Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Code-switching en los foros latinos hispanohablantes de *Reddit*

AUTOR: JESSICA FEGGIO

TUTOR: MELISA ELFENBAUM

OCTUBRE 2023

Resumen

El objetivo principal de la inteligencia artificial es lograr un diálogo natural con los usuarios, lo cual requiere desarrollar técnicas de procesamiento del lenguaje natural que aborden el fenómeno del *code-switching*, es decir, la alternancia entre dos o más idiomas en una conversación. Esto es especialmente relevante para aplicaciones multilingües como asistentes virtuales o sistemas de traducción en tiempo real.

En este contexto, el tema propuesto a desarrollar en el presente trabajo es analizar los países de habla hispana de Latinoamérica en la red social *Reddit*[1]. El objetivo es identificar las variaciones en el uso de anglicismos entre estos países, analizar las tendencias temporales y evaluar el impacto de factores como el turismo de Estados Unidos y las relaciones comerciales.

Para lograrlo, se aborda el concepto de *code-switching* y *spanglish*, se describe el proceso de obtención de los datos y se presentan las técnicas a implementar. Luego, se analizan las tendencias de los anglicismos en los países, agrupándolos. Finalmente, se investiga la posible correlación entre la cercanía a Estados Unidos, la recepción de turistas estadounidenses y la frecuencia de anglicismos.

A través de la aplicación de técnicas de minería de datos, se busca obtener un indicador de las tendencias lingüísticas en las regiones analizadas, contribuyendo al desarrollo de modelos de lenguaje más precisos y robustos en contextos multilingües.

Palabras clave

Procesamiento de lenguaje natural, *Code-switching*, *Spanglish*, Minería de datos, *Reddit*

[1] Para más información: https://www.reddit.com/

Índice

Introducción

Con el transcurso del tiempo, ocurren cambios en la sociedad que son imposibles de evitar, debido a los continuos intercambios entre los diferentes países y sobre todo con la globalización, que ha transformado una parte considerable de la vida diaria. El idioma español no resulta incólume a ese fenómeno: en los periódicos, en las redes sociales, en el trabajo, como también en las conversaciones entre amigos, el idioma ha ido cambiando e incluye cada vez más anglicismos.

El *code-switching* es el fenómeno por el cual se cambia entre múltiples idiomas durante la comunicación oral o escrita, por ejemplo: “Tengo un *meeting* con mi jefe”. La importancia de desarrollar tecnologías del lenguaje capaces de procesar el lenguaje *code-switched* es inmensa, dada la influencia del idioma inglés en los programas televisivos, en la música y demás.

Uno de los objetivos de la inteligencia artificial es permitir un dialogo natural con los usuarios, para lograrlo es necesario desarrollar técnicas de procesamiento de lenguaje natural que tengan en cuenta el *code-switching*, siendo este un tema que todavía no ha sido muy investigado. El *code-switching* representa un reto, debido a que los modelos que están desarrollados para tratar un idioma se rompen al mezclarse con otro.

La finalidad de este trabajo es analizar la recurrencia, tipología y contexto de uso de anglicismos en la red social "Reddit" en los países de habla hispana en Latinoamérica durante el período 2016-2023.

Para poder llevar adelante el objetivo planteado, el trabajo se estructura en tres capítulos. En el primer capítulo, se tratará de la relevancia del idioma inglés en los países hispanohablantes de Latinoamérica. Se presentarán los datos más salientes sobre las relaciones comerciales entre Estados Unidos y las regiones interesadas, como también la cantidad de turistas procedentes del país norteamericano, ya que se trata de factores que se ha decido considerar para este trabajo. En la segunda parte se presentará el procesamiento de lenguaje natural y, finalmente, se abordará el tema del *code-switching* para ofrecer una visión general de esta ocurrencia.

En el segundo capítulo, se realizará el relevamiento de las publicaciones y su posterior análisis. Este relevamiento se llevará a cabo extrayendo datos de la red social *Reddit*, específicamente aquellas publicaciones que están relacionadas con la región que se decidió analizar. Para esto, se usará la API (*Application Programming Interface*) de *Reddit*. Amazon define las API como:

> [..] mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos. Por ejemplo, el sistema de software del instituto de meteorología contiene datos meteorológicos diarios. La aplicación meteorológica de su teléfono "habla" con este sistema a través de las API y le muestra las actualizaciones meteorológicas diarias en su teléfono. (Amazon, 2023)

Se ejecutará sobre un entorno de desarrollo usando el lenguaje *Python*[2] . Mediante el uso de técnicas de minería de datos se procesarán los datos obtenidos para una mejor interpretación y la detección de tópicos.

Finalmente, en el tercer y último capítulo se procederá a describir los diferentes modelos de *clústering* y se implementará uno para determinar si las relaciones comerciales con Estados Unidos y/o la llegada de 1.000.000 o más turistas procedentes de este país influyen en los resultados. Como último apartado, se realizará el análisis de los resultados obtenidos al aplicar métodos estadísticos.

[2] Lenguaje de programación orientada a objetos, para más información consultar: *https://www.python.org/*

Capítulo 1: Influencia del Idioma Inglés y *Code-Switching* en la Región

El idioma inglés ha adquirido una importancia sin precedentes en el ámbito global. Como lengua franca de los negocios, la tecnología y la cultura popular, el inglés se ha convertido en una herramienta esencial para la comunicación internacional y el intercambio cultural. Estados Unidos, como líder mundial en muchos aspectos, ha jugado un papel fundamental en la difusión y promoción del idioma inglés en todo el mundo. Esta tendencia se atribuye, en gran medida, a la presencia e influencia de Estados Unidos tanto en términos de su cultura como de su poder económico. Sin embargo, es importante destacar que este fenómeno no es exclusivo de la actualidad, sino que se ha venido intensificando a lo largo de los años.

Un testimonio que ilustra esta situación data de la década de 1950 y proviene de un profesor de español estadounidense llamado Savaiano (Savaiano, 1950). Durante su desempeño como docente de inglés en el Instituto bilingüe Panamericano, Savaiano constató que dicha institución contaba con una matrícula de más de 900 alumnos, lo que generaba la necesidad de rechazar a numerosos solicitantes debido a la demanda existente. Este hecho demuestra la creciente demanda de aprendizaje del inglés en aquel periodo. Asimismo, el profesor Savaiano relata su sorpresa al mudarse a Lima, donde observó el énfasis que los padres de sus alumnos daban al aprendizaje del inglés. Esto se debía a la importancia que les atribuían a las relaciones con Estados Unidos, ya sea para obtener mejores oportunidades laborales en el futuro o para perseguir estudios avanzados en dicho país.

Con la difusión de los anglicismos, se empiezan a notar contextos conversacionales en los que puede aparecer un cambio de código entre las conversaciones generadas en un contexto social monolingüe (Cantero & De Arriba, 1996). Los autores además afirman que el *code-switching* se utiliza principalmente para cumplir una función enfática, es decir:

> [..] llamar la atención del interlocutor sobre una idea, un medio de subrayar, de focalizar la porción de discurso codificada en otra lengua. La alternancia de código funciona, por lo tanto, como marcador de énfasis y como focalizador.
>
> A menudo, la alternancia de código consiste en una cita, que puede ser una frase hecha, un refrán, una porción conocida de otro discurso, etc. En tales casos el énfasis consiste en emplear como elemento discursivo recurrente un enunciado en otra lengua fijado y compartido por ambos interlocutores: un lugar común. Si la cita no

> forma parte del contexto compartido por los interlocutores, pierde toda efectividad y se convierte en un elemento innecesario y pedante. (p. 3)

En los siguientes subapartados se realizará un análisis general de la relevancia del idioma inglés en los países hispanohablantes de Latinoamérica y la influencia que Estados Unidos ejerce en la región. Además, se abordará en detalle el tema del lenguaje natural, destacando sus características distintivas y diversos casos de aplicación. Por último, se profundizará en el fenómeno del *code-switching*, resaltando su importancia en el contexto actual.

1.1. Influencia del Idioma Inglés y de EE. UU en la Región

Según el British Council (2018) 1,75 billones de personas hablan inglés, casi un cuarto de la población mundial; la mayoría de estas personas no son hablantes nativos, y su número supera por mucho a los que sí lo son. En 2021, Fishman destacó que, aunque el inglés es hablado como lengua materna por una población relativamente pequeña de aproximadamente 370 millones de personas, es el idioma predominante en la escritura de la gran mayoría de libros y artículos académicos a nivel mundial y observa:

> Tanto si consideramos el inglés una "lengua asesina", o si entendemos su difusión como una globalización benigna o como un imperialismo lingüístico, es innegable su alcance expansivo y, por el momento, imparable. En la historia de la humanidad nunca tanta gente había hablado (y no digamos medio hablar) una lengua de forma tan amplia. (Fishman, 2001, pág. 1)

A pesar de la importancia de este idioma, los resultados de las pruebas lingüísticas hechas por *EF Education First* en Latinoamérica revelan que los resultados de la región quedan en la categoría "bajo" o "muy bajo" a excepción de Argentina con un elevado puntaje de 58.40 a la par de países europeos o asiáticos como lo son Alemania (61.58), las Filipinas (60.33) y República Checa (59.09) (Salomé Guardione, 2019); (Cronquist & Fiszbein, 2017).

Carolina Quezada (2011) añade que el inglés es considerado como el idioma mundial de las telecomunicaciones, usado por un 80% de usuarios de Internet. En la actualidad, según la Fundación Universia (2020), más del 80% de las convocatorias de empleos para puestos de

medio rango y directivos tiene como requisito fundamental que el candidato hable una segunda lengua y en la mayoría de los casos corresponde al inglés. Las razones de la importancia y difusión de este idioma se deben no solo a lo laboral, sino también al mundo del entretenimiento, las mayores y más estrechas relaciones internacionales y sociales entre los países de habla española y los de habla inglesa y, por último, la enorme preponderancia económica y científica y política de los países anglosajones en el mundo contemporáneo. (Alfaro, 1948)

A pesar de que Argentina haya sido el país latinoamericano que obtuvo el puntaje más elevado en la encuesta de EF, según los datos de la Oficina del Representante Comercial de los Estados Unidos (2022), no se encuentra entre los países que tienen más relaciones económicas con EE. UU. Estos son:

- México: es el principal socio comercial de Estados Unidos en América Latina, con 614.5 mil millones de dólares en comercio de bienes (bidireccional) durante 2019. Las exportaciones de bienes totalizaron 256.6 mil millones de dólares.
- Colombia: es el vigésimo quinto mayor socio comercial de Estados Unidos en términos de comercio de bienes, con un comercio total de 28,9 mil millones de dólares en 2019. Las exportaciones de bienes ascendieron a 14,7 mil millones de dólares, mientras que las importaciones fueron de 14,2 mil millones de dólares.
- Chile: es actualmente el 29º socio comercial más grande de Estados Unidos en términos de comercio de bienes, con un total de 26.1 mil millones de comercio de bienes (bidireccional) en 2019. Las exportaciones de bienes sumaron 15.7 mil millones, mientras que las importaciones de bienes sumaron 10.4 mil millones.
- Perú: el comercio de bienes y servicios entre Estados Unidos y Perú ascendió a un total de 21.3 mil millones en el año 2019.

Los resultados son diferentes cuando se analiza la situación desde un punto de vista turístico, ya que los países que recibieron la mayor cantidad de turistas estadounidenses en 2019 según la Organización Mundial del Turismo (OMT) fueron México (36.9 millones); República Dominicana (2.7 millones); Costa Rica (1.3 millones); Colombia (810.000) y Panamá con 940.000 turistas recibidos (World Tourism Organization, 2019).

Es importante resaltar que México ocupa un lugar destacado en ambos aspectos mencionados. De manera significativa, según un análisis realizado en 2018, el 25% de los inmigrantes que residen en Estados Unidos nació en México, lo que representa el mayor porcentaje en comparación con otras regiones. En particular, en el caso de Asia, este porcentaje fue del 28%, mientras que para otros países latinoamericanos fue del 25%, y para otras partes del mundo fue del 9% (Budiman, Tamir, Mora, & Noe-Bustaman, 2020). Este dato no debe considerarse una mera casualidad, ya que refleja una conexión estrecha y continua entre México y Estados Unidos.

1.2. Procesamiento de Lenguaje Natural

El procesamiento de lenguaje natural (NLP)[3] combina modelos de lingüística computacional, de aprendizaje automático, y aprendizaje profundo para procesar el lenguaje humano. Este explora cómo se puede utilizar la tecnología, para entender y manipular lenguaje natural escrito o hablado y emplearlo de manera útil (Chowdhury, 2003). A través de la descomposición del texto en componentes (p. ej. adjetivos, verbos, sustantivos, etc.), estas técnicas computacionales permiten realizar diferentes tipos de análisis. A continuación, se tratarán algunos ejemplos dados por el blog de IBM *Cloud Education* (s.f):

- Reconocimiento de mensajes spam[4]: la clasificación textual es utilizada para analizar correos electrónicos y detectar la presencia de términos comúnmente utilizados en mensajes de spam. Esto incluye el uso de palabras en idiomas extranjeros o la presencia de errores ortográficos y gramaticales.
- Traducción automática: el uso de tecnologías como el *deep learning*[5] ha permitido desarrollar modelos capaces de ofrecer traducciones más naturales, como los que utiliza Google Traductor[6]. A diferencia de las primeras aproximaciones que se limitaban a traducir literalmente palabra por palabra, estos modelos son capaces de comprender el contexto y ofrecer resultados más precisos. Un ejemplo es la

[3] Acrónimo inglés: *Natural Language Processing*
[4] Anglicismo: mensaje no deseado
[5] Anglicismo: aprendizaje profundo
[6] Programa de traducción automática, para más información visitar: https://translate.google.com/. Véase también como ejemplo adicional, el programa de traducción automática de la Unión Europea, llamado *eTranslation*: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

metodología de traducción de cero fuentes (*zero-resource translation*) de Google, que permite traducir idiomas minoritarios para los que no hay muchos recursos (Caswell & Bapna, 2022).

- Asistentes virtuales y *chatbots*: los asistentes virtuales como Siri (asistente virtual de Apple)[7] o Alexa (asistente virtual de Amazon)[8] utilizan el reconocimiento de patrones lingüísticos aprendidos para responder a comandos de voz y llevar a cabo acciones específicas. Los *chatbots* programas de computadora diseñados para interactuar con seres humanos a través de una interfaz de chat como lo es ChatGPT[9]. Estas aplicaciones pueden simular una conversación con una persona utilizando lenguaje natural y responder a preguntas o proporcionar información.
- Análisis sobre "los sentimientos" en las redes sociales: con el objetivo de descubrir nuevos datos, el análisis "de sentimientos" analiza el lenguaje utilizado en las publicaciones, comentarios, reseñas y otros para extrapolar tendencias y emociones en respuesta a productos, promociones y eventos; datos que podrán ser aprovechados por las empresas.
- Resumen textual: para crear sinopsis de volúmenes de datos muy grandes.

En el ámbito empresarial, el procesamiento del lenguaje natural ha transformado significativamente el panorama de los negocios al derribar las barreras lingüísticas y promover operaciones internacionales, lo cual ha impulsado el comercio global. La implementación de esta tecnología ha brindado numerosos casos de uso, entre ellos, para la interacción con los clientes y los análisis empresariales. Para el primer caso, se puede aportar como ejemplo la creación de *chatbots*, que permiten resolver eficientemente los problemas de los clientes sin la necesidad de interactuar con un agente para cada consulta o duda y que además ofrecen sugerencias personalizadas basadas en palabras clave específicas. En cuanto la tipología de análisis que se pueden realizar gracias a estas técnicas, se puede mencionar el análisis de las reseñas de los productos, mediante el cual se pueden identificar sus defectos y cualidades. Así mismo, se puede emplear el análisis de las emociones en los comentarios de los clientes, para obtener una percepción informada sobre productos y/o servicios.

[7] Para más información: https://www.apple.com/siri/
[8] Para más información: https://developer.amazon.com/it-IT/alexa
[9] Para más información: https://openai.com/blog/chatgpt

Para realizar análisis de lenguaje natural y obtener resultados satisfactorios, es fundamental contar con fuentes lexicales adecuadas, como diccionarios y textos que sean legibles por las máquinas. En este sentido, el concepto de *corpus* desempeña un papel primordial.

Sierra Martínez (2015) define un *corpus* como un conjunto estructurado y organizado de textos escritos en un determinado idioma o dominio. Esta colección puede abarcar diversos tipos de documentos, como libros, artículos, periódicos, páginas web e incluso mensajes intercambiados entre individuos. Estos textos se recompilan y procesan con el objetivo de utilizarlos como una base de datos lingüística para el estudio y análisis del lenguaje. Se destaca que una biblioteca, sin importar su tipo, no constituye en sí misma un *corpus* debido a que el material no ha sido seleccionado aún con criterios bien definidos.

El más simple de estos conjuntos de texto, no contiene ninguna estructura, mientras que otros *corpus* más complejos se agrupan por categorías que pueden corresponder, entre otros al género, al autor o al idioma.

La más famosa de estas fuentes, es una colección de textos estructurados llamada *Brown corpus*, que contiene alrededor de un millón de palabras de la variante americana del idioma inglés, redactada en la universidad Brown entre los años 1960 y 1970 (Manning & Schütze, 1999).

En la era digital, no resulta sorprendente que la noción de "colección de textos" ya no se limite únicamente a textos físicos, como libros o periódicos. En cambio, muchas empresas, como *Twitter* o el *New York Times*, ofrecen la posibilidad de acceder a sus datos a través de APIs (Interfaces de Programación de Aplicaciones, por sus siglas en inglés), que actúan como espacios de almacenamiento de datos. En el caso de este trabajo en particular, se utilizará la API de la red social *Reddit*[10] para recopilar el texto necesario y cumplir con los objetivos establecidos.

Una vez que se ha encontrado un corpus adecuado, se inician las tareas de procesamiento, las cuales se dividen en cinco fases principales, que abarcan desde el análisis a nivel de palabras hasta el análisis de oraciones completas. Según el investigador Moreno de la Universidad Autónoma de Madrid (2022), se pueden distinguir diferentes tipos de análisis en el procesamiento del lenguaje natural.

En primer lugar, se encuentra el análisis morfológico o léxico, que se centra en las palabras que conforman las oraciones y permite extraer lemas, rasgos flexivos y unidades

[10] Para más información: https://www.reddit.com/dev/api/

léxicas compuestas, brindando información básica para el procesamiento. A continuación, se encuentra el análisis sintáctico, que se encarga de la estructura de las oraciones de acuerdo con el modelo gramatical utilizado. Por otro lado, se encuentra el análisis semántico, cuya función es interpretar las oraciones y eliminar ambigüedades morfosintácticas, proporcionando un significado más preciso. Por último, se menciona el análisis pragmático, que incorpora el contexto de uso en la interpretación final, teniendo en cuenta aspectos como el lenguaje figurado y el conocimiento específico necesario para comprender textos especializados.

Es importante destacar que no todos los análisis mencionados deben aplicarse en todas las situaciones, ya que las técnicas utilizadas dependen de los objetivos de la aplicación en cuestión.

1.3. Code-Switching

El *code-switching* es un fenómeno lingüístico que ocurre cuando una persona alterna dos o más idiomas en una misma conversación. Es particularmente frecuente en Europa y en la India, y se asocia a conversaciones informales o coloquiales como en las redes sociales o en las charlas entre amigos. Centero y De Arriba (1996) afirman que la intromisión de elementos en otro "código", término usado como sinónimo de "lengua", debida a errores del hablante, no deben considerarse *code-switching* y aportan un caso muy característico de falso cambio de código, el de un hablante extranjero en cuyo discurso aparecen elementos de su lengua propia sin que pueda evitarlo, porque no es consciente de ello. (p.1) Se puede entonces afirmar que el *code-switching* se trata en efecto de un modo bilingüe de comunicación.

Los autores además sugieren que el cambio de código puede tener distintas funciones discursivas, dependiendo del contexto social donde aparece, monolingüe o bilingüe. En un contexto monolingüe podría considerarse común la coexistencia de dialectos que generan situaciones de cambio de código, parecidas a las que ocurren en contextos sociales bilingües, siendo algunas regiones de Alemania o Italia buenos ejemplos de este caso, debido a que las variaciones dialectales difieren tanto como lo harían si fueran idiomas distintos. En contextos sociales bilingües, el code-switching es frecuente entre las lenguas en contacto; mientras que en países donde oficialmente se habla solo un idioma generalmente se produce con la influencia de un idioma extranjero. En un contexto monolingüe, el code-switching tiene una

función de énfasis, en un contexto bilingüe el cambio de código puede verificarse dependiendo del tema enfrentado o del interlocutor con quien se entra en contacto, y que puede ir cambiando en base al contexto. Se puede afirmar entonces, que la función principal que cumple en este ámbito es de tipo expresivo debido a que los hablantes cambian a su primera lengua en situaciones emocionales.

Múltiples investigaciones se han encargado de analizar este tipo de lenguaje mixto desde un punto de vista sociolingüístico. Según Poplack (1980) el *code-switching* es la alternancia de dos idiomas en una misma oración y es un indicador de capacidades bilingües. Muysken (2000) prefiere el termino *code-mixing* y se refiere a *code-switching* solo en los casos cuando coexisten múltiples lenguas en una oración. En este trabajo se decidió utilizar el término *code-switching* para referirse a cualquiera de estas instancias.

Existen diferentes formas de *code-switching*, las más comunes son el cambio intersentencial, que se refiere a un cambio de lengua entre dos oraciones consecutivas y el cambio intrasentencial, que indica una alternancia dentro de una misma oración.

Se puede suponer que casi el 100% de la población global perteneciente a la Generación Y o *Millennials* nacidos entre 1983-2000 y la Generación X o *Cenanarians* nacidos después el año 2000 son bilingües, lo cual hizo que se generaran y/o difundieran idiomas mezclados como lo son por ejemplo el *franglais*, el *runglish* o el *spanglish*. (Krasina & Jabballa Mahmoud, 2018)

Sobre el *spanglish*, la profesora Ana Celia Zentella (Hablacultura, s.f) afirma: "Todo inmigrante inmediatamente al llegar aprende unas palabras en inglés y empieza a adaptarse al léxico del inglés, pero no hacen esa alternancia creativa, que es la que nos distingue a nosotros, en la cual nos sentimos más cómodos. Yo, al hablar con usted ahora, me siento cómoda pero no soy Ana Celia completa. La Ana Celia completa habla inglés a veces, español a veces, pero con la gente con quien más comparto, más afines, hablo los dos." Como lo es Zentella, se puede asumir que hay muchos hablantes de *spanglish*. Solo en los Estados Unidos, más de 50 millones de personas hablan español (Perez, 2015) y entre los hispanos, el 59% son bilingües (Krogstad & Gonzales-Barrera, 2015). Este tipo de lenguaje se puede encontrar también en la política, un ejemplo es el *tweet*[11] publicado por la diputada Alexandria Ocasio-Cortez (2019):

[11] Mensaje publicado en Twitter que contiene texto, fotos, GIF o video. Para más información sobre Twitter: https://twitter.com/

Alexandria Ocasio-Cortez. (2019, diciembre, 18). *I'm nervous for this all-Spanish town hall, but I also know that the only way I'm going to improve my Spanish is by practicing it!*
Nevada: Únete a nosotros este Domingo para un... *town hall*(?) en español, y probablemente con un poquito de "*spanglish*" también. [Tweet]. Twitter. https://twitter.com/AOC/status/1207370145573867521?s=20

Aguilera, periodista de "El País", afirma que el uso del *spanglish* es difuso en el sector del entretenimiento. En las canciones que llegaron al top 100 de *Billboard*[12] de EE. UU., aumentó considerablemente desde la década de 1980, cuando se registraron solo 12 canciones que tenían esta mezcla, hasta la década de los 2000 cuando aparecieron 62 temas en esta lista de popularidad estadounidense (2019).

El *spanglish* es una forma de *code-switching* y se puede tomar a título ejemplificativo para demostrar que no necesariamente es indispensable hablar inglés o buscar activamente estar en un entorno de habla inglesa para practicar este idioma. El lenguaje interesado por el *code-switching*, difiere del lenguaje estándar utilizado en contextos más formales, como por ejemplo en libros (Baheti, Sitaram, Choudhury, & Bali, 2017) y se encuentra mayormente en las plataformas de redes sociales (Bali, Sharma, Choudhury, & Vyas, 2014). Esto representa un problema, siendo la mayoría de los sistemas de procesamiento del lenguaje natural hoy en día monolingües y tienen un desempeño pobre con este tipo de datos.

Según Martínez Soto (2020) para desarrollar modelos aptos para este tipo de lenguaje, hay que enfrentarse a cuatro desafíos principales:

- La falta de *corpus* suficientemente extensas que contengan lenguajes de tipo *code-switched*, necesarios para construir los modelos y que puedan ser ajustados en futuro para diferentes tipos de tareas.
- Las anotaciones lingüísticas necesarias para los algoritmos de aprendizaje supervisado usadas para entrenar los modelos de aprendizaje automáticos, son costosas debido a que deben ser tomadas por profesionales bilingües.
- Encontrar la manera más eficaz de integrar los *corpora* monolingües a tareas para el *code-switching* para poder aprovechar al máximo su potencial.

[12] Lista de éxitos musicales de los sencillos más vendidos en Estados Unidos. Para más información: https://www.billboard.com/

- La incorporación del conocimiento preexistente derivado de otros campos, como por ejemplo de la lingüística, para mejorar el desempeño de los modelos.[10]

Debido a que las redes sociales han hecho posible obtener datos generados por el usuario que reflejan el uso de instancias de *code-switching*, en este trabajo se va a tratar de analizar la relevancia de este tipo de lenguaje mixto en la red social Reddit y en particular en los *subreddit* pertenecientes a la región hispanohablante de Latinoamérica. El objetivo es identificar, si existiera, la frecuencia y recurrencia de los anglicismos y concluir si existe la necesidad de utilizar modelos de procesamiento de lenguaje natural de *code-switching* para futuros análisis.

Capítulo 2: Procesamiento y Análisis de Datos

La generación de información a nivel mundial presenta un crecimiento exponencial, superando nuestra capacidad de consumo. Considerables volúmenes de datos circulan constantemente a través de la red, compuestos por las actividades cotidianas de individuos inmersos en un entorno digital. Esta enorme cantidad de información genera base de datos cada vez más complejas y voluminosas que volvieron obsoletos a los tradicionales softwares de procesamiento de datos. Estos datos se conocen con el nombre de *big data,* un término derivado de *business intelligence*[13], entendido como la capacidad de comprender las interrelaciones entre los datos presentados, de manera que pudieran orientar la toma de decisiones hacia un objetivo deseado (Luhn, 1958).

A pesar de que el concepto de *big data* sea relativamente nuevo, el uso de grandes conjuntos de datos se remonta a los años 1960-1970 y experimentó un crecimiento significativo en los años ochenta, con la utilización de computadoras personales, bases de datos relacionales y del lenguaje SQL (Structured Query Language). El aumento constante de la cantidad de datos en los años noventa, se debió a la popularización del comercio electrónico y motores de búsqueda, que continuaron generando una amplia cantidad de información (Oracle Cloud Infrastructure, 2023). En esta época se introdujeron los almacenes de datos (*data warehouses*), que funcionaban como repositorios centrales de información y que permitían un análisis más profundo para tomar decisiones más informadas.

Pocos años después en 2005, se acuñó el término Web 2.0 para referirse al rápido crecimiento de contenido generado por el usuario en redes sociales como Facebook o YouTube (Firican, 2020) y Roger Mougalas utiliza por primera vez el término *big data* (Halevi & Moed, 2012). Los datos generados en las redes sociales son de tipo no estructurado que, contrariamente a los de tipo estructurado, no tienen modelos predefinidos y pueden ser de cualquier índole: imágenes, audio, datos de texto y mucho más.

Guadalupe Moreno (2019), periodista, afirma que:

> La cantidad de datos creados en todo el mundo en 2018 alcanzó los 33 *zettabytes* (un *zettabyte* equivale a 1.000 millones de *terabytes*), 16,5 veces más que solo hace nueve años. No obstante, gracias a los nuevos desarrollos tecnológicos, como el internet de las cosas, se estima que la cantidad de información digital generada en 2035 ascienda a los 2.142 *zettabytes*. (A la espera de un Big Bang de datos, párrafo 1)

[13] Anglicismo, inteligencia de negocios.

Según Dialani (2020), aproximadamente el 80% de esta información se considera de tipo no estructurado. La predominancia de los datos no estructurados hace que las técnicas de minería de textos sean extremadamente valiosas para las organizaciones, ya que permiten analizarlos y obtener información relevante a partir de ellos.

La minería de texto, también conocida como *text mining*, es una disciplina que se enfoca en el análisis matemático para descubrir patrones y tendencias presentes en los datos recopilados. Su objetivo es establecer un modelo de minería de datos (*data mining*) que permita obtener nueva información a través del uso de métodos estadísticos (Universidad de Málaga, s.a). Utilizando técnicas de minería de texto, se logra la capacidad de examinar diversos tipos de documentos con el fin de desentrañar información que resultaría difícil de identificar de otro modo.

Un caso ilustrativo es el estudio de las proteínas, un campo de investigación crucial para el avance de medicamentos que puedan alterar las conexiones entre las proteínas responsables de ciertas enfermedades (JISC, 2008). El doctor Weeber y su equipo emplearon técnicas de minería de datos para investigar el uso de la talidomida[14], y encontraron que, a pesar de estar comercializado, podría ser beneficioso para los pacientes con lepra. Además, afirmó que estas técnicas fueron fundamentales para su investigación, debido a que cuando se busca "talidomida", se obtienen entre 2.000 y 3.000 resultados. Por otro lado, si se busca "enfermedad", se obtienen 40.000 resultados y gracias al uso de herramientas automatizadas de minería de texto, solo fue necesario leer de 100 a 200 resúmenes y 20 o 30 artículos completos (Weber, 2003).

Otras aplicaciones son: el análisis de sentimientos o minería de opiniones, utilizada para detectar nuevas tendencias o determinar la opinión del publico acerca de un servicio/producto; la personalización de la experiencia cliente para adaptar experiencias personalizadas para diferentes segmentos de clientes dependiendo se sus hábitos (AWS, s.f) o la elaboración de resúmenes.

Las principales tareas de la minería de texto se pueden combinar juntas en un único *workflow* compuesto por cuatro pasos: recolección de la información, procesamiento del

[14] Para más información: https://es.wikipedia.org/wiki/Talidomida

lenguaje natural (NLP), extracción de la información y *data mining* (minería de datos) (Zanini & Dhawan, 2015).

En los subapartados a continuación se tratarán las primeras tres fases del proceso de *text mining*: recolección de información, procesamiento del lenguaje natural y extracción de la información.

2.1. Obtención y Procesamiento de Publicaciones

Como se menciona en la introducción de este capítulo, el primer paso esencial en el proceso de minería de textos consiste en recolectar los datos que serán utilizados. Para llevar a cabo esta tarea, existen dos técnicas conocidas: el *web scraping* y el API *scraping*. En un artículo elaborado por Paulina Tobella en 2021, se analizan las diferencias entre ambas herramientas. Según la autora, el *web scraping* se refiere a la práctica de extraer datos de una o varias páginas web utilizando herramientas automatizadas, y destaca que es un método sencillo que no requiere la intervención de un programador. Además, Tobella señala que este enfoque es rápido en cuanto a la actualización de datos, ya que se ajusta automáticamente a cualquier modificación que ocurra en línea. Por otro lado, se destaca que:

> [..] la interfaz de programación de aplicaciones (API) actúa como un puente que conecta la consulta con la solución. Las reglas de transferencias son fijas y solo se pueden alterar cuando un programador cambia el software API. [..] las regulaciones estrictas solo permitirán que la empresa recopile datos específicos y solo puedan acceder a algunos campos de datos particulares (Las Diferencias Entre Web Scraping y API, párrafo 1).

En este trabajo, se optó por obtener la información mediante la API de la plataforma en cuestión, lo que implicó recolectar los datos de una única fuente. Esta elección se hizo con el propósito de facilitar el proceso de recopilación y obtener resultados bien estructurados.

Reddit es una plataforma de discusión de contenido web y agregación de noticias que fue fundada en 2005. Los usuarios conocidos como "Redditors", tienen la capacidad de compartir contenido como imágenes, enlaces y publicaciones de texto. Los demás usuarios pueden votar a favor o en contra, comentar y compartir dicho contenido. La plataforma está

organizada en "subreddits", que son foros creados por los usuarios y abarcan una amplia variedad de temas.

En la actualidad, un reporte de redactado por Statista (2022) afirma que la plataforma se ha convertido en una de las redes sociales más populares a nivel global, ocupando un lugar destacado en términos de número de usuarios, solo superada por Pinterest y Twitter. Hasta principios de 2020, se estima que había aproximadamente 430 millones de usuarios activos mensuales en todo el mundo. Es importante destacar que la plataforma ha experimentado un crecimiento significativo en el número de usuarios activos mensuales desde 2019 hasta 2021, con un aumento del 30%, superando a actores establecidos en el mercado como Facebook, Snapchat, Instagram y Twitter.

Para el desarrollo de este trabajo se consideran los títulos de las publicaciones, los comentarios de los *subreddits* con más de 500 miembros relacionados con los países hispanohablantes de Latinoamérica (ver el Apéndice 1 para la lista completa) y su fecha de creación para poder hacer la distinción entre años. Con la librería Pandas[15] y el uso de Python, se extrae la información descrita y se generan *dataframe*[16] por cada país a analizar; con el objetivo de obtener un dataframe de dos columnas: una que contiene el año de creación y la otra que contiene el texto extraído.

Antes de poder comenzar con el análisis textual, es necesario realizar una limpieza de los textos, para filtrar datos inútiles. Para eso, se utilizaron spaCy[17] y NLTK (*Natural Language Toolkit*)[18], paquetes especificadamente pensados para el desarrollo de aplicaciones relacionadas con el procesamiento de lenguaje natural. Gracias a estas librerías es posible acceder fácilmente a una extensa colección de corpus, generados por múltiples fuentes, como, por ejemplo: libros, noticias, artículos y redes sociales (Moez, 2023); recursos léxicos y librerías dedicadas al procesamiento textual para clasificación, *tokenization*[19], *stemming*[20], *tagging*[21] y *parsing*[22] (NLTK, s. f.).

---

[15] Para más información: pandas - Python Data Analysis Library (pydata.org)
[16] Paneles bidimensionales compuestos por filas y columnas, que permiten destacar las relaciones entre las distintas variables de la serie de datos. (DataScientest, s.a.)
[17] Para más información: spaCy · Industrial-strength Natural Language Processing in Python
[18] Para más información: NLTK :: Natural Language Toolkit
[19] Anglicismo: el acto de dividir un texto en elementos más cortos llamados *tokens*. (Aravindpai, 2020)
[20] Anglicismo: método para reducir una palabra a su raíz. (Wikipedia, 2021)
[21] Anglicismo: es una técnica que consiste en etiquetar cada palabra de un documento en su correspondiente categoría gramatical. (KeepCoding, 2023)
[22] Anglicismo: proceso de analizar una secuencia de símbolos a fin de determinar su estructura gramatical definida. (Alarcón, 2017)

La primera tarea que se decide realizar es la normalización de los textos, que incluye diferentes tareas. Importante es la exclusión de *stopwords*, palabras que son comúnmente utilizadas pero que no comunican información importante, como por ejemplo “a”, “lo”, “por”, “del” y que pueden distorsionar el análisis si no se remueven (Moez, 2023). La transformación de todas las letras en minúsculas (*case-folding*) para que, por ejemplo, la palabra Auto y auto coincidan. Esto podría causar problemas en casos de nombres de organizaciones y marcas (*Día*, cadena argentina de supermercados) que se diferencian de un nombre común (día) por la capitalización de la primera letra.

A pesar de lo expuesto, se consideró que la transformación en minúsculas y la eliminación de los diacríticos de todas las palabras sea la solución más practica en el caso del presente trabajo, debido a que especialmente en las redes sociales, los usuarios tienden a no capitalizar las palabras y a no utilizar los acentos (Manning, Raghavan, & Schütze, 2008).

Se crearon tokens de estos textos y a continuación, se cargaron los *corpus* en idioma inglés y español para poder identificar los elementos lingüísticos en español. Posteriormente, se retuvieron únicamente aquellos términos que también se encontraron en el corpus en inglés, es decir, los anglicismos. Una vez más, se excluyeron las palabras comunes (*stopwords*) y aquellas que pertenecen exclusivamente al corpus en español; manualmente se creó un array de palabras para excluir aquellas que podrían generar ambigüedad, por ser lemas aceptados en el idioma inglés.

2.2. Análisis de Publicaciones

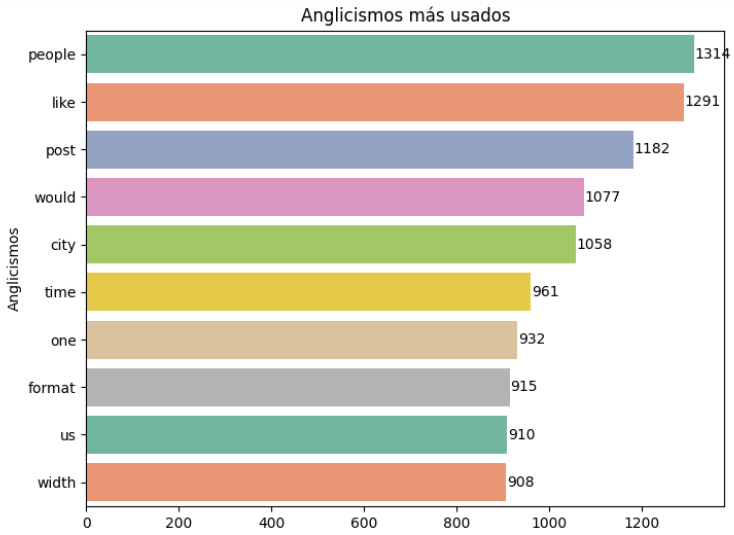
El procesamiento de datos mencionado en la sección anterior contribuye a optimizar el análisis necesario para realizar la minería de textos de manera más eficiente. Sin embargo, con el propósito de adquirir conocimiento del comportamiento histórico, es fundamental llevar a cabo un análisis descriptivo que facilite la comprensión de los eventos pasados y permita realizar un diagnóstico preliminar. Para lograr esto, es posible emplear la computación de métricas estadísticas y la utilización de representaciones gráficas que brinden una visualización del patrón de los datos. En este trabajo, para visualizar los análisis realizados para la detección de los anglicismos más frecuentes, se utilizan las herramientas matplotlib[23] y Seaborn[24].

[23] Para más información: Matplotlib — Visualization with Python
[24] Para más información: seaborn: statistical data visualization — seaborn 0.12.2 documentation (pydata.org)

Con el objetivo de visualizar solo los tokens más relevantes, se decide incluir solo los 10 anglicismos más frecuentes (Figura 1).

Figura 1

Top 10 anglicismos más frecuentes

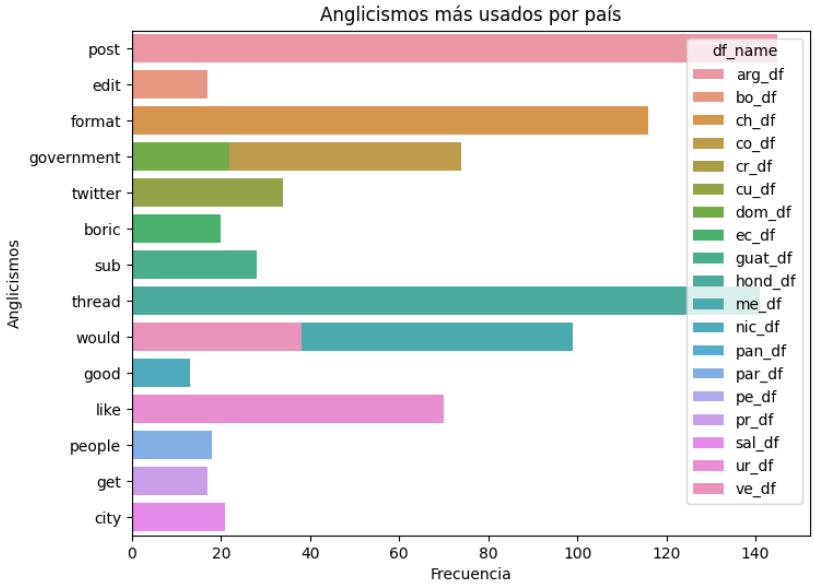


Fuente: Elaboración propia en Python

De la figura anterior se observa que las palabras con mayor frecuencia son: “people” (gente, con 1314 apariciones) y “like” (me gusta, con 1291 apariciones) términos que se pueden vincular a las redes sociales y a la interacción con otros usuarios. Como complemento de este primer diagnóstico, se decide analizar también las palabras más comunes por *subreddit* (Figura 2, para más información ver Apéndice 2. Frecuencia anglicismos más frecuentes por *subreddit*).

Figura 2

Anglicismos más frecuentes por subreddit



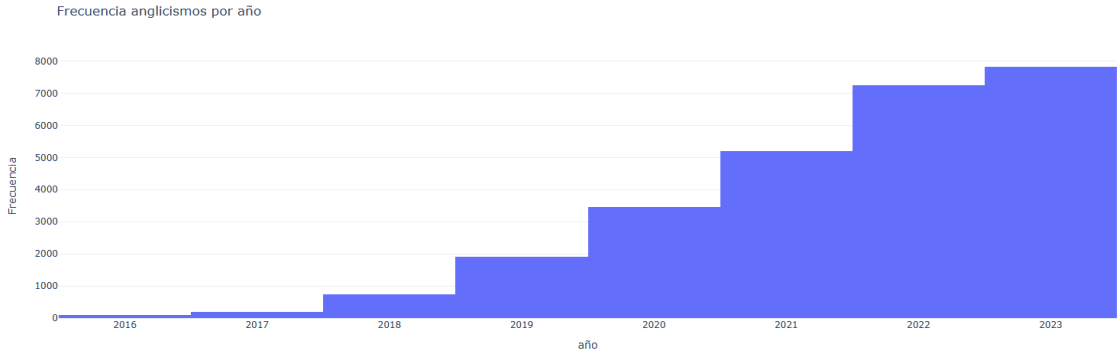
Fuente: Elaboración propia en Python

Se destaca que las palabras más frecuentes encontradas en un *subreddit* son “post” (publicación/publicar, con 145 apariciones) en el *subreddit* argentino y “thread” (tema, 141 apariciones) en el *subreddit* hondureño. En general, aparecen términos relacionados con el mundo digital (*thread*, *twitter*, *post*, *edit*, *sub*, *like*, *format*) o con la sociedad (*government*, *people*, *city*).

En la Figura 3, se puede notar un fuerte incremento del número de anglicismos en el periodo 2016-2023.

Figura 3

Frecuencia de los anglicismos por año

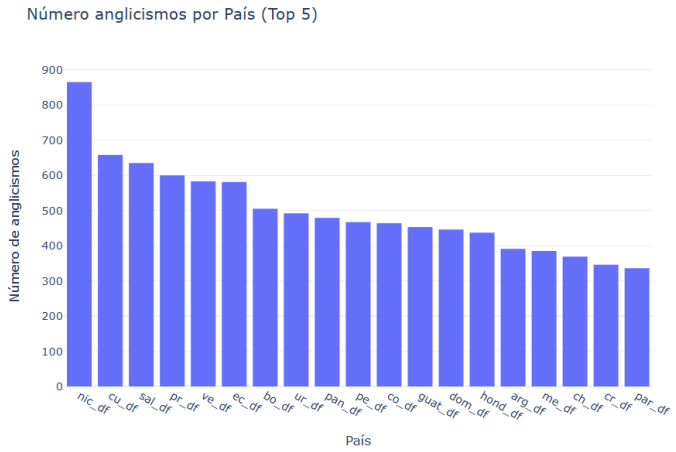


Fuente: Elaboración propia en Python

Para definir en cual *subreddit* se utiliza el mayor número de anglicismos, se toma una muestra de 1000 casos de cada uno. Se decidió este número, ya que en algunos *subreddits* se pudo obtener hasta 12000 tokens mientras que en otros casos hay tan solo 1000 tokens (Figura 4).

Figura 4

Anglicismos por subreddit



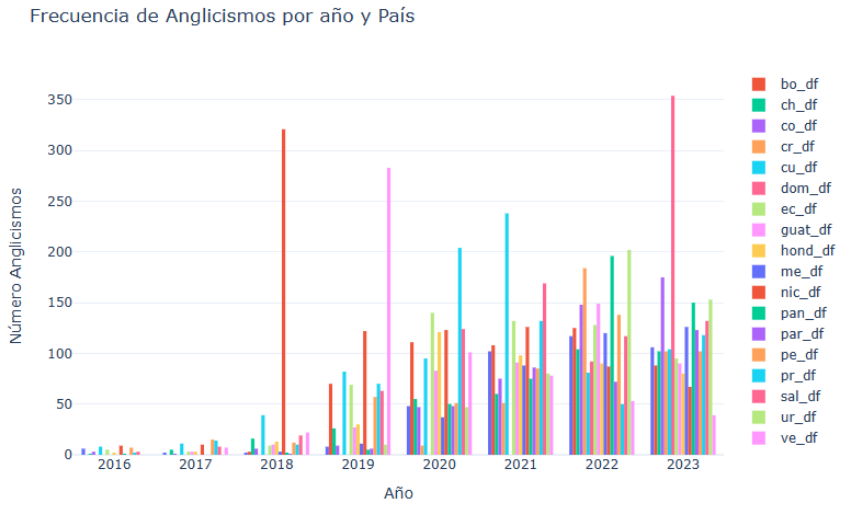
Fuente: Elaboración propia en Python

De este análisis resulta evidente que el subreddit nicaragüenses es el que tiene la mayor cantidad de anglicismos (865 sobre 1.000) seguido por el de Cuba (658 sobre 1.000) y El Salvador (635 sobre 1.000), siendo el subreddit de Paraguay el que tiene la menor cantidad (336 sobre 1.000). Resalta que hay una diferencia substancial entre el primer y segundo lugar, mientras que entre los demás lugares la diferencia es mucho menor.

A lo largo de los años, Nicaragua resulta ser el subreddit con el mayor número de anglicismos en el año 2018, destacándose ampliamente frente a los demás. En los años 2020-2021 sobresale el subreddit cubano; en 2020 el venezolano y en 2023 el de República Dominicana (Figura 5).

Figura 5.

Frecuencia de anglicismos por año



Fuente: Elaboración propia en Python

2.3. Algoritmos de Minería de Datos y Detección de Tópicos

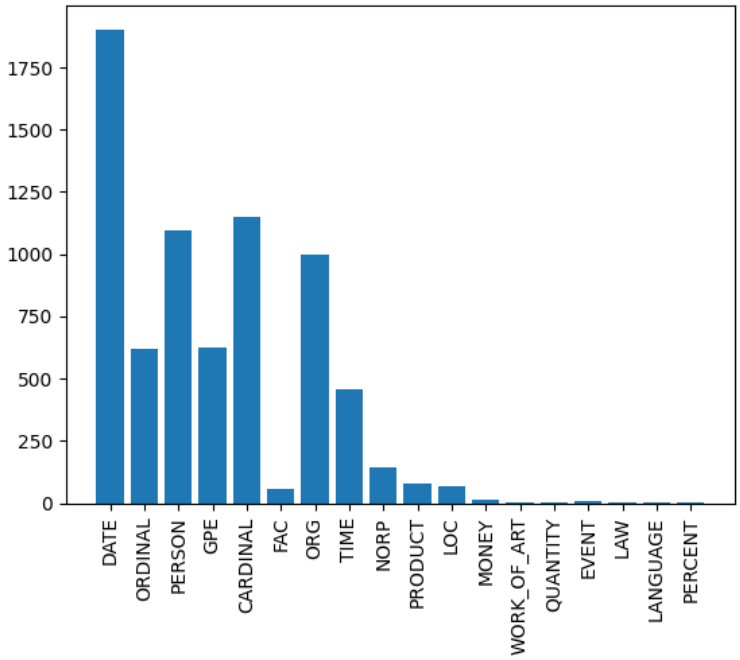
Las palabras se pueden agrupar por clases morfológicas (*part of speech*, PoS). De acuerdo con la gramática tradicional, no son muchas: sustantivos, verbos, adjetivos, preposiciones, adverbios, entre otras; modelos más recientes cuentan con más clases (Penn Treebank, 45; Brown Corpus, 87; C7 tagset, 146) (Maggini, s.f.). Según la gramática, palabras como Janet, Colorado o Universidad de Buenos Aires son nombres propios; de una perspectiva semántica estos sustantivos se refieren a diferentes entidades: Janet es una persona, Universidad de Buenos Aires es una organización y Colorado es una ubicación (Jurafsky & Martin, 2023)

El reconocimiento de las entidades sustantívales (*Named entity recognition*, NER) es un proceso en el cual una oración o parte de texto se analiza para encontrar entidades y asignar categorías como por ejemplo nombres, organizaciones, ubicaciones, cantidades, valores monetarios, porcentajes, etc. (DeepAI, s.f.).

Utilizando la librería spaCy, se analizaron los anglicismos recolectados (Figura 6), donde resaltó que la tipología más utilizada es de tipo fecha (*date*) con 1904 casos, seguida por la categoría de números cardinales (1150) y personas (*person*, 1097).

Figura 6

Frecuencia de la tipología de los anglicismos



Fuente: Elaboración propia en Python

Para analizar los temas más recurrentes de los anglicismos utilizados en Reddit, se recurre a la detección de tópicos (*topic modelling*). El *topic modelling* es una aplicación de la minería de textos, que permite clasificar documentos en función de su temática, descubriendo el tema subyacente en una colección de documentos (KeepCoding, 2023).

El algoritmo que se decidió utilizar para detectar los temas es el *Latent Dirichlet Allocation* o LDA, que busca visualizar los temas subyacentes mediante probabilidades de palabras. Introducido por Blei, Ng y Jordan en 2003, se trata de un método probabilístico generativo no supervisado para modelar un corpus, donde se asume que cada documento puede ser representado como una distribución probabilista sobre temas latentes y se asume que las distribuciones de temas en todos los documentos comparten una distribución de Dirichlet común (Jelodar, y otros, 2019).

El LDA, al igual que todos los modelos de temas probabilísticos, es un modelo generativo de un corpus de documentos que, a partir de tres distribuciones de probabilidad, extrae las palabras que conformarán cada documento. Estas tres distribuciones son las siguientes: una distribución de Poisson con el parámetro lambda (λ), que representa la

cantidad de palabras en los documentos (N_d); una distribución multinomial (θ) que describe las distribuciones de probabilidad de los K temas en los D documentos. Finalmente, la distribución tópicos-palabras (φ), que describe la probabilidad de que las palabras pertenezcan a un tópico (Cao, 2019).

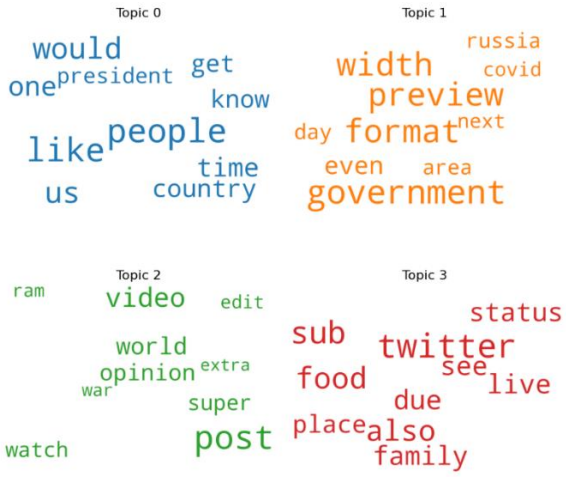
Para construir un documento es necesario elegir una cantidad palabras (N_d) que siguen la distribución de Poisson (λ) en el documento; para cada documento en el corpus y por cada palabra (w_n) se escoge un tema de acuerdo a la distribución multinomial (θ). Finalmente, a cada palabra extraída del vocabulario (V), se le asigna una probabilidad multinomial de pertenencia en cada uno de los tópicos (Proto, 2018).

Después de evaluar la distribución de los anglicismos en varios grupos, se decidió que la mejor división sería en cuatro grupos temáticos que se visualizan a través de una nube de palabras; una herramienta visual que ayuda a resaltar las palabras más frecuentes (Figura 7).

- Tema 0 relacionado con Política y Liderazgo, que contiene principalmente las palabras: *would* (condicional del verbo ser), *president* (presidente), *get* (agarrar), *know* (saber), *people* (gente), *like* (gustar), *time* (tiempo), *us* (nosotros o EE. UU.), *country* (país);
- Tema 1 relacionado con Eventos Actuales y Asuntos Internacionales, que contiene principalmente las palabras: *Russia* (Rusia), Covid, *width* (ancho), *preview* (previsualización), *day* (día), *format* (formato), *next* (siguiente), *even* (incluso), *area* (área), *government* (gobierno);
- Tema 2 relacionado con Tecnología y Medios de Comunicación, que contiene principalmente las palabras: *ram* (memoria RAM), video (vídeo), *edit* (editar), *world* (mundo), extra, *opinion* (opinión), *war* (guerra), super, *watch* (ver), post (publicación);
- Tema 3 relacionado con Redes Sociales y Vida Personal, que contiene principalmente las palabras: *sub* (subreddit), *status* (estado), twitter, *see* (ver), *live* (en vivo), *food* (comida), *due* (para el tal día), place (lugar), *also* (también), *family* (familia).

Figura 7

Grupos temáticos de los anglicismos



Fuente: Elaboración propia en Python

Capítulo 3: Relación entre Anglicismos y EE. UU.

La influencia de los Estados Unidos ha tenido un enfoque primordialmente económico, surgido de su marcada supremacía en los ámbitos industriales y tecnológicos, y fuertemente relacionado con la expansión de negocios estadounidenses y la inversión en el extranjero. Después de la segunda guerra mundial, los Estados Unidos emergieron como la principal y más próspera potencia industrial a nivel mundial. Lillibridge (1966) afirma que, en el período posterior a la guerra, el país poseía recursos de carácter liberal esenciales para la reconstrucción de economías devastadas, así como para sentar las bases de nuevas economías en sociedades anteriormente atrasadas. Como consecuencia, se generó un flujo considerable de recursos económicos desde los Estados Unidos, tanto de origen público como privado, una situación sin precedentes en la historia.

Además, comenta que, desde el final de la guerra hasta mediados de 1962, se destinaron al extranjero más de sesenta y seis mil quinientos millones de dólares en forma de asistencia económica, y la afluencia de capital privado estadounidense también fue sustancial. Solo durante la década de 1950, se invirtieron alrededor de treinta mil millones de dólares en el extranjero. Para 1960, la inversión de capital en el extranjero ascendió a unos cuarenta mil millones de dólares. Los Estados Unidos no solo brindaban fondos públicos y privados, también ofrecían importantes recursos filantrópicos (por ej. el plan Marshall).

Se puede entonces afirmar que este país tiene un papel fundamental en la escena internacional y por eso se decidió analizar la influencia de esta Estados Unidos en Latinoamérica donde, por temas geográficos e históricos, se comparten similitudes en términos de idioma, cultura, historia, clima y lugares de interés turístico.

Sin embargo, estas naciones presentan diferencias marcadas en sus economías, que se han desarrollado de manera divergente a lo largo del último siglo (Eugenio-Martin, Morales Martin, & Scarpa, 2004).

En este contexto, en este apartado se explicarán las técnicas de *clústering* y se utilizará un modelo de aprendizaje no supervisado, específicamente el algoritmo K-means, con el propósito de agrupar naciones con características económicas y turísticas similares. Los resultados obtenidos a través de este análisis permitirán evaluar si la influencia comercial y turística de los Estados Unidos influyó en el uso de los anglicismos en los países de interés.

3.1. Modelos de *Clústering*

El *clústering* es una técnica ampliamente utilizada y de relativa simplicidad en el análisis multivariado. Esta metodología opera en un enfoque de aprendizaje no supervisado, lo que implica la ausencia de conocimiento previo, y su propósito es adaptarse a las observaciones al agrupar elementos similares. En este contexto, un clúster denota una agrupación de objetos con rasgos comunes o correlacionados entre sí y que difieren de los objetos en otros grupos. En diversas situaciones, un clúster puede ser visto como una suerte de conjunto que reúne elementos con similitudes, mientras se distinguen de otros conjuntos. (Universidad Mediterránea de Reggio Calabria, s.f.).

Las funciones esenciales de este enfoque de aprendizaje no supervisado se centran en la comprensión y resumen de datos. Sus aplicaciones son variadas: puede ser empleado como un análisis independiente, como paso inicial previo a otras técnicas de análisis, como componente integrado en algoritmos para distintos tipos de análisis, o para la preparación de datos, lo que involucra la identificación y eliminación de valores atípicos y la reducción de la dimensionalidad de los datos (Amato, 2009). En un ámbito empresarial, como el económico, el *clústering* resulta útil para que los operadores identifiquen grupos diferenciados de clientes según sus patrones de compra. En el campo de la biología, se utiliza para establecer clasificaciones taxonómicas de plantas y animales, categorizar genes con funcionalidades similares y explorar diversas características de grupos específicos. En un motor de búsqueda, se pueden agrupar las respuestas parecidas entre sí, para poder presentar menos alternativas a los usuarios.

Según el profesor Amato de la Universidad de Chieti-Pescara (2009), los algoritmos de *clústering* se aplican típicamente en las siguientes estructuras de datos:

- Una matriz de datos (Figura 8): esta matriz representa n objetos, como por ejemplo personas, con *p* variables (también llamadas medidas o atributos), como edad, altura, peso, etnia, y así sucesivamente. La estructura adopta la forma de una tabla relacional o una matriz $n \times p$ (*n* objetos por *p* variables).

Figura 8

Matriz de datos x_{ij}

$$\begin{pmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1p} & \dots & \dots & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{pmatrix}$$

Fuente: Adaptado de "Analisi di Raggruppamento" (p.7), por G. Amato, 2009, Universidad "G. D'Annunzio" de Chieti-Pescara

- Una matriz de disimilitud: esta matriz almacena el grado de disimilitud entre cada par de objetos involucrados. A menudo se representa en forma de una tabla $n \times n$, donde $d(i, j)$ representa la diferencia o disimilitud medida entre los objetos i y j. Es importante notar que $d(i, j) = d(j, i)$ y que $d(i, i) = 0$.

Figura 9

Matriz de las distancias

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{pmatrix}$$

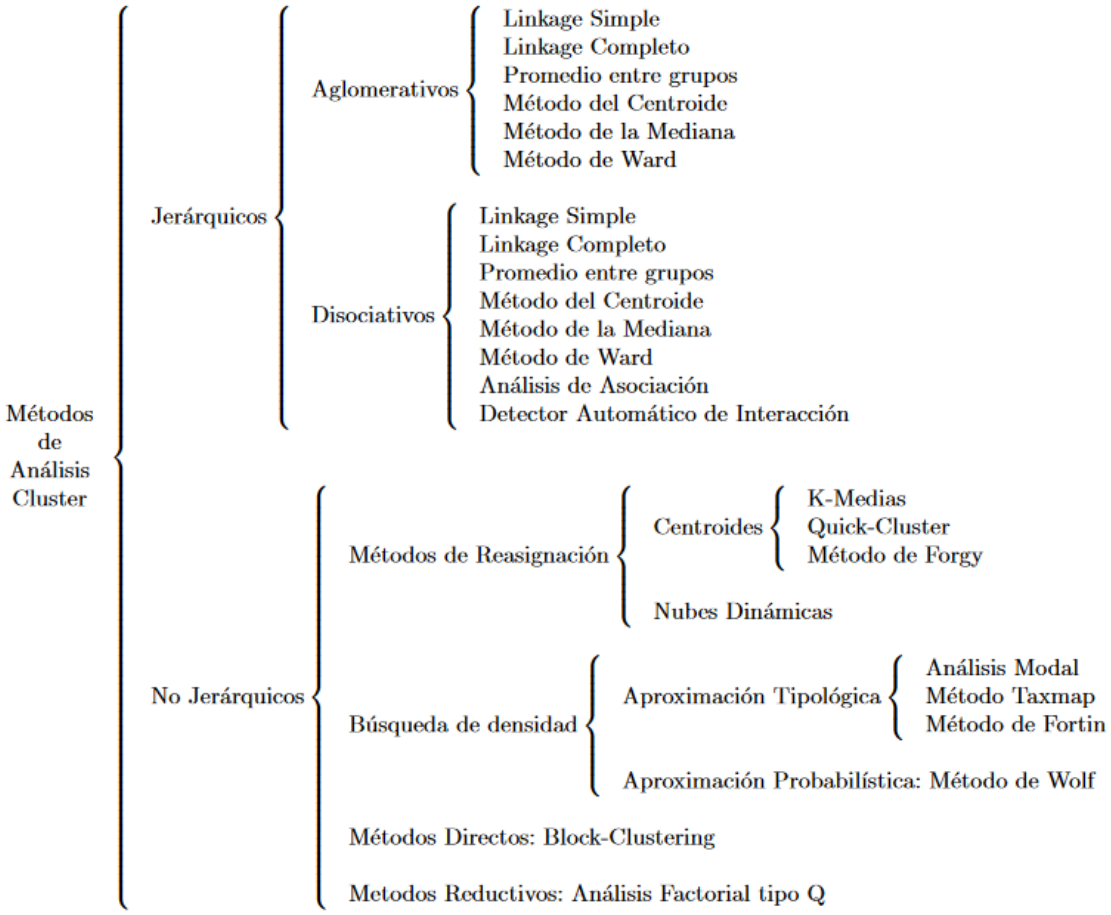
Fuente: Adaptado de "Analisi di Raggruppamento" (p.7), por G. Amato, 2009, Universidad "G. D'Annunzio" de Chieti-Pescara

Frecuentemente, la matriz de datos es conocida como matriz bidireccional, mientras que la matriz de disimilitud es denominada matriz unidireccional. Esta distinción se debe a que las filas y columnas de la primera representan distintas entidades, a diferencia de la segunda, que representa una misma entidad. Muchos algoritmos de *clústering* operan empleando la matriz de disimilitud. En caso de que los datos estén presentados en forma de una matriz de datos, se requiere realizar una conversión para obtener una matriz de disimilitud antes de aplicar dichos algoritmos.

A grandes rasgos se distinguen dos grandes categorías de métodos *clúster*: métodos jerárquicos y métodos no jerárquicos.

Figura 10

Clasificación de las técnicas clúster



Fuente: Adaptado de "Introducción al Análisis Clúster. Consideraciones generales" (p.63), por la Universidad de Granada, s.f., Universidad de Granada

Según lo descrito por la Universidad de Granada sobre las técnicas de *clústering* (s.f.), el método jerárquico crea una descomposición jerárquica de un conjunto dado de objetos que se pueden clasificar en aglomerativos o disociativo, según cómo se realice esta descomposición. En el enfoque aglomerativo cada objeto forma inicialmente un grupo separado; luego, los objetos o grupos cercanos entre sí se fusionan hasta obtener un único grupo (el nivel más alto de la jerarquía) o hasta que se cumpla una condición de finalización. En el enfoque disociativo, se comienza con todos los objetos en un mismo clúster. En cada iteración subsiguiente, un clúster se divide en clústeres más pequeños hasta que cada objeto se encuentre en un clúster diferente o hasta que se cumpla una condición de finalización específica. Los docentes de la Universidad de Granada siguen relatando que los métodos

jerárquicos tienen la limitación de que una vez que se ha realizado un paso (fusión o división), no se puede deshacer. Esta rigidez es beneficiosa ya que conlleva menores costos computacionales al no permitir un número combinatorio de opciones diferentes. Sin embargo, un gran problema de estas técnicas es que no pueden corregir decisiones incorrectas. En los métodos no jerárquicos también conocidos como partitivos o de optimización, es necesario fijar de antemano el número de clúster deseado. Otras diferencias con el método jerárquico es que la agrupación se realiza para optimizar el criterio de selección y que se trabaja con la matriz de datos original, sin necesitar su conversión a una matriz de distancias (Universidad de Granada, s.f.).

Con ambas técnicas es importante establecer como determinar la similitud entre dos objetos para que se puedan formar los *clústeres*, para calcular estas distancias las funciones más utilizadas son la Manhattan y la euclídea que arrojan valores más altos para objetos que son más diferentes el uno al otro (Howard, 2009).

3.2. Implementación de un Modelo de Clústering

En el marco de la creciente interconexión global, resulta importante examinar en detalle la influencia que Estados Unidos ejerce en distintos países, tanto desde una perspectiva comercial como turística. Este estudio tiene como objetivo principal llevar a cabo una evaluación exhaustiva de dicha influencia, utilizando datos provenientes de fuentes oficiales y confiables. Para tal fin, se han empleado dos fuentes clave: la página oficial de la Oficina del Representante Comercial de Estados Unidos (USTR) y los datos suministrados por la Organización Mundial del Turismo (UNWTO).

Desde la perspectiva comercial, la USTR desempeña un rol central en la coordinación y supervisión de la política comercial internacional de Estados Unidos. Su plataforma en línea alberga información detallada sobre las relaciones comerciales con distintos países, abordando aspectos como importación, exportación, balance comercial e inversiones. Para este estudio, se ha adoptado un enfoque que se basa en la clasificación ordinal disponible en las fichas de países de la USTR. Esta clasificación, que ordena a los países según el volumen total de productos intercambiados con Estados Unidos, proporciona una visión estructurada de la relación comercial entre Estados Unidos y cada país en cuestión.

En cuanto a la influencia turística, el análisis se apoya en los datos recopilados por la UNWTO, una entidad reconocida a nivel mundial por sus estadísticas turísticas exhaustivas. En específico, se ha establecido un umbral de un millón de turistas estadounidenses como criterio para identificar los destinos turísticos de mayor relevancia. Aquellos países que superan este umbral se identifican con el número 1, mientras que los demás reciben la categoría número 2. Mediante este enfoque, se logra trazar con precisión los destinos que experimentan una afluencia significativa de turistas estadounidenses, lo que a su vez denota una fuerte influencia turística proveniente de Estados Unidos (para consultar la tabla completa, consultar el apéndice 3).

Para dividir los datos en *clústeres*, se decidió utilizar el método de reasignación K-Means que busca lograr una alta similitud dentro de los *clústeres* y baja similitud entre ellos. Esta metodología posibilita que un individuo *k* inicialmente asignado a un grupo de objetos *n* durante una etapa específica del proceso, pueda ser posteriormente designada a otro grupo en una fase posterior, siempre y cuando esta reasignación mejore el criterio de selección. El procedimiento finaliza en el momento en que ya no queden individuos cuya reasignación pueda perfeccionar el logro obtenido hasta ese punto (Universidad de Granada, s.f.). El criterio utilizado generalmente es el error cuadrático, definido como:

$$Err = \sum_{i=1}^{k} \sum_{p \in C_i} d_e(p, m_i)^2$$

Donde *E* es la suma del error cuadrático de todos los objetos en la base de datos, *p* es el punto que representa el objeto en el espacio, y m_i es el promedio del clúster C_i (*p* y m_i son multidimensionales).

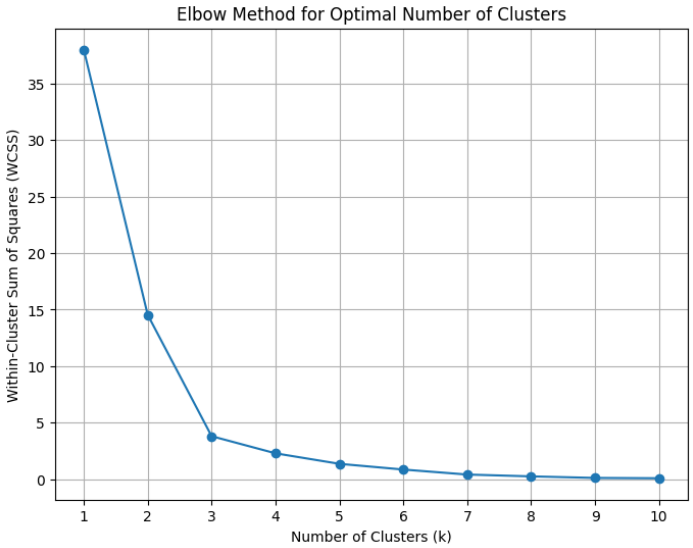
Con el objetivo de graficar el numero óptimo de *clústeres*, se opta por emplear el método del codo, el cual se apoya en la métrica conocida como "Suma de Cuadrados Dentro de los Clústeres" (WCSS). El WCSS es una medida que cuantifica la dispersión interna de los puntos en un clúster, midiendo qué tan cerca están los puntos entre sí en relación con su centroide. Su cálculo se basa en la suma de las distancias euclidianas al cuadrado entre cada punto y el centroide del clúster al que pertenece (ODSC - Open Data Science, 2018). Al crear el gráfico de la métrica WCSS con respecto al número de clústeres (*k*) en los ejes *y* y *x*, respectivamente, se observa una disminución en el valor de WCSS a medida que aumenta *k*. Sin embargo, esta disminución tiende a aplanarse en un punto, formando un codo en el

gráfico, el punto en el que la curva del gráfico muestra este codo es el número óptimo de clústeres. Este número implica un equilibrio entre la reducción de la dispersión interna en los clústeres y la minimización del número de clústeres, evitando así un exceso de fragmentación.

El método del codo sugiere tres como el número óptimo de clústeres según se ilustra en la Figura 11.

Figura 11

Método del codo



Fuente: Elaboración propia en Python

Sin embargo, para asegurar la validez de esta elección, se opta por verificar también mediante el método de la silueta (*silhouette*).

El coeficiente de la silueta se puede definir como la distancia promedio de un punto (*x*) a todos los otros puntos en el mismo clúster *a(x)* y la distancia promedio de un punto (*x*) a todos los otros puntos en el clúster más cercano al que (*x*) no pertenece *b(x).* Con base en estas definiciones, el coeficiente se calcula utilizando la siguiente fórmula:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

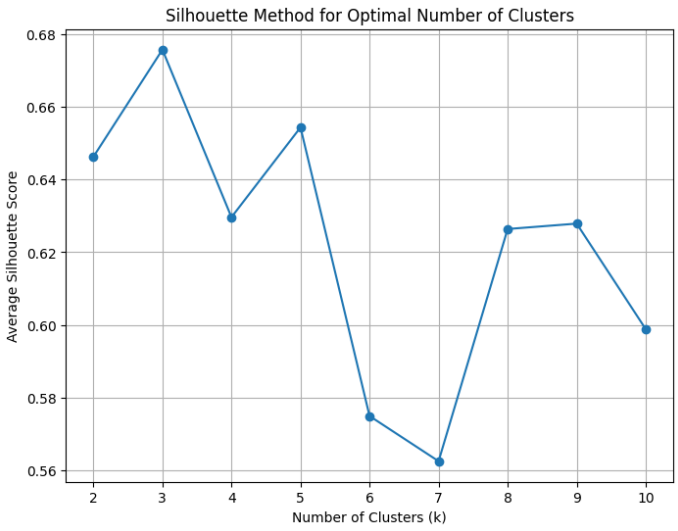
El valor de *s(x)* se encuentra en el rango entre -1 y 1 el coeficiente para todo el agrupamiento es:

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x)$$

Según el método de la silueta, tres clústeres resulta nuevamente ser el número óptimo a escoger (Figura 12).

Figura 12

Método de la silueta



Fuente: Elaboración propia en Python

3.3. Análisis de la Influencia de EE. UU. en las Publicaciones

Siguiendo las recomendaciones brindadas por el método del codo y el coeficiente de la silueta, se decide utilizar el método K-Medias con tres *clústeres* (Figura 13), cuyos resultados son:

- Clúster 0: Costa Rica; República Dominicana; México; Puerto Rico.

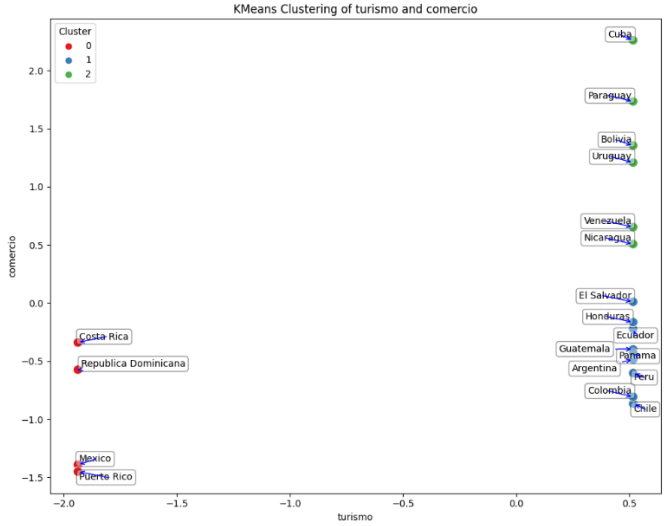
  Este grupo engloba a países que reciben más de 1 millón de turistas de Estados Unidos por año (valor 1 en la columna "Turismo"). En lo que respecta al comercio, no todos los países en este grupo son considerados socios comerciales relevantes, ya que algunos tienen un valor igual o menor a 30 en la columna "Comercio".
- Clúster 1: El Salvador; Honduras; Ecuador; Guatemala; Panamá; Argentina; Perú; Colombia; Chile.

  Este clúster agrupa a países que reciben pocos turistas de Estados Unidos, pero que son considerados socios comerciales relevantes debido a valores significativos en la columna "Comercio".
- Clúster 2: Cuba; Paraguay; Bolivia; Uruguay; Venezuela; Nicaragua.

Aquí se encuentran países que reciben un número bajo de turistas de Estados Unidos. En cuanto al comercio, hay pocos socios comerciales relevantes para Estados Unidos en este grupo.

Figura 13

Clustering de los países (y= comercio; x=turismo)



Fuente: Elaboración propia en Python

Una vez efectuada esta división, se decide entonces analizar si el turismo estadounidense o las relaciones comerciales con este país influyen en la frecuencia de los anglicismos (Figura 14).

Figura 14

Frecuencia de anglicismos por año y clúster

Fuente: Elaboración propia en Python

Se observa en la Figura 14 que, a partir del año 2020, los *subreddits* de los países pertenecientes al clúster 1 presentan un aumento notable en la cantidad de anglicismos en comparación con los clústeres de los demás países. Durante los años 2018 y 2019, los países del clúster 2 encabezaron la lista en términos de la frecuencia de anglicismos.

Conclusión

En base a los resultados analizados, se puede concluir que la cantidad de anglicismos está influenciada por las relaciones comerciales con Estados Unidos. Los resultados resaltan la importancia de considerar tanto los aspectos económicos como culturales al analizar la adopción de anglicismos, y cómo estos factores pueden interactuar de manera significativa en el lenguaje utilizado en línea. Además, estos resultados contribuyen a la comprensión de cómo la influencia económica y turística de Estados Unidos puede relacionarse con el uso de anglicismos en contextos lingüísticos variados.

En cuanto a la transferencia de estos hallazgos, se vislumbran implicaciones en múltiples campos. Desde una perspectiva lingüística, los resultados ofrecen una visión más profunda de cómo las relaciones comerciales y turísticas pueden moldear la evolución del lenguaje en comunidades digitales. En términos económicos, estos descubrimientos pueden ser de relevancia para estrategias comerciales y de marketing, al considerar cómo el uso de anglicismos puede influir en la interacción con países de habla inglesa.

Mirando hacia el futuro, este estudio deja entrever varias líneas de investigación por explorar. Sería valioso profundizar en el análisis de las razones detrás de los aumentos en la frecuencia de anglicismos en momentos específicos, lo que permitiría una comprensión más rica de los factores subyacentes. Además, se podría investigar cómo otros aspectos culturales y sociales interactúan con el uso de anglicismos en contextos digitales. Ampliar la metodología para analizar la influencia de otros idiomas o culturas en el lenguaje en línea podría arrojar nuevas luces sobre la dinámica global del lenguaje. En última instancia, este estudio subraya la necesidad de continuar explorando la relación entre la economía, el turismo y el lenguaje en un mundo cada vez más interconectado.

Bibliografía

Aguilera, R. (23 de septiembre de 2019). *La música en 'spanglish' conquista los oídos estadounidenses*. Recuperado el 18 de 7 de 2023, de El País: https://elpais.com/sociedad/2019/09/21/actualidad/1569102725_803175.html

Alarcón, N. (24 de noviembre de 2017). *¿Qué es Parseo (Parsing)?* Recuperado el 1 de 8 de 2023, de Alarcón Nelson: https://www.alarconnelson.com/2017/11/que-es-parseo-parsing.html

Amato, G. (2009). *Analisi di Raggruppamento.* Chieti. Pescara: Università "G. D'Annunzio" di Chieti-Pescara. Obtenido de https://www.sci.unich.it/~amato/teaching/old/datamining08/lucidi/07-clústering.pdf

Amazon. (22 de junio de 2023). *aws*. Obtenido de Amazon: https://aws.amazon.com/it/what-is/api/

Aravindpai, P. (26 de mayo de 2020). *What is Tokenization in NLP? Here's All You Need To Know*. Recuperado el 1 de 8 de 2023, de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/

AWS. (s.f de s.f de s.f). *¿Qué es el análisis de textos?* Recuperado el 29 de 7 de 2023, de AWS: https://aws.amazon.com/es/what-is/text-analysis/

Baheti, A., Sitaram, S., Choudhury, M., & Bali, K. (2017). Curriculum Design for Code-switching: Experiments with Language Identification and Language Modeling with Deep Neural Networks. *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)* (págs. 65-74). Kolkata: NLP Association of India.

Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). Proceedings of the First Workshop on Computational Approaches to Code Switching. *"I am borrowing ya mixing?" An Analysis of English-Hindi Code Mixing in Facebook* (págs. 116-126). Doha: Association for Computational Linguistics.

Budiman, A., Tamir, C., Mora, L., & Noe-Bustaman, L. (20 de agosto de 2020). *Facts on US immigrants*. Recuperado el 10 de 7 de 2023, de Pewresearch: https://www.pewresearch.org/hispanic/2020/08/20/facts-on-u-s-immigrants/

Cantero, F. J., & De Arriba, C. (1996). El cambio de código: contextos, tipos y funciones. En F. J. Arriba, J. Otal, & I. F. Codina (Edits.), *Estudios de Lingüística Aplicada* (págs. 587-596). Barcelona: Publicacions de la Universitat Jaume I.

Cao, N. (s.f. de s.f. de 2019). Modelli Latent Dirichlet Allocation ed applicazioni in psicologia. *Modelli Latent Dirichlet Allocation ed applicazioni in psicologia*. Padova, Veneto, Italia: Unipd.

Caswell, I., & Bapna, A. (11 de mayo de 2022). *AI Google Blog*. Recuperado el 10 de 7 de 2023, de Unlocking Zero-Resource Machine Translation to Support New Languages in Google Translate: https://ai.googleblog.com/2022/05/24-new-languages-google-translate.html

Chowdhury, G. (2003). Natural language processing. *The Annual Review of Information Science and Technology*, 51-89.

Cronquist, K., & Fiszbein, A. (2017). *English Language Learning in Latin America.* s.r: The Dialogue.

DataScientest. (27 de mayo de s.a.). *¿Qué es un DataFrame?* Recuperado el 8 de 1 de 2023, de DataScientest: https://datascientest.com/es/que-es-un-dataframe#:~:text=A%20diferencia%20de%20las%20Series,Pandas%20indexadas%20por%20un%20valor.

DeepAI. (s.f. de s.f. de s.f.). *Named Entity Recognition Explained*. Recuperado el 15 de 8 de 2023, de DeepAI: https://deepai.org/machine-learning-glossary-and-terms/named-entity-recognition

Dialani, P. (29 de octubre de 2020). *The Future of Data Revolution will be Unstructured Data*. Recuperado el 28 de 7 de 2023, de Analytics Insights: https://www.analyticsinsight.net/the-future-of-data-revolution-will-be-unstructured-data/

Eugenio-Martin, J. L., Morales Martin, N., & Scarpa, R. (2004). *Tourism and Economic Growth in Latin American Countries: A Panel Data Approach.* Milano: Fondazione Eni Enrico Mattei.

Firican, G. (28 de Mayo de 2020). *The history of big data*. Recuperado el 28 de 7 de 2023, de LightsOnData: https://www.lightsondata.com/the-history-of-big-data/#:~:text=In%202005%2C%20Big%20Data%20was,handle%20Big%20Data%2C%20was%20created.

Fundación Universia. (17 de junio de 2020). *La importancia del inglés en el mundo laboral*. Recuperado el 10 de 7 de 2023, de Universia: https://www.universia.net/pe/actualidad/estudiar-en-el-extranjero/importancia-ingles-mundo-laboral-842796.html

Hablacultura. (s.f de s.f de s.f). *Spanglish o Espanglish*. Recuperado el 18 de 7 de 2023, de Hablacultura: https://hablacultura.com/cultura-textos-aprender-espanol/cultura/espanglish-o-spanglish/

Halevi, G., & Moed, H. (2012). The evolution of big data as a research and scientific topic: Overview of the literature. *Research Trends*, 3-6.

Howard, H. (s.f. de s.f. de 2009). *Computer Science 831: Knowledge Discovery in Databases. Clústering*. Recuperado el 28 de agosto de 2023, de URegina: https://www2.cs.uregina.ca/~dbd/cs831/notes/clústering/clústering.html

IBM Cloud. (s.f de s.f de s.f). *What is natural language processing (NLP)?* Recuperado el 10 de 7 de 2023, de IBM: https://www.ibm.com/topics/natural-language-processing

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 15169-15211.

JISC. (2008). *Text Mining.* Manchester: JISC. Recuperado el 28 de 7 de 2023, de https://www.webarchive.org.uk/wayback/archive/20140613220103/http://www.jisc.ac.uk/media/documents/publications/bptextminingv2.pdf

Jurafsky, D., & Martin, J. H. (2023). Sequence Labeling for Parts of Speech and Named Entities. En D. Jurafsky, & J. H. Martin, *Speech and Language Processing* (pág. 6). Upper Saddle River, New Jersey: Pearson. Prentice Hall. Obtenido de https://web.stanford.edu/~jurafsky/slp3/8.pdf

KeepCoding. (23 de Febrero de 2023). *PoS tagging con spaCy*. Recuperado el 1 de 8 de 2023, de KeepCoding: https://keepcoding.io/blog/como-funciona-el-pos-tagging-con-spacy/

Krasina, E. A., & Jabballa Mahmoud, M. X. (2018). Code switching: State of Art. *RUDN Journal of Language Studies, Semiotics and Semantics*, 403-415.

Krogstad, J. M., & Gonzales-Barrera, A. (24 de marzo de 2015). *A majority of English-speaking Hispanics in the U.S. are bilingual*. Recuperado el 14 de 7 de 2023, de Pew Research Center: https://www.pewresearch.org/short-reads/2015/03/24/a-majority-of-english-speaking-hispanics-in-the-u-s-are-bilingual/

Lillibridge, G. D. (1966). The American Impact Abroad: Past and Present. *The American Scholar*, 39-63.

Luhn, H. P. (1958). A Business Intelligence System. *IBM Journal*, 314-319. Obtenido de https://www.ibm.com/watson/assets/pdfs/ibmrd0204H.pdf

Maggini, M. (s.f.). *Natural Language Processing. Parte 2: Part of Speech Tagging.* Bologna: Università di Bologna - Tecnologie per l'elaborazione del linguaggio. Recuperado el 15 de 8 de 2023, de

https://www3.diism.unisi.it/~maggini/Teaching/TEL/slides/06%20-%20NLP%20-%20PoS%20Tagging.pdf

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural language processing.* Londres: The MIT Press.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge: Cambridge University Press. Obtenido de https://nlp.stanford.edu/IR-book/

Moez, A. (s.d de Mayo de 2023). *NLTK Sentiment Analysis Tutorial for Beginners*. Recuperado el 1 de 8 de 2023, de Datacamp: https://www.datacamp.com/tutorial/text-analytics-beginners-nltk

Moreno, G. (17 de Abril de 2019). *A la espera de un Big Bang de datos*. Recuperado el 28 de 7 de 2023, de Statista: https://es.statista.com/grafico/17734/cantidad-real-y-prevista-de-datos-generados-en-todo-el-mundo/

Muysken, P. (2000). *Bilingual speech a typology of code-mixing.* Cambridge: Cambridge University Press.

NLTK. (s. f. de s. f. de s. f.). *NLTK*. Recuperado el 1 de 8 de 2023, de NLTK: https://www.nltk.org/

ODSC - Open Data Science. (17 de diciembre de 2018). *Unsupervised Learning: Evaluating Clústeres*. Obtenido de Medium: https://odsc.medium.com/unsupervised-learning-evaluating-clústeres-bd47eed175ce

Office of the United States Trade Representative. (10 de julio de 2022). *https://ustr.gov/*. Recuperado el 10 de 7 de 2023, de Western Hemisphere: https://ustr.gov/countries-regions/americas

Oracle Cloud Infrastructure. (22 de julio de 2023). *The Evolution of Big Data and the Future of the Data Platform.* Obtenido de Oracle: https://www.oracle.com/a/ocom/docs/big-data/big-data-evolution.pdf

Perez, C. (29 de junio de 2015). *US has more Spanish speakers than Spain*. Recuperado el 14 de 7 de 2023, de New York Post: https://nypost.com/2015/06/29/us-has-more-spanish-speakers-than-spain/

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: toward a typlogy of code-switching. *Linguistics*, 581-618.

Proto, S. (s.f. de Abril de 2018). Enhancing topic modeling through Latent Dirichlet Allocation with self-tuning strategies. *Enhancing topic modeling through Latent Dirichlet Allocation with self-tuning strategies*. Torino, Piemonte, Italia: Politecnico di Torino.

Quezada Narvaéz, C. (2011). La popularidad del inglés en el siglo XXI. *Tlatemoani - Revista academica de investigación*, 4. Obtenido de https://www.eumed.net/rev/tlatemoani/05/cqn.pdf

Salomé Guardione, M. (8 de Enero de 2019). *¿Qué países de América Latina tienen un mejor y peor inglés?* Recuperado el 10 de 7 de 2023, de EF English Live: https://englishlive.ef.com/es-mx/blog/ingles-en-la-vida-real/america-latina-paises-hablan-ingles/

Savaiano, G. F. (1950). *The Teaching of English in Latin America. The Modern Language Journal*, 51–54. doi:https://doi.org/10.2307/318962

Sierra Martínez, G. E. (2015). *Introducción a los corpus lingüísticos.* Ciudad de México: Instituto de Ingeniería. UNAM. Obtenido de http://www.corpus.unam.mx/cursocorpus/LibroCorpus.pdf

Soto Martínez, V. (2020). *Identifying and Modeling Code-Switched Language.* Columbia University, Computer Science. Columbia: Columbia University. Recuperado el 18 de 7 de 2023, de https://academiccommons.columbia.edu/doi/10.7916/d8-2zmf-9t73

Statista. (13 de septiembre de 2022). *Reddit - Statistics & Facts*. Recuperado el 31 de 7 de 2023, de Statista: https://www-statista-com.bibliopass.unito.it/topics/5672/reddit/#topicOverview

Tobella, P. (28 de julio de 2021). *¿API es lo mismo que web scraping?* Recuperado el 31 de 7 de 2023, de Octoparse: https://www.octoparse.es/blog/api-es-lo-mismo-que-web-scraping

Twitter. (14 de julio de 2019). Recuperado el 14 de 7 de 2023, de Twitter: https://twitter.com/AOC/status/1207370145573867521?s=20

Universidad de Granada. (s.f.). *Introducción al Análisis Clúster*. Granada: Universidad de Granada. Obtenido de https://www.ugr.es/~gallardo/pdf/clúster-g.pdf

Universidad de Málaga. (1 de Mayo de s.a). *¿Qué es el Text Mining y cuáles son sus aplicaciones?* Recuperado el 28 de 7 de 2023, de Máster en Formación Permanente en Big Data e Inteligencia Artificial: https://www.bigdata.uma.es/que-es-el-text-mining-y-cuales-son-sus-aplicaciones/

Universidad Mediterránea de Reggio Calabria. (s.f.). *Il Clústering*. Reggio Calabria: Università Mediterránea di Reggio Calabria. Obtenido de https://www.unirc.it/documentazione/materiale_didattico/599_2008_93_1623.pdf

Weber, M. (2003). Text mining suggests new uses for thalidomide. *J Am Med Inform Assoc.*, 252-259.

Wikipedia. (1 de julio de 2021). *Stemming*. Recuperado el 1 de 8 de 2023, de Wikipedia: https://es.wikipedia.org/wiki/Stemming

World Tourism Organization. (10 de julio de 2019). *Dashboard de datos turísticos de la OMT*. Recuperado el 10 de 7 de 2023, de UNWTO: https://www.unwto.org/es/omt-dashboard-datos-turisticos

Zanini, N., & Dhawan, V. (2015). Text Mining: An introduction to theory and some applications. *Research Matters: A Cambridge Assessment publication*, 38-44. Recuperado el 28 de 7 de 2023, de https://www.cambridgeassessment.org.uk/Images/466185-text-mining-an-introduction-to-theory-and-some-applications-.pdf

Apéndice 1. Lista de los *Subreddits* Utilizados

Argentina:
- republica_argentina; alianzaargentina; buenosaires; argentina; cordoba; bahiablanca; bariloche; chubut; corrientes; lapampa; mendoza; rosario; salta; tucuman; neuquen.

Bolivia:
- bolivia

Chile:
- republicadechile; chile; santiago; yo_ctm; noeslalegal; anormaldayinchile; chilefit; clubdelecturachile; chileambiental; chileorganico.

Colombia:
- colombia; bogota; medellin; barranquilla; cali; bucaramanga; manizales; pereira; santamarta.

Costa Rica:
- ticos

Cuba:
- cuba

Ecuador:
- ecuador

El Salvador:
- el salvador

Guatemala:
- guatemala

Honduras:
- honduras

México:
- mexico; monterrey; guadalajara; mexicocity; tijuana; puebla; videojuegosmx; memexico; mexicofinanciero; derechomexicano; somosmexico.

Panamá:
- panama

Paraguay:
- paraguay

Perú
- peru; cusco; machupicchu; arequipa; cumbiaperuana; pokemongoperu.

Puerto Rico:
- puerto rico

República Dominicana:
- dominicanos

Uruguay
- uruguay; uruguay marketplace; burises; uruguay libre; uruguay crypto; uruguay circle jerk; uruguay verde; charruadevs.

Venezuela
- venezuela; vzla.

Apéndice 2. Frecuencia de los Anglicismos más Frecuentes por *Subreddit*

Tabla 1

Frecuencia de los anglicismos por subreddit

País	Anglicismo	Frecuencia anglicismo	Total *tokens*
Argentina	post	145	12206
Bolivia	edit	17	1171
Chile	format	116	6576
Colombia	government	74	4996
Costa Rica	government	16	1151
Cuba	twitter	34	1051
República Dominicana	government	22	1245
Ecuador	boric	20	1110
Guatemala	sub	28	1126
Honduras	thread	141	1107
México	would	99	13811
Nicaragua	good	13	1214
Panamá	like	22	1469

País	Anglicismo	Frecuencia anglicismo	Total *tokens*
Paraguay	people	18	1236
Perú	like	21	1824
Puerto Rico	get	17	1076
El Salvador	city	21	1322
Uruguay	like	70	4956
Venezuela	would	38	2232

Apéndice 3. Relaciones de EE. UU. con los Países de Interés

Tabla 2

Relaciones de EE. UU. con los países de interés

País	Turismo	Comercio
Argentina	2	33
Bolivia	2	96
Chile	2	20
Colombia	2	22
Costa Rica	1	38
Cuba	2	127
República Dominicana	1	30
Ecuador	2	42
El Salvador	2	50
Guatemala	2	36
Honduras	2	44
Nicaragua	2	67
México	1	2
Paraguay	2	109
Perú	2	29
Panamá	2	35
Puerto Rico	1	0
Uruguay	2	91

Venezuela	2	72

.UBAeconómicas posgrado

ENAP Escuela de Negocios y Administración Pública

Evaluación de la Mentora

Trabajo Final Integrador de Especialización
Code-switching en los foros latinos hispanohablantes de Reddit

Autora: Jessica Feggio
Mentora: Melisa Elfenbaum

Reporte:

El crecimiento sostenido de acceso a la información de los últimos veinte años está generando nuevos desafíos y necesidades en el ámbito organizacional. Frente a estos nuevos retos, y sabiendo que la cantidad de datos seguirá aumentando, es necesario desarrollar herramientas que permitan realizar la recolección, análisis y predicción de datos con formas que anteriormente no existían.

En este contexto, el Code-switching en los foros latinos hispanohablantes de Reddit es un tema interesante que todavía no ha sido muy investigado, en relación a otras áreas de la inteligencia artificial. En cuanto a la definición del problema, se realiza de forma clara y concreta, con la ventaja de saber desde el principio con que set de datos se va a realizar el trabajo de investigación.

Con respecto a los objetivos, tanto el general como los específicos son adecuados y consistentes a las hipótesis planteadas, las cuales también se encuentran bien estructuradas. Por otra parte, con el tema elegido y la correspondiente identificación del problema, logra incorporar el conocimiento adquirido en distintas asignaturas de la especialización.