entrenar y evaluar estos modelos usando películas por fuera de la muestra con la que se construyan originalemnte las temáticas, para no sobreestimar su desempeño.

Índice de inmigración usando clasificación

La principal mejora en la construcción de un índice de inmigración usando clasificación sería mejorar la comparabilidad entre los clasificadores entrenados a partir de $F2V$ y RoBERTa (ver Anexo B).

También se podría probar y comparar con el uso de *zero-shot* o *few-shot learning* en grandes modelos de lenguaje como forma de predecir el grado de contenido de inmigración.

Finalmente, otra forma alternativa para medir el contenido de inmigración en las películas sería usar métricas de la correspondencia de los subtítulos con

otros corpus de inmigración, como el corpus COMMIRE [17]. Una limitación de este enfoque alternativo es que las películas pueden tratar la temática de inmigración de manera distinta a como es tratada en textos de otro género.

Aplicación: métricas de consumo de contenido inmigratorio

En este trabajo se presentó un análisis exploratorio inicial de asociación entre la inmigración real en distintos países y regiones receptoras y el consumo de contenido inmigratorio. Deberían especificarse mejor los modelos -por ejemplo, los rezagos temporales relevantes- y desarrollarse mejor los supuestos para que dicho análisis tenga interpretación causal, esto es, que no sean correlaciones espúrias debido a factores inobservados que afectan tanto a la inmigración real como al contenido inmigratorio.

5.2. Etiqueta de inmigración "verdadera"

Se muestra en la Sección 4.3 que la etiqueta de IMDb parece ser imperfecta. Para mejorar el etiquetado, algunas opciones son:

- No usar sólo la búsqueda por palabras clave de IMDb sino también otras fuentes (por ejemplo, películas de inmigración según Wikipedia, listados en otras plataformas como Letterboxd).
- Anotadores manuales, a los que se les asigna un subconjunto aleatorio de películas para etiquetar. Incluso podría aprovecharse para crear etiquetas continuas. Opciones para hacerlo: armar una encuesta con slider del 1 al 10, definir la probabilidad de que sea de inmigración según el grado de acuerdo entre anotadores, presentarles películas de a pares y que marquen sucesivamente cual tiene más contenido inmigratorio para construir dicha medida.
- Usar grandes modelos de lenguajes como anotadores. Un problema que tiene esta opción es que no se sabe cómo construyen la etiqueta ni con qué datos y por lo tanto pueden (i) introducir sesgos respecto de qué películas se clasifican o no como de inmigración o (ii) haber usado los mismos datos que luego se quieren usar para el análisis, lo que sesga los resultados de desempeño de los modelos.

5.3. Nuevas preguntas

Hay muchas otras preguntas interesantes a explorar, como por ejemplo:

- Comparar el tratamiento de la temática de inmigración en distintos subgéneros de películas. En particular, sería interesante ver el tratamiento de estas temáticas en películas infantiles.
- Analizar en qué momento de la trama las películas se vuelven más intensivas en contenido de inmigración. Por ejemplo, usando un modelo entrenado a nivel de grupo de líneas como RoBERTa, se podría graficar la relación entre el momento de la película (usando las marcas temporales de los subtítulos) y la probabilidad promedio de contenido de inmigración.

- Comparar la trama de las películas de inmigración versus de no-inmigración, y de las películas que tratan la inmigración desde distintas temáticas o perspectivas. Para eso, una opción posible es aprovechar las marcas temporales y hacer *sentiment analyisis* a medida que avanza la película y mostrar los arcos narrativos promedio de cada tipo de películas, al modo de [47].
- Comparar el tratamiento de la temática de inmigración en películas de distintos orígenes e idiomas.
- Analizar la relación entre el consumo de contenido de inmigración y la salida de emigrantes (en vez de la llegada de inmigrantes, es decir, por país emisor en vez de por país receptor). Analizar la relación con distintos tipos de inmigración: refugiados por guerras versus por motivos económicos, de distintos países de origen, y en distintos momentos del tiempo.
- Una importante limitación de la medición del consumo de películas de inmigración es que se limita a los consumos en el cine, debido a restricciones en los datos. A futuro, sería interesante recopilar datos para medir el consumo de las películas en plataformas de *streaming*.

CAPÍTULO 6

Conclusiones

Caracterizar el contenido de inmigración en las películas es relevante para entender las representaciones sociales de los inmigrantes en distintas épocas, que pueden verse reflejadas en y a la vez ser influenciadas por los consumos culturales. Este trabajo explora si hay temáticas sistemáticamente asociadas a la inmigración en las películas, cuáles son, y su importancia en cada película, y cuantifica el contenido de inmigración de más de 27 mil películas a partir de sus subtítulos, usando técnicas de procesamiento de lenguaje natural y una combinación de modelos de aprendizaje automático supervisados y no supervisados.

Desde ya, hay parte del contenido de inmigración en las películas que por definición los subtítulos y su significado literal no pueden capturar, como la comunicación a través de las imágenes, el lenguaje corporal, las expresiones y la emocionalidad que acompañan a los diálogos. Ahora bien, los resultados muestran que el texto de los subtítulos sí tiene información relevante para distinguir entre películas de inmigración y de no-inmigración, y para definir distintos ejes temáticos desde los cuales se trata la inmigración en las películas.

Primero, la combinación de los métodos de clustering jerárquico, Random Forest y Fast K-Medoids a partir de la matriz de distancias coseno de las representaciones vectoriales unos 27 mil lemas únicos presentes en el corpus de subtítulos permiten encontrar temáticas sistemáticamente asociadas a la inmigración y con un sentido semántico claro: contenido asociado a lo británico, a los conflictos de medio oriente, a la economía y empleo, a referencias a Europa y otros lugares del mundo, a lo latino, al lenguaje, a la ley inmigratoria, al nazismo, a Nueva York y Estados Unidos, a la religión, ideología o cosmovisión, a la tecnología. Cambiando la iniciación aleatoria y el número de clústers elegido inicialmente se encuentran temáticas similares. Algunas tendencias históricas que se observan son el pico del contenido referido a Europa y al nazismo en los años '40, un ascenso del contenido asociado a lo latino y a la tecnología en los últimos años, y un descenso de las referencias a Nueva York y Estados Unidos y a la economía y el empleo.

Segundo, el uso de un modelo de aprendizaje supervisado interpretable como la Regresión Logística permite cuantificar el contenido de inmigración de las películas a partir de la probabilidad predicha por el modelo de que una película sea de inmigración. Aunque predice muchas más películas de inmigración que las identificadas como tales por IMDb, lo cierto es que algunas de esas películas adicionales ("falsos positivos") en realidad sí parecen contener alto contenido inmigratorio si se hace una verificación manual. Esto es, el modelo

predictivo parece resumir la información de los subtítulos referida a contenido de inmigración de forma tal que permitiría mejorar el etiquetado inicial. El *finetuning* de RoBERTa permite predecir películas de inmigración de forma muy precisa. En ambos casos, usar aprendizaje automático mejora enormemente la capacidad de discernir entre películas de inmigración y de no-inmigración respecto de reglas de decisión fijas *ad-hoc*, por ejemplo basadas en la presencia de determinadas palabras en los subtítulos.

Tercero, la exploración inicial de la asociación entre el consumo de películas con contenido de inmigratorio en cines y las dinámicas reales de recepción de inmigrantes en distintas épocas y geografías arroja resultados mixtos; se encuentra una asociación negativa en la mayoría de los casos.

Finalmente, se han listado en detalle las limitaciones, mejoras metodológicas y nuevas preguntas a explorar a futuro partiendo del trabajo realizado.

Bibliografía

[1] Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470.*

[2] Miserlis Hoyle, A. M., Goel, P., Sarkar, R. y Resnik, P. (2022). Are Neural Topic Models Broken?. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[3] Allen, W. y Blinder, S. (2013). Migration in the News: Portrayals of Immigrants, Migrants, Asylum Seekers and Refugees in National British Newspapers, 2010-2012. *Migration Observatory.*

[4] Ballesteros, I. (2015). Immigration cinema in the New Europe. *Intellect.*

[5] Banerjee, A., La Ferrara, E., y Orozco, V. (2019). Entertainment, education, and attitudes toward domestic violence. *AEA Papers and Proceedings.* Vol. 109, pp. 133-137.

[6] Banerjee, A., La Ferrara, E., y Orozco-Olvera, V. H. (2019). The entertaining way to behavioral change: Fighting HIV with MTV. *National Bureau of Economic Research.* No. w26096.

[7] Bansak, K., Hainmueller, J.,y Hangartner, D. (2023). Europeans' support for refugees of varying background is stable over time. *Nature*, 620(7975), 849-854.

[8] Beltagy, I., Peters, M. E., y Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150.*

[9] Bojanowski, P., Grave, E., Joulin, A. y Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606.*

[10] Bougiatiotis, K., y Giannakopoulos, T. (2016). Content representation and similarity of movies based on topic extraction from subtitles. *Proceedings of the 9th Hellenic Conference on Artificial Intelligence* (pp. 1-7).

[11] Byrne, C., Horak, D., Moilanen, K. y Mabona, A. (2022). Topic Modeling With Topological Data Analysis. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11514–11533, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[12] Cape, G. S. (2003). Addiction, stigma and movies. *Acta Psychiatrica Scandinavica*, 107(3), 163-169.

[13] Chao, B., y Sirmorya, A. (2016). Automated movie genre classification with LDA-based topic modeling. *International Journal of Computer Applications*, 145(13), 1–5.

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

[15] Dahl, G., y DellaVigna, S. (2009). Does movie violence increase violent crime?. *The Quarterly Journal of Economics*, 124(2), 677-734.

[16] Erjavec, Tomaž; et al., 2021, Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1, *Slovenian language resource repository CLARIN.SI*, ISSN 2820-4042.

[17] Furtado, A. B. D., y Teixeira, E. D. (2022). Corpus Multilíngue sobre Migração e Refúgio (COMMIRE): planejamento, compilação e conteúdo, em linhas gerais. Texto Livre, 15, e36965.

[18] Gálvez, R. H., Tiffenberg, V., y Altszyler, E. (2019). Half a century of stereotyping associations between gender and intellectual ability in films. *Sex Roles*, 81(9-10), 643-654.

[19] Garg, N., Schiebinger, L., Jurafsky, D., y Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635-E3644.

[20] Geena Davis Institute on Gender in Media. (2015). The reel truth: Women aren't seen or heard. An automated analysis of gender representation in popular films. Recuperado de *seejane.org/research-informs-empowers/data/*.

[21] Goschenhofer, J., Ragupathy, P., Heumann, C., Bischl, B. y Aßenmacher, M. (2022). CC-Top: Constrained Clustering for Dynamic Topic Discovery. *Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)*, 26–34, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[22] Gretz, S., Halfon, A., Shnayderman, I., Toledo-Ronen, O., Spector, A., Dankin, L., Katsis, Y., Arviv, O., Katz, Y. Slonim, N. y Dor, L. E. (2023). Zero-shot Topical Text Classification with LLMs-an Experimental Study. *Findings of the Association for Computational Linguistics: EMNLP 2023* (9647-9676).

[23] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

[24] Hainmueller, J., t Hopkins, D. J. (2014). Public attitudes toward immigration. *Annual Review of Political Science*, 17, 225-249.

[25] Haluszka E, Niclis C, Pareja Lora A, Aballar LR (en prensa). Application of Natural Language Processing for the recognition of obesity-related topics in the discourses of Argentine Twitter users. *Lodz Papers in Pragmatics*.

[26] Hasan, M. M., Dip, S. T., Kamruzzaman, T. M., Akter, S., y Salehin, I. (2021). Movie Subtitle Document Classification Using Unsupervised Machine Learning Approach. *2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)* 219-224. IEEE.

[27] Hastie, T., Tibshirani, R., Friedman, J. H., y Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* .

[28] Hesham, M., Hani, B., Fouad, N., y Amer, E. (2018). Smart trailer: Automatic generation of movie trailer using only subtitles. *2018 First International Workshop on Deep and Representation Learning (IWDRL)* 26-30. IEEE.

[29] Información cortesía de IMDb (https://www.imdb.com). Usada con permiso.

[30] Islentyeva, A. (2020). *Corpus-based analysis of ideological bias: Migration in the British press.* Routledge.

[31] Jacobsen, G. D. (2011). The Al Gore effect: an inconvenient truth and voluntary carbon offsets. *Journal of Environmental Economics and Management*, 61(1), 67-78.

[32] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., y Suchomel, V. (2013). The TenTen corpus family. *7th International Corpus Linguistics Conference CL* (pp. 125-127).

[33] James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning* New York: Springer.

[34] Jurafsky, D., y Martin, J. H. (2024). *Speech and language processing (draft of February 3, 2024)*. Chapter 10: Transformers and Large Language Models. Último acceso el 4 de febrero, 2024.

[35] Jurafsky, D., y Martin, J. H. (2024). *Speech and language processing (draft of February 3, 2024)*. Chapter 11: Fine-Tuning and Masked Language Models. Último acceso el 4 de febrero, 2024.

[36] Kaufman, L. y Rousseeuw, P. J. (1990). Partitioning Around Medoids (Program PAM). *Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 68–125.

[37] Kim, K., y Lee, Y. (2023). DRAFT: Dense Retrieval Augmented Few-shot Topic classifier Framework. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2278-2294.

[38] Lison, P. y Tiedemann, J. (2016) OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC-2016), 2016.

[39] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. y Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint* arXiv:1907.11692

[40] Malik, M., Hopp, F. R., y Weber, R. (2022). Representations of Racial Minorities in Popular Movies: A Content-Analytic Synergy of Computer Vision and Network Science. *Computational Communication Research*, 4(1).

[41] Mangolin, R. B., Pereira, R. M., Britto Jr, A. S., Silla Jr, C. N., Feltrim, V. D., Bertolini, D., y Costa, Y. M. (2022). A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, 81(14), 19071-19096.

[42] Matthews, P., y Glitre, K. (2021). Genre analysis of movies using a topic model of plot summaries. *Journal of the Association for Information Science and Technology*, 72(12), 1511-1527.

[43] Mintz, S. Historical Context: Movies and Migration. *The Gilder Lehrman Institute of American History.* https://www.gilderlehrman.org/history-resources/teaching-resource/historical-context-movies-and-migration. Descargado el 1 de abril de 2024.

[44] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., y Dehak, N. (2019). Hierarchical transformers for long document classification. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 838-844). IEEE.

[45] Pennington, J., Socher, R. y Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)* 1532-1543.

[46] Ramakrishna, A., Martínez, V. R., Malandrakis, N., Singla, K., y Narayanan, S. (2017). Linguistic analysis of differences in portrayal of movie characters. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vol. 1, pp. 1669-1678.

[47] Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., Y Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. E*PJ Data Science*, 5(1), 1-12.

[48] Ren, C., y Bloemraad, I. (2022). New Methods and the Study of Vulnerable Groups: Using Machine Learning to Identify Immigrant-Oriented Nonprofit Organizations. *Socius*, 8.

[49] Sarkar, S., Feng, D.,y Santu, S. K. K. (2023). Zero-Shot Multi-Label Topic Inference with Sentence Encoders and LLMs. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 16218-16233.

[50] Schubert, E., y Rousseeuw P.J. (2019). Faster k-Medoids clústering: Improving the PAM, CLARA, and CLARANS Algorithms. *12th International Conference on Similarity Search and Applications (SISAP 2019)*, 171-187.

[51] Schubert, E., y Rousseeuw P.J. (2021). Fast and Eager k-Medoids clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms. *Information Systems* (101) 101804.

[52] Sia, S., Dalmia, A., y Mielke, S. J. (2020). Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914.*

[53] Siegel, A. A., Laitin, D., Lawrence, D., Weinstein, J., y Hainmueller, J. (2022). Tracking Legislators' Expressed Policy Agendas in Real Time.

[54] Stammbach, D., Zouhar, V., Hoyle, A., Sachan, M.,y Ash, E. (2023). Revisiting Automated Topic Model Evaluation with Large Language Models. *arXiv preprint arXiv:2305.12152.*

[55] Steele, L. G., y Abdelaaty, L. (2019). Ethnic diversity and attitudes towards refugees. *Journal of ethnic and migration studies*, 45(11), 1833-1856.

[56] Suchomel, V. Better Web Corpora For Corpus Linguistics And NLP. (2020). Doctoral thesis. Masaryk University, Faculty of Informatics, Brno. Supervised by Pavel Rychly.

[57] Tiedemann, J. (2012), Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012).

[58] *The British National Corpus*, version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

[59] Thompson, L., y Mimno, D. (2020). Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626.*

[60] United Nations Department of Economic and Social Affairs, Population Division (2020). *International Migrant Stock 2020.*

[61] United Nations High Commissioner for Refugees (2023). *UNHCR Refugee Population Statistics Database.* Base actualizada al 24 de octubre de 2023. Datos extraídos el 4 de diciembre de 2023.

[62] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. y Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[63] von Boguszewski, N., Moin, S., Bhowmick, A., Yimam, S. M., y Biemann, C. (2021). How hateful are movies? a study and prediction on movie subtitles. *arXiv preprint arXiv:2108.10724.*

[64] Wang, F., Beladev, M., Kleinfeld, O., Frayerman, E., Shachar, T., Fainman, E., Lastmann Assaraf, K., Mizrachi, S. y Wang, B. (2023). Text2Topic: Multi-Label Text Classification System for Efficient Topic Detection in User Generated Content with Zero-Shot Capabilities. *arXiv preprint arXiv:2310.14817, aceptado en EMNLP 2023.*

[65] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, p., Ravula, A., Wang, Q., Yang, L. y Ahmed, A. (2020). Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33, 17283-17297.

Anexo A: temáticas de inmigración

Lemas en los clústers finales (ver secciones 3.4 y 4.2):

- **Británico**: belfast bertie clare cork derry dublin edinburgh gaelic glasgow ifa ireland irish limerick london mayo paddington paisley scotland scots scottish shamrock ulster westminster
- **Conflictos medio oriente**: 9/11 abbas afar afghan afghanistan ankara arab ariel avi aviv baghdad barrier bedouin beirut benjamin bethlehem blair blockade border britain buildup bulldozer civilian closure coalition crescent curfew damascus deploy deployment destruction dispatch egyptian elon embargo fallujah galilee gaza gideon gulf halt har humanitarian hussein ina incursion indirect invasion iranian iraq iraqi iraqis israel israeli israelis jericho jerusalem jewish jordan kippur kurdish kurds kuwait lebanese lebanon levy libyan mass military moshe multinational occupation palestine palestinian palestinians plo pows rabbi reconstruction refugee saddam settler shalom sharon sinai somali strip sunni syria syrian tel terrorism transitional turkish u.s un urgent uri uzi war withdrawal yom
- **Economía y empleo**: accountant acquire adventurer advertise alderman architect auction baht banker banknote bargain bidder billionaire bookmaker brainchild brand branson businessman businesswoman buy buyer centime cheap cheaply citizen congressman currency dealer dinar diplomat distribute dollar donate earner edmond entrepreneur erm euro exclusive expensive farmer fetch filipino financier founder franc guilder industrialist invest ipo krona krone landowner lasalle lawyer lira lobbyist locally lombard lucrative magnate manufacture market maverick merge middleman millionaire mogul monopoly multimillionaire offload outbid owner ownership peddle peg peseta peso philanthropist politician pound privately procure profitable prominent property purchase rancher rand resale resell rmb rouble ruble rupee sale select sell shilling socialite speculator stake statesman sterling stockbroker swap takeover tycoon unload versus wealthy yen
- **Europa**: alba albania amsterdam ankara antwerp armenia attila austria austrian baden baltic belgian belgium berlin berliner bern bohemia bolshoi bonn boris brandenburg brussels bucharest budapest bulgaria bulgarian chancellor chechen chechnya cologne copenhagen crimea croatia cyprus czech czechoslovakia danish denmark dresden dusseldorf dutch

dynamo enlargement estonia estonian finland finnish fischer flanders frankfurt frisk georgia georgian german germany gothenburg greece halle hamburg hanover havana heidelberg helmut helsinki holland hungarian hungary iceland icelandic istanbul kazakhstan kazan kiev klaus kohl kovacs kremlin krone laszlo latvia latvian leipzig leningrad lithuania lithuanian luxembourg malta maltese mol montenegro moscow munich netherlands nordic norway norwegian nuremberg odessa oslo paris perm petersburg petra platz poland polish portugal prague putin republic reykjavik riga romania romanian rotterdam russia russian salzburg sas scandinavian seb seoul sofia stockholm stuttgart swede sweden swedish swiss switzerland timea transylvania tt ukraine ukrainian vienna vladimir vladivostok warsaw washington zurich

- **Gentilicios y lugares del mundo**: abroad afghans africans albania albanian albanians algerians america americans americas asia asian asians australians austrians autonomy balkan balkans baltic belgians belgrade bloc bosnia bosnian brazilians breakaway britain britons brits buffer canadians chechen chechnya colombians continent continental croatian cubans czechs danes deployment egyptians englishmen enlargement envoy ethnic eu europe european europeans filipinos france frenchmen gentile germans germany global globally greeks hemisphere hungarians independence internationally iranians italian italians italy japan koreans kosovo latin latinos luka mexicans middle migration milo montenegro nation nato nigerians norwegians notably outnumber parisians partition pows prevalent refugee republic ricans romanians russians sarajevo serb serbia serbian serbs sicilians spaniards territory thais trans transatlantic turks u.n un world worldwide yugoslavia

- **Latino**: acapulco adolfo ajax alberto alejandro alfonso alfredo alvarez alvaro amazon america american americas andalusia andes andres angel antonio aragon argentina argentine argentinean armando arroyo arturo aruba augusto az azul bahamas baja barbados barcelona batista belize benito beto bilbao blanco boa boca bolivia bolivian bravo bremen cadiz camilo cancun caribbean carlos castillo castro catalan cesar chavez chihuahua chile chilean chinchilla chivas clemente colombia colombian cordoba cortez costa cuba cuban delgado diaz diego domingo dominguez dominican duran ecuador edgar eduardo elias emery enrique ernesto escobar esp esteban facto felipe felix fernandez fernando figueroa fiji filipino flores francisco franco freddy fuentes gabriel garcia gaspar gil gomez gonzalez gonzalo granada gregorio grenada guadalajara guadalupe guatemala guerrero guevara guillermo gulf gustavo gutierrez guzman haiti hector hernandez herrera hidalgo hispanic honduras hugo ignacio immigrant ismael jacinto jaime jamaica javier jimenez joaquin jorge jose juan juanito juarez julio laredo latino leon lima liverpool lobo lopez lorena lorenzo lucio luis luna machu madrid majorca malaga manuel marbella marcos mariachi mariana mariano marin mario marquez martinez martinique matias mauricio medina mendez mendoza mexican mexicans mexico migrant miguel miranda molina mora morales moreno navarro nicaragua nogales oaxaca ortega ortiz pablo paco palma pamplona panama pancho papua paraguay paso paz pedro pena pepe perez peru peruvian philippine picchu pinto pique ponce puerto rafa rafael ramirez ramiro

ramon ramos raul rey reyes rica rican ricardo rico rios rivera roberto robles rodrigo rodriguez rojas romero ruben ruiz salamanca salazar salvador sanchez sandoval santana santiago sergio serrano seville sonora soto spanish suarez tabasco thorne tijuana toledo tomas tonga torres trinidad undocumented uruguay valdez valencia vallarta vargas vega venezuela veracruz vicente xavier zapata

- **Lenguaje** : afghani afrikaans alphabet arabic aramaic bengali bilingual braille calligraphy cantonese creole dialect english farsi fluent fluently greek hebrew hieroglyphic hindi ibn interpreter koran language malay mandarin mongolian oriental persian pidgin pronunciation punjabi quran sanskrit slavic swahili translation translator turk urdu vocabulary voiceover yiddish yom

- **Ley inmigratoria**: allow applicant apply apprehend asylum authorisation authority authorization authorize ban blacklist citizen citizenship clamp clearance clemency combatant confiscate consent cremate demolition deport deportation detain discharge discretion dispensation disqualify eavesdrop eligible emigrate entry evacuate evict exempt exemption expel expressly extradite fcc feds forbid forcibly grant hospitalize identification illegally immigrate immunity impound imprison incarcerate indefinite ins interpol interrogate lawful learner legally licence license modification nab objection obtain overstay passport permanent permission permit proceed prohibit prosecute register registration relocation renew renewal repatriate repatriation require requirement residency restrict restriction revoke temporary undocumented unlimited unseal unsupervised uproot valid visa waive waiver wiretap

- **Nazismo**: adolf aryan auschwitz concentration dachau doomsday extermination fascist german gestapo ghetto himmler hitler holocaust infamous klan klux nazi nazis neo pow puppet reich semitic semitism slave supremacist survivor swastika sympathizer vichy war wartime wehrmacht

- **Nueva York y Estados Unidos**: 9/11 abner adjacent administration aide airport albany alley anchorage angeles arroyo atlanta avenue baltimore barack battleground bedford bergen blair borough boston boulevard brentwood bridgeport bronx brooklyn brownsville buffalo bush bushwick busy camden capitol charlotte cheney chicago chinatown cincinnati city civic cleveland clinton coliseum columbus coney corner courthouse dallas denver detroit dole downtown durst eastside flatbush fulton gentrification george gonzales gore gov greenwich hadley hampton harlem harrisburg hartford havana hillary hoboken houston immigrant indianapolis jacksonville jeb jersey kerry lafayette laguardia liberty los louisville madison mall manhattan mccain memphis metropolitan miami midtown milwaukee minneapolis montreal nashville nassau neighborhood newark newt nhl nyc oakland obama orlando orleans ot philadelphia philly phoenix pittsburgh plaza policy portland powell precinct predecessor president providence proximity quincy reagan redevelopment republicans residential riverdale rochester rollins rove rudolph scarsdale schroeder schwarzenegger scranton seaport seattle section sidewalk skyline skyscraper speech square staten storefront street suburban sunset surrounding

syracuse tampa toboggan toronto trenton upstate uptown urban venue w. washington waterfront weld westchester white whitman williamsburg ymca yonkers york

- **Religión, ideología, cosmovisión**: abstain activism afghans afterlife agnostic alchemy alike apocalyptic apparition arabs armenian armenians astral astrological astrology atheist belong biblical buddhists capitalism capitalist catholic catholicism catholics celestial christian christianity christians communion communism constitute contemplative cosmic creed critique demonic devout doctrine dogma earthly enlightenment esoteric ethnicity faithful fanatical fascism feminism feminist filipino godly haram healer heresy hierarchy hindu hindus hypnotism idealism ideological ideology immortality indigenous intangible islam jehovah jew jews judaism kurds legacy magical mainstream manifesto martyr martyrdom marxism marxist materialism mecca mentality metaphorical metaphysical mindset minority miraculous mormons mosque muslim muslims mysterious mystic mystical mysticism mythical mythological negroes observant occult orthodox outnumber pacifist paranormal partisan pharisees philosophical philosophy platonic potion preach predominantly primordial propaganda prophecy prophet prophetic protestant protestants psychic quasi ramadan reactionary realm rebirth reformation religion religious resurrection rhetoric rigid secular sharia sikh sikhs sisterhood socialism spiritual spiritualist spirituality squarely stereotype sunni supernatural superstition symbolism telekinesis telepathic telepathy teleportation templar tendency transcend transcendent transformative unorthodox vocation zionist
- **Tecnología**: absorption altimeter ambient antenna beacon beam binoculars brightness camera cctv cryogenic detect detection detector electromagnetic finder flashlight gps gravitational holographic illumination infrared invisible ir kev laser microwave navigation neutron optic optical passive photon prism projector radar radiation readout rearview seeker sensor shutter sighting sonar speckle spectral spectrum strobe tactile telescope thermal transmit transmitter ultraviolet uplink uv wavelength zoom

Anexo B: *finetuning* RoBERTa

Metodología

RoBERTa es un modelo de lenguaje basado en redes neuronales con la arquitectura *transformer* bidireccional, entrenado para la tarea de predicción de tokens enmascarados en el texto [35, 39]. Un elemento central de esta arquitectura es el uso de un mecanismo de autoatención que indica cómo combinar la representación de una palabra en la capa anterior con información de las representaciones de las palabras vecinas, para computar una representación contextualizada de la palabra en la siguiente capa. Es decir, las representaciones del significado de cada palabra incorporan información sobre la relación entre palabras en una ventana arbitrariamente larga del texto [34]. El *transformer* usado en RoBERTa es bidireccional porque el mecanismo de autoatención toma en consideración información del contexto de la palabra tanto a su izquierda como a su derecha. El modelo RoBERTa preentrenado usado (https://huggingface.co/FacebookAI/roberta-base) tiene 12 capas y 125M parámetros y fue entrenado a partir de 5 corpus en inglés de distintos tamaños y dominios que totalizan más de 160 GB de texto descomprimido [39]. Luego, el *finetuning* toma como base la red aprendida en el modelo RoBERTa pre-entrenado y realiza unos pasos de entrenamiento adicionales. Ajusta los parámetros preentrenados agregando capas finales en la red para aprender alguna tarea específica. Esto permite comenzar con representaciones mucho más ricas del texto aprendidas por el modelo preentrenado, sin tener que entrenarlo desde el inicio (lo que sería demasiado costoso en cómputo y tiempo) y mejorar su desempeño en la tarea específica de interés, en este caso, clasificación [35].

Se usaron 402 películas de inmigración y una muestra aleatoria de 402 películas de no-inmigración entre el conjunto de películas finales (luego del paso 4 de la limpieza descrita en la Sección 3.2), pero tomando todo el texto del subtítulo luego de haber quitado las marcas propias de los subtítulos (es decir, luego del paso 2 de la Sección 3.2). A continuación, se agruparon de a 40 las líneas de cada película. Esto permite tener textos más largos e informativos, lo más cercano posible a tomar todo el texto de la película, pero sin superar el límite de *input* admitido por el modelo RoBERTa preentrenado que es de 512 tokens. Se obtiene así un dataset de aproximadamente 34700 textos (grupos de 40 líneas), y a cada cual se le asigna la etiqueta de inmigración según la película a la que pertenecen. A continuación, se dividieron los textos en tres conjuntos: entrenamiento, testeo y validación.

Para el *finetuning* se tomó el modelo pre-entrenado roberta-base de Hugging Face. Luego, se tokenizaron los textos usando el tokenizador del modelo pre-entrenado y como límite de número de tokens por texto se usó el percentil 75 del conjunto de entrenamiento, aplicando *padding* para los textos mas cortos que dicho límite. Partiendo de los parámetros del modelo pre-entrenado, se lo entrenó por 3 épocas más, con *batches* de 48 textos y una tasa de aprendizaje de 2e−5. La elección del número de épocas y tasa de aprendizaje se hizo para evitar el sobreajuste a los datos de entrenamiento, basado en los resultados en el conjunto de testeo. Por eso, a la hora de evaluar el desempeño del modelo se usa el conjunto restante de validación.

El modelo permite predecir, para cada grupo de 40 líneas, si pertenecen a una película de inmigración o no. Para obtener una medida de contenido de inmigración a nivel de película, se calcula el porcentaje de grupos de líneas de los conjuntos de testeo y de validación clasificados como de inmigración en cada una de ellas, excluyendo del promedio los grupos de líneas en el conjunto de entrenamiento[1] .

Comparación entre modelos

Los clasificadores entrenados a partir de $F2V$ en 3.5 y el clasificador usando RoBERTa no son directamente comparables. A continuación se proponen opciones para mejorar la comparabilidad.

Hay dos grandes diferencias en como fueron entrenados que los vuelven incomparables:

- Para los primeros se usaron 27709 películas únicas y se sobremuestreó con SMOTE para solucionar el problema del desbalance. Esto se hizo así para aprovechar al máximo la información de todas las películas disponibles. Ahora bien, RoBERTa se entrenó y evaluó con solamente 804 películas: se partió de las 402 películas de inmigración disponibles y se submuestreó aleatoriamente el películas de no-inmigración. Esto se hizo así por dos motivos. Primero, por la mayor complejidad a la hora de introducir un sobremuestreo como SMOTE en el *finetuning* del modelo pre-entrenado. Para el análisis presentado, se usó el módulo de Transformers de Hugging Face que permite realizar el *finetuning* a partir del texto tokenizado. En cambio, para poder aplicar SMOTE sería necesario modificar directamente el input de la primer capa de neuronas en cada *batch*. El segundo motivo es que usar 34 veces más datos requeriría usar mucha más capacidad y tiempo de cómputo, que sobre todo escalaría a la hora de la predicción final, una vez entrenado el modelo. De cualquier modo, estos son cambios que podrían implementarse tratando de optimizar el código a futuro. Ahora bien, también debe tenerse en cuenta que no es de particular interés que el modelo prediga de forma perfecta la etiqueta de IMDb, si no que se busca capturar mayor contenido de inmigración, o de resumirlo mejor. Por lo tanto, el costo-beneficio de realizar estas mejoras era bajo, y hasta puede ser perjudicial para el objetivo buscado.

[1] Hay 5 películas que sólo tuvieron grupos de líneas en el conjunto de entrenamiento, que por lo tanto se excluyen del análisis.

Cuadro 1: Entrenamiento RoBERTa

Época	**Pérdida en datos de entrenamiento**	**Pérdida en datos de testeo**
1	0.648000	0.588645
2	0.494700	0.453681
3	0.345000	0.436659

Cuadro 2: RoBERTa - *balanced accuracy* en conjunto de validación pre versus post *fine-tuning*

RoBERTa base	**RoBERTa *finetuneado***
0.5	0.8

- Los resultados y la división en conjuntos de entrenamiento y testeo de los clasificadores entrenados a partir de $F2V$ son a nivel de película. En cambio, los resultados de RoBERTa son a nivel de grupo de 40 líneas y la misma película puede tener algunos grupos de líneas en el conjunto de entrenamiento y otros en el conjunto de testeo, lo que vuelve más parecidos a ambos conjuntos, dándole ventajas de desempeño a RoBERTa. Para poder usar el RoBERTa preentrenado necesariamente deben usarse como input textos más cortos (de máximo 512 tokens). La opción elegida fue dividir a las películas en conjuntos de líneas y entrenar al nivel de grupo de linea; opciones alternativa hubieran sido hacer una selección aleatoria de una parte del subtítulo de cada película, o entrenar un modelo que primero resuma el contenido de los subtítulos en un texto más corto. Ahora bien, una mejora simple a futuro, manteniendo el entrenamiento a nivel de grupos de 40 líneas, es modificar el reparto de los grupos de líneas entre los conjuntos de entrenamiento y testeo de modo de que sea a nivel de película, asignando los grupos de líneas pertenecientes a la misma película a un mismo conjunto. Esto resolvería el problema de la sobre-estimación del desempeño de RoBERTa para predecir si una película *nueva* es de inmigración o no. Otra opción sería utilizar modelos de *transformers* que permitan inputs de texto más largos, como *transformers* jerárquicos, Longformer o Big Bird [44, 8, 65].

Resultados

El modelo RoBERTa *finetuneado* para predecir si una película es o no de inmigración está públicamente disponible en Hugging Face (wenbrau/roberta-base_immifilms). El Cuadro 1 muestra la reducción de la pérdida en los datos de testeo en cada época del *finetuning*[2], y el Cuadro 2 muestra la mejora en el *balanced accuracy* en el conjunto de datos de validación antes y después del *finetuning*: el *finetuning* generó una mejora notoria.

Aunque no son directamente comparables (ver subsección previa), si se toma la clasificación de 799 películas usando RoBERTa versus su clasificación Regresión Logística, los modelos arrojan la misma predicción para el 71 % de

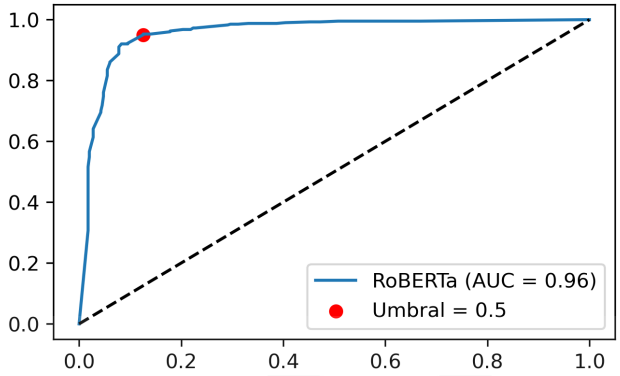
[2]Se probó aumentar la cantidad de épocas a 5 pero los resultados en testeo no mejoraban, sino que había un *overfitting* a los datos de entrenamiento.

las películas de inmigración y para el 81 % de las películas de no-inmigración, pero el desempeño de RoBERTa parece ser mejor. Identifica más películas de inmigración con menor cantidad de falsos positivos, y la diferenciación en los valores del índice de contenido de inmigración entre películas de inmigración versus de no inmigración es mayor (ver Figura 1 y comparar con Figuras 4.7 y 4.8)[3].

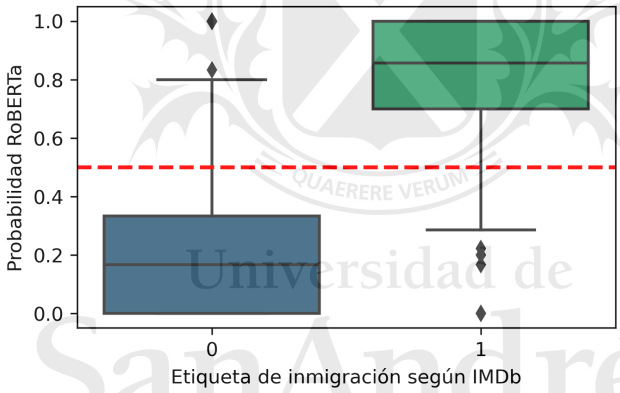
Una desventaja del uso de RoBERTa, es que la inferencia para toda la muestra de películas tomaría mucho más tiempo. Por otra parte, si se parte de la base de que la etiqueta de IMDb es imperfecta, no es más deseable un modelo que se ajuste mejor a dicho etiquetado. Se mostró cómo la Regresión Logística logra detectar con alta probabilidad películas de inmigración no etiquetadas como tales por IMDb pero que sí parecen serlo al hacer una verificación manual. Sería necesario hacer un nuevo etiquetado en el que sí se tenga plena confianza para basar la comparación del desempeño de los modelos (se detallan algunas opciones en este sentido en la Sección 5.2).

[3]Para la regresión logística, el 75 % de esas películas formaban parte de la muestra de entrenamiento, con lo cual el desempeño está sobrestimado. Para RoBERTa, como otras líneas de las mismas películas sí se usaron para el entrenamiento, su desempeño a la hora de clasificar películas también está sobrestimado (ver subsección previa).

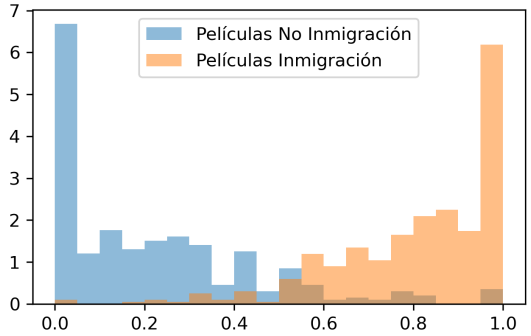
Figura 1: RoBERTa: curva ROC e índice de contenido de inmigración inmigración según etiqueta real



(a) Curva ROC



(b) Boxplots



(c) Histogramas