UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

ESTUDIO DE LOS CAMBIOS DE PERCEPCIÓN RELACIONADOS A EMPRENDIMIENTO UTILIZANDO PROCESAMIENTO DEL LENGUAJE NATURAL.

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

IGNACIO ALEJANDRO MEZA DE LA JARA

PROFESOR GUÍA:
Felipe Bravo Márquez

MIEMBROS DE LA COMISIÓN:
Jorge Silva Sánchez
Andrés Caba Rutte

Este trabajo ha sido parcialmente financiado por:
FONDECYT 11200290

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: **IGNACIO ALEJANDRO MEZA DE LA JARA**
FECHA: 2022
PROF. GUÍA: Felipe Bravo Marquez

ESTUDIO DE LOS CAMBIOS DE PERCEPCIÓN RELACIONADOS A EMPRENDIMIENTO UTILIZANDO PROCESAMIENTO DEL LENGUAJE NATURAL.

El análisis del cambio de la percepción del emprendimiento es un desafió que depende de las barreras culturales y del lenguaje o modismos presentes en cada región. Debido a que hoy en día los emprendimientos son un pilar fundamental en las economías mundiales, la comprensión de este concepto se vuelve una tarea relevante para visualizar tanto impactos económicos de un país, como la evolución que han experimentado los grupos sociales que componen diferentes comunidades.

En esta memoria se tiene como objetivo abstraer información relevante para conceptos relacionados a emprendimiento utilizando documentos noticiarios de habla inglesa. Para esto, se aplicarán técnicas de inteligencia computacional, en busca de visualizar una potencial variación en la interpretabilidad de este concepto para los diferentes países que componen el corpus, y observar la evolución de sesgo que han experimentado diferentes grupos sociales a lo largo de los años.

Para lograr el objetivo señalado, el código es elaborado en el lenguaje computacional *python*, donde son aplicadas técnicas de procesamiento del lenguaje natural sobre los documentos noticiarios. Con esto, se propone la desambiguación de conceptos relacionados a emprendimiento destacando el año de su aparición, para luego obtener la representación vectorial de las palabras presentes en los documentos a través de *Word2Vec*, para finalmente trabajar con las características obtenidas utilizando técnicas clásicas de *machine learning*.

Los resultados obtenidos logran relacionar eventos históricos con la variación de la percepción sentimental asociada a emprendimiento a lo largo de los años. Además, es posible visualizar una interpretación desigual de conceptos relacionados a emprendimiento para los diferentes países estudiados, y una clara tendencia de sesgos relacionados a género y religión.

Study hard what interests you the most in the most undisciplined, irreverent and original manner possible.

Richard P. Feynman

Agradecimientos

Tras vacilar muchos años sobre mi futuro, logro por fin ver el comienzo de mil caminos que se vienen por delante. Si bien, este trabajo forma un final de un ciclo, no significa mas que un comienzo de lo que me han enseñado las experiencias que he vivido estos dos ultimo años. Es impresionante como las personas correctas pueden influir en tu vida, mostrándote caminos que jamás habías pensado y por eso quiero agradecerles a todos ustedes.

En primer lugar, quiero reconocer todo el apoyo y la crianza dada por mis dos madres: Mi madre biológica Solange y mi segunda madre de crianza, mi abuela Carmen. Sin el apoyo que me han brindado día a día no hubiese podido llegar a ser lo que soy hoy en día, y sin duda mis caídas hubiesen sido más dolorosas sin su soporte. Por otro lado, quiero agradecer a mi abuelo, quien es un segundo padre para mí, quien desde pequeño me ha brindado todo el amor que ha podido y hasta el día de hoy soporta mis molestias.

Quiero agradecer a mi polola Vanesa. Su compañía, su cariño y su apoyo incondicional a mis indecisiones de que estudiar o que hacer con mi vida. Las experiencias vividas juntos sin duda han aportado en mi crecimiento como persona y te agradezco de corazón el apoyo que constante me das cuando tengo un problema.

En tercer lugar, quiero dar mis agradecimientos a Luis y Cristian, amigos del alma, que quizás por diferentes circunstancias perdimos el contacto por un tiempo. Ustedes fueron un impulso importante para en la forja de este camino, ya que gracias a ustedes par de payasos escogí salir del fracaso y actuar por lo que quería.

No se quedan afuera los amigos que hice durante este viaje, con quienes compartí risas, tocatas en las que fueron palos blancos, partidas de LOL, APEX (inserté videojuegos que sacan todo lo malo de nosotros) y muchas alegrías y/o decepciones durante nuestro paso en la universidad. Los quiero un montón y muchas gracias por su compañía estos años.

Finalmente agradecer la oportunidad entregada por Amanda Williamson y mi profesor guía (y mentor) Felipe Bravo. Especialmente quiero agradecer a Felipe Bravo, persona que admiro notablemente y me ha enseñado un montón durante el año y medio que lo he conocido; dándome casi un intensivo de conceptos relacionados a inteligencia computacional que hicieron reencantarme por un área que me había decepcionado.

Como diría Cerati: ¡Gracias Totales!

Tabla de Contenidos

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

En el presente capitulo son expuestos los lineamientos seguidos durante este trabajo de investigación. Para esto, se comienza describiendo el problema que motiva a realizar el trabajo investigativo, para luego presentar los objetivos generales, específicos y los alcances que tendrá este trabajo. Así, finalmente se presenta la estructura con la que se organiza el documento, señalando de forma breve los contenidos que se abordaran en cada uno de los tópicos.

1.1. Motivación y Antecedentes

La percepción que tenemos de nuestro vocabulario está en constante construcción y depende de la sociedad en la que estamos envueltos, el lenguaje que se habla [1] y el desarrollo de los sesgos sociales que han sido aplicados a lo largo de los años en ella [2]. El efecto de la constante variación de la percepción del lenguaje genera en las sociedades importantes impactos que afectan en el desarrollo de estas. Este factor es relevante en temas de interacción entre los ciudadanos que conforman a estas sociedades, pudiendo afectar directamente al desarrollo económico del país, debido al impacto generado por la interpretación de las oportunidades para generar negocios [4].

Hoy en día uno de los principales instrumentos que expone la percepción social son los medios de comunicación. De ellos se reconoce como uno de los más relevantes al periódico, quien es utilizado en la mayoría de los países, logrando registrar de una forma "objetiva" las opiniones y los eventos acontecidos. Actualmente este medio ha evolucionado conjunto a las formas de comunicarse, pasando a una plataforma *online* que permite la libre lectura de sus artículos y el desarrollo en tiempo real de eventos importantes. Las noticias se caracterizan por poseer excelentes formas de captar a los lectores [5], lo que produce que un gran calibre de la población se informe por estos medios y genere una percepción asociada a las noticias expuestas [6].

Actualmente uno de los principales pilares en las economías mundiales son los emprendimientos, quienes según un estudio realizado por la Agencia Federal para el Desarrollo de la Pequeña Empresa de los Estados Unidos [3], señala que estos tienen un rol importante en la entrega de trabajos y generación de ingresos para el país. Es por esta razón que comprender el cómo, por qué y cuando se generan los emprendimientos se vuelve una tarea relevante para entender factores de crecimiento en un país. Producto que los emprendimientos son generados en su mayoría por personas particulares, la toma de decisiones para generar negocios va

de la mano de como los medios de comunicación exponen la información e influencian a las personas. Aquí yace la relevancia de estudiar las noticias expuestas por los medios de comunicación, especialmente para comprender un tópico tan relevante como es el emprendimiento.

Por lo comentado, en este trabajo será analizada la percepción de conceptos relacionados a *entrepreneurship* (emprendimiento en inglés) en base a diferentes corpus de noticias de habla inglesa. Para esto se busca responder, ¿cómo varia la percepción asociada a emprendimiento a lo largo de los años?, ¿Existen sesgos étnicos, religiosos o de género en la generación de emprendimientos? y ¿cómo es percibido el concepto *entrepreneurship* entre los diferentes países que compone nuestro corpus?. Es pertinente señalar, que el conjunto de corpus utilizado en este trabajo es una recopilación de noticias enfocadas en emprendimiento de los países: Australia, Reino Unido, Estados Unidos, Nueva Zelanda, Irlanda y Canadá. El corpus en total suma un conjunto de 370.363 artículos noticiarios y maneja un desbalance considerable entre países. La diferencia en el número de datos por corpus refleja una diferencia en las décadas de noticias que se manejan para cada uno de los países.

Si bien la tarea de abstraer información sobre miles de documentos noticiarios pudo haber resultado una odisea para una persona hace años atrás; ya que, esto hubiese implicado la lectura de múltiples textos y toma de apuntes sobre potenciales relaciones entre todas las palabras presentes en los documentos. Hoy en día gracias al desarrollo en el área de *machine learning* de procesamiento del lenguaje natural (PLN), es posible obtener información relevante desde los textos a través de múltiples técnicas. Una de las primeras técnicas que permitió abstraer información lingüística desde textos es conocida con el nombre *bag of word* en [8], quien propone el conteo de la frecuencia de las palabras en un corpus para obtener potenciales similitudes entre documentos con un número similar de palabras. Un método más elaborado es propuesto en [9] donde se propone la obtención de *word embeddings*, los cuales son representaciones vectoriales de las palabras obtenidas a través de un entrenamiento realizado sobre una red neuronal superficial. Técnicas similares a esta han sido propuestas en [11], quienes proponen la obtención de n-gramas de las letras que componen a cada una de las palabras del corpus, para obtener una representación vectorial en un conjunto de palabras más amplio y sin la necesidad que aparezcan en el corpus. Finalmente técnicas más elaboradas en [12] y [13] proponen el uso de técnicas de *deep learning* para la obtención de características sobre documentos; estos algoritmos son capaces de obtener características lingüísticas más ricas pero requieren de entrenamientos y corpus más grandes para la obtención de resultados.

Uno de los principales objetivos de este proyecto es abstraer la interpretabilidad asociada a conceptos relacionados a emprendimiento a lo largo de los años, con esto, si bien no existen trabajos que desarrollen un estudio basado en la evolución de la interpretabilidad del tópico de interés, existen múltiples estudios que logran extraer información relevante de los textos que pueden servir de referentes. Un trabajo destacado es el realizado por Hamilton et. al [14], quien expone una evolución de la semántica en décadas, utilizando *word embeddings* y unas matrices de coocurrencia para visualizar los cambios de la semántica de los textos, el resultado obtenido en este trabajo señala que el vocabulario evoluciona, visualizando diferentes interpretaciones de un mismo concepto en el transcurso de los años. Un trabajo similar es el realizado por Jatowt et. al [20], donde los autores demuestran la evolución de la semántica a través de la similitud de coseno, validando este proceso por medio de un recurso léxico que agrupa el sentimiento percibido de un conjunto de palabras por décadas.

A continuación son expuestos algunos trabajos relevantes en la visualización de sesgos presentes en emprendimientos, cabe señalar que estos trabajos no poseen la aplicación de algoritmos de *machine learning*, pero sirven como referentes para abstraer ideas y estar al tanto de conclusiones relevantes obtenidas en áreas externas. Unos primeros trabajos son expuestos en [32] y [33], donde a través de entrevistas encuentran que el emprendimiento se ve típicamente como un campo masculino; señalando que tanto hombres como las mujeres ven a los emprendimientos como un campo de ocupación masculina. En estudios más recientes, [34] propone evaluar la percepción que se tiene de los hombres y mujeres respecto a los emprendimientos, para esto utiliza evaluaciones realizadas por gerentes y los efectos marginales generados por sus evaluaciones, llegando a visualizar diferencias significativas en las características asociadas a hombres y mujeres.

Por otro lado, algunos estudios de sesgo realizados en el área de Inteligencia computacional por Caliskan et al. [16] demuestran a través de técnicas estadísticas de *machine learning* que los sesgos históricos en el texto se ven heredados del corpus; con esto proponen técnicas para identificar estas huellas. Por otro lado, en Garg et al. [15] se propone la métrica de *relative norm distance* (RND) para medir la distancia entre un concepto de consulta (palabra o conjunto de palabras a revisar el sesgo) y grupos sociales de interés para identificar el sesgo asociado a estos grupos; dentro de los resultados este es uno de los más relevantes, ya que logran exponer como la variación de los sesgos étnicos y de genero varían a través del tiempo. En Badilla et al. [21] se propone un *framework* con el que es capaz de obtener fácilmente la huella de sesgo de un corpus a través de múltiples métricas en base a word-embeddings pre-entrenados. Nuestro trabajo utiliza la idea de variación temporal de sesgos para encontrar hallazgos relevantes en el cambio de percepción de emprendimiento.

Para el caso del análisis sentimental de textos, son múltiples los trabajos que realizan esto. En Kolchyna et al. [17] y Taj [20] predicen el sentimiento asociado a oraciones en base a clasificadores previamente entrenados con *word embeddings* y lexicones. Por otro lado, trabajos más complejos son desarrollados en Liu et al. [19], donde proponen la clasificación *multi-label* de texto, comparando múltiples algoritmos de clasificación de múltiple etiqueta.

Acorde a lo señalado, el presente trabajo toma elementos relevantes de cada uno de los estudios señalados para obtener información significativa desde de los corpus de noticias. De esta forma, el trabajo consistirá en la aplicación de múltiples técnicas de inteligencia computacional para realizar desarrollar un estudio social relacionado a emprendimiento.

1.2. Descripción del problema

En este trabajo se tiene interés en encontrar características relevantes que nos permitan extraer información de la evolución que ha desarrollado conceptos relacionados a emprendimiento en los últimos años. Para esto, proponemos la aplicación de técnicas de procesamiento del lenguaje natural como *word embeddings*, para abstraer atributos relevantes desde corpus de noticias de habla inglesa. En base a los atributos obtenidos, se espera aplicar modelos clásicos de *machine learning* para clasificar o encontrar tendencias significativas en los datos. El resultado de los métodos aplicados en esta investigación pretende dar una mejor comprensión de la interpretabilidad de términos relacionados a emprendimiento para los diferentes países

que componen el corpus, como también visualizar potenciales sesgos sociales generados en torno a este tópico a lo largo de los años (e.g., encontrar que los emprendimientos están más relacionados al género masculino que al femenino).

Parte de los desafíos para realizar este estudio se encuentra la aplicación de una metodología robusta que nos permitan abstraer información relevante de los datos, con esto, la elección equivocada de metodologías podría significar la abstracción de conceptos errores de los datos, generando un análisis irrelevante de estos mismos. Por otro lado, se encuentra la valoración de los resultados obtenidos, por lo que al tratarse de un tema del área comercial se necesita opinión de expertos relacionados a esta área para discernir si estos son o no relevantes, por lo que el trabajo multidisciplinario será esencial en este trabajo.

1.3. Objetivos

1.3.1. Objetivos Generales

Obtener información descriptiva sobre la evolución de la semántica y sesgo social asociado al concepto de emprendimiento, visualizando un estudio a lo largo de años en países de habla inglesa, utilizando técnicas de procesamiento del lenguaje natural.

1.3.2. Objetivos Específicos

- Entrenar un modelo de *Word Embeddings* en base a un corpus de noticias para obtener representaciones vectoriales de las palabras.
- Visualizar la variación de interpretabilidad asociada a términos relacionados a emprendimiento para diferentes países que comparten el mismo idioma.
- Verificar a través de un clasificador de sentimientos la presencia de periodos con percepción sentimental negativa para años de recesión económica.
- Evaluar a través de métricas de sesgo la asociación de emprendimiento a grupos de intereses como: genero, religión y etnia.

1.4. Aporte de la memoria

1.4.1. Alcances

El estudio realizado en este trabajo es enfocado en la búsqueda de características relevantes desde periódicos de habla inglesa. Con esto, se aprovecharán los algoritmos presentes en la literatura para encontrar conceptos relevantes que permitan comprender mejor la interpretabilidad, tanto temporal como atemporal, que existe en torno a términos relacionados a emprendimientos. Es importante señalar que no es parte del objetivo generar un algoritmo para predecir eventos futuros en donde sea más conveniente generar un emprendimiento.

Cada una de las aplicaciones realizadas en el desarrollo de este proyecto son desarrolladas utilizando el lenguaje computacional *Python*, esto debido al gran desarrollo que poseen

las librerías asociadas a inteligencia computacional para este lenguaje. Por esta razón, no será desarrollado ningún modelo matemático de inteligencia computacional, sino que serán aplicados solamente estos conceptos a través de librerías.

1.5. Estructura de la Memoria

El desarrollo de este avance de memoria está organizado en cuatro capítulos, donde en cada uno serán abordados los siguientes tópicos:

- Capítulo 2 - Marco Teórico y Estado del Arte: En este capítulo se presenta una breve descripción de los diferentes conceptos utilizados durante el trabajo de esta memoria. Para esto, se comienza haciendo una pequeña introducción de las áreas del saber involucradas en esta investigación, se introducen conceptos de redes neuronales y *word embeddings*, para terminar con la presentación de algunas aplicaciones relevantes con *word embeddings* para este trabajo.
- Capítulo 3 - Metodología: Se exponen los pasos a seguir para obtener los objetivos propuestos en el capítulo 1 de esta memoria.
- Capítulo 4 - Resultados y Análisis: Se exponen los resultados obtenidos para cada uno de los experimentos señalados en el capítulo anterior, comentando y analizando los aspectos más relevantes de los resultados obtenidos.
- Capítulo 5 - Conclusiones: Se verifica si los objetivos señalados fueron logrados en el trabajo realizado, especificando con ello las características más relevantes que se pudieron extraer desde el corpus de noticias. Además, se describe el trabajo a futuro que se podría realizar en base al tópico estudiado, señalando mejoras o enfoques que se podrían seguir.

Capítulo 2

Marco Teórico

Este capítulo está compuesto de tres temas principales. El primer tema para revisar está enfocado en introducir al lector a las áreas de conocimiento que aborda este trabajo. En segundo lugar, se describen arquitecturas primordiales para comprender como técnicas modernas abstraen información desde los textos. Mientras que, en tercer lugar, se elabora una descripción detallada de los conceptos utilizados para el desarrollo de esta investigación de acuerdo con el estado del arte. Por lo anterior, la estructura de este capítulo está dada por: una pequeña introducción de las áreas involucradas en esta investigación, revisión de la teoría existente para obtener características desde los textos, y para terminar se presentan técnicas de machine learning que aprovechan la información obtenida de los textos.

Cada uno de los puntos se desarrolla en el siguiente esquema: En la sección 2.1 se explica algunos aspectos relevantes del área de conocimiento de Machine Learning 2.1.1, para luego en 2.1.2 referirnos a cuáles son los objetivos de la disciplina de procesamiento natural del lenguaje.

Luego en la sección 2.2 se introducen los conceptos básicos de redes neuronales, presentado desde la estructura básica hasta una breve explicación del entrenamiento que se realiza sobre estas arquitecturas. Esta introducción, se realiza con el fin de explicar conceptos esenciales para más tarde comprender la obtención de vectores a través de *word embeddings* en la sección 2.3.

Finalmente, en la sección 2.5.5 se expone la motivación que posee la utilización de algoritmos de clasificación, exponiendo en 2.5.5.1, 2.5.5.2 y 2.5.5.3 algunos modelos de clasificación clásicos, los que acompañados de lexicones emocionales nos permitirán predecir sentimientos asociados a los *word embeddings*. Finalmente en 2.5.6 se exponen algunas métricas para evaluar el sesgo presente en documentos de texto.

Cabe señalar que este marco teórico se encuentra en desarrollo y son presentados solo los avances solicitados por el cuerpo docente del curso Introducción al Trabajo de Título. El trabajo final debería presentar una mayor elaboración de los tópicos expuestos y la inclusión de nuevos temas.

2.1. Áreas del conocimiento

En esta sección se describe brevemente las áreas del conocimiento relacionadas al trabajo realizado.

2.1.1. Machine Learning

Machine learning es un área de la inteligencia computacional y *computer science* que está enfocada en la programación de las computadoras para que estas aprendan características relevantes desde los datos. Una definición ingenieril de esta ciencia es dada por Tom Mitchell quien señala que:

> *"Se dice que un programa de ordenador aprende de la experiencia E con respecto a alguna tarea T y alguna medida de rendimiento P, si su rendimiento en T, medido por P, mejora con la experiencia E."* –Tom Mitchell, 1997

Con esto, el gran valor de estos algoritmos está en generar modelos matemáticos robustos que sean capaces de abstraer información relevante desde los datos, sin la necesidad de modelarlos. Esto a diferencia de un algoritmo basado en conjuntos de reglas para solucionar un problema nos permite abstraer características relevantes de los datos, permitiéndonos generar algoritmos con una mayor capacidad de generalización.

Respecto al entrenamiento de estos modelos, estos se caracterizan por encontrar a través de funciones de optimización los parámetros ideales que logran solucionar el problema deseado. Los parámetros por optimizar dependerán del modelo utilizado y del tipo de entrenamiento realizado.

Las formas de clasificar las técnicas de aprendizaje de *machine learning* son variadas, estas dependen de la cantidad de datos disponibles y de la supervisión que se tenga del entrenamiento. Con esto, se reconocen cuatro tipos de entrenamiento:

- Aprendizaje supervisado: Este tipo de aprendizaje se caracteriza por entregarle al algoritmo inteligente la solución deseada a través de un etiquetado. Estos algoritmos tienen el objetivo de predecir o clasificar nuevos datos de entrada.
- Aprendizaje no supervisado: Los datos de entrenamiento se encuentran sin etiquetados. Por lo general este tipo de aprendizaje se caracteriza por encontrar relaciones entre los datos de entrada.
- Aprendizaje semisupervisado: Es una combinación entre los dos aprendizajes ya anunciados. Este tipo de aprendizaje se da cuando solo parte de los datos de entrenamiento este etiquetado, siendo necesarios algoritmos no supervisados para agrupar los datos no etiquetados y de esta forma etiquetar los datos para generar finalmente un entrenamiento supervisado.
- Aprendizaje reforzado: Este tipo de algoritmos de aprendizaje se caracteriza por aprender por sí solos, para a través de un sistema de recompensas"guía al modelo a aprender la mejor estrategia para solucionar el problema.

2.1.2. Procesamiento del Lenguaje Natural

Una de las grandes ventajas que tenemos como seres humanos es el lenguaje. La capacidad de comunicarnos les entrego a nuestros antepasados ventajas competitivas en la caza, recolección y generación de comunidades frente a otras especias. Hoy en día, el humano ha desarrollado maquinas que utilizan sus propios lenguajes y son capaces de procesar grandes cantidades de información en pocos segundos. Esta capacidad genera la necesidad de producir métodos para hacer que el lenguaje humano sea accesible y/o entendible para las computadoras, aprovechando su capacidad de cómputo para solucionar problemas que involucran el lenguaje.

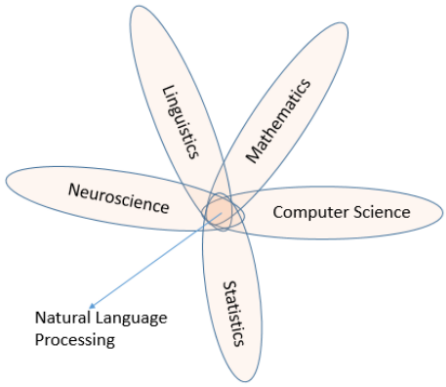


Figura 2.1: Áreas cubiertas por el procesamiento del lenguaje natural.

El procesamiento del lenguaje natural es un área interdisciplinaria que utiliza inteligencia computacional para la resolución de tareas que involucran el lenguaje humano en forma de texto o voz. Estas tareas son acotadas y poseen un conocimiento de la salida esperada del problema, algunos ejemplos son la traducción de texto, software de reconocimiento de voz, etc. Por otro lado, PLN no se debe confundir con la Computación Lingüística, la cual es un área que busca comprender el lenguaje humano a través de la computación, planteándose preguntas más profundas y científicas del lenguaje. O sea, el lenguaje es el objeto de estudio para la computación Lingüística.

Si bien PLN resulta una tarea interesante, esta no es una tarea simple. Ya que a pesar de que el ser humano es habilidoso para aprender lenguajes, hay una pobreza en el entendimiento y descripción de las reglas que rigen a estos sistemas. Por esto, trabajar con el lenguaje humano se vuelve una tarea ambigua, ya que el cambio de una palabra en una oración puede generar un cambio importante en la semántica de esta. Otro de los problemas que se le suma es el dinamismo que posee el lenguaje, ya que este posee una evolución gradual, que nos hace variar la forma de cómo comunicamos nuestras ideas dependiendo de factores como la cultura, época, rango etario, etc.

2.2. Redes Neuronales Artificiales

A lo largo de la historia el humano se ha inspirado en múltiples elementos de la naturaleza para generar inventos revolucionarios; vimos a un pájaro volar y quisimos volar, observamos a la planta de loto no mojarse y quisimos generar una prenda para no mojarnos. Es por esto por lo que no parece extraño la intención de simular el funcionamiento de nuestro propio cerebro. En 1943 el neurocientífico Warren McCulloch y el matemático Walter Pitts proponen en [22] la primera red neuronal artificial, la que de forma rudimentaria intentaba simular el funcionamiento de las redes neuronales animales a través de lógica proposicional. A pesar del gran impacto logrado por el trabajo señalado, las redes no lograron un eficiente desarrollo hasta 1986 con la aparición del algoritmo de backpropagation [23].

Tras el largo invierno vivido por las redes neuronales, gracias a los avances que ha experimentado el área de la computación, una red neuronal básica se representa como se expone en la figura 2.2. De esta es posible reconocer tres partes: La primera es una capa de entrada, que tiene el fin de recibir los datos con que se desea entrenar la red neuronal. En segundo lugar, se reconoce una capa oculta que puede estar conformada por una o más unidades básicas reconocidas como neuronas (algunos autores prefieren llamarla perceptrón). Finalmente se reconoce la capa de salida, quien entrega el resultado tras las aplicaciones de múltiples operaciones matemáticas en la red.

Uno de los puntos importantes a señalar de una red neuronal son las conexiones. Las conexiones en una red neuronal son representadas por la letra W simbolizando la palabra *weigth* (peso en español). Estos pesos entregan a cada una de las conexiones ponderaciones que se interpretan como la relevancia que tendrá el valor anterior para los estados próximos. Por otro lado, a cada una de las conexiones se le anexa un *bias*, quien es un ajuste adicional que entrega una no dependencia de los pesos provenientes de la capa anterior. Llegados a este punto, otros elementos relevantes es la nomenclatura que reciben las neuronas, estas están dadas por la letra h, las que a través de un súper índice señalan la capa en la que están ubicadas y con un sub-índice el número de neurona que representan en la capa.

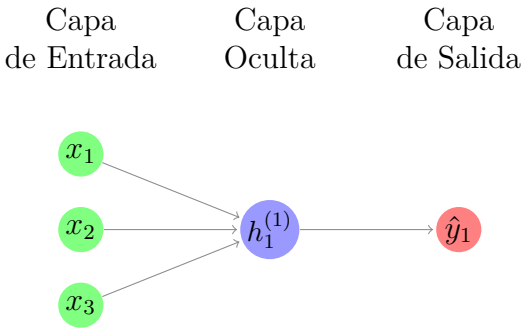


Figura 2.2: Red neunoral simple con cuatro entradas.

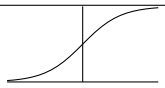
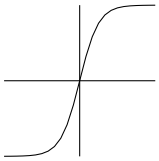
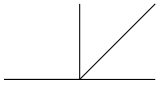
En base a lo señalado, la representación matemática de una red neuronal vendrá dada por el producto punto entre un vector de entrada de dimensionalidad n ($x = \{x_1, x_2, x_3, \ldots, x_n\}$) y una matriz de pesos con dimensión igual a los valores de entrada (n) por el número de neuronas a las que está conectada. Esto valor es sumado por el parámetro *bias*, quien se simboliza por la letra b y posee una dimensión igual al número de neuronas conectadas.

$$h_1^{(l)} = \vec{x}W + \vec{b} \tag{2.1}$$

Donde $\vec{X} \in R^{d_{in}}$, $\vec{W} \in R^{d_{in} \times d_{out}}$ y $\vec{b} \in R^{d_{out}}$

Obtenido $h_1^{(l)}$, se aplica sobre el resultado de la red neuronal funciones no lineales conocidas con el nombre de funciones de activación. Estas funciones cumplen el rol de exigir un umbral numérico para considerar activa la operación realizada en la ecuación 2.1. En otras palabras, estas funciones apagan o mantienen encendidas las neuronas con resultados sobre el umbral. La aplicación de estas funciones de activación depende mucho del tipo de aplicación, pero una de las más reconocidas es la función ReLU, quien entrega excelentes desempeños a los modelos de redes neuronales. A continuación, se listan las funciones de activación más relevantes:

Tabla 2.1: Funciones de activación no lineales.

Nombre	Función	Derivada	Gráfico
Sigmoide	$\sigma(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1 - f(x))^2$	
tanh	$\sigma(x) = \frac{e^x - e^{-x}}{e^z + e^{-z}}$	$f'(x) = 1 - f(x)^2$	
ReLU	$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0. \end{cases}$	$f'(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 1 \geq 0. \end{cases}$	
Softmax	$f(x) = \frac{e^x}{\sum_i e^x}$	$f'(x) = \frac{e^x}{\sum_i e^x} - \frac{(e^x)^2}{(\sum_i e^x)^2}$	

Con lo anterior y definiendo una función de activación cualquiera a través de la variable g, la salida una neurona estará dada por:

$$Z_1^l = g^l(\vec{x}W^l + \vec{b}^l) \tag{2.2}$$

Definidos los conceptos básicos de una red neuronal, podemos definir estructuras más complejas como las dadas por las redes *Multi Layer Perceptron* (MLP). Estas redes tienen como principal característica la presencia de múltiples capas ocultas y neuronas, entregando a la red una mayor robustez para solucionar problemas. Esta capacidad queda descrita en el trabajo realizado en [35], donde se señala que las redes neuronales multicapas son aproximadores universales capaces de generar cualquier función; en otras palabras, una arquitectura de red neuronal lo suficientemente grande puede aproximar cualquier función continua con un mínimo de grado de error.

De la figura 2.3 podemos visualizar la estructura que posee una MLP. Producto del aumento en las neuronas presentes en las capas ocultas son generadas matrices de pesos de mayor dimensionalidad, este factor producirá que el entrenamiento de las variables presentes en la arquitectura sea más complejo de realizar. Sin embargo, actualmente existen algoritmos de aprendizaje lo suficientemente robustos para poder encontrar los óptimos locales que minimizan el error de entrenamiento en estas redes.

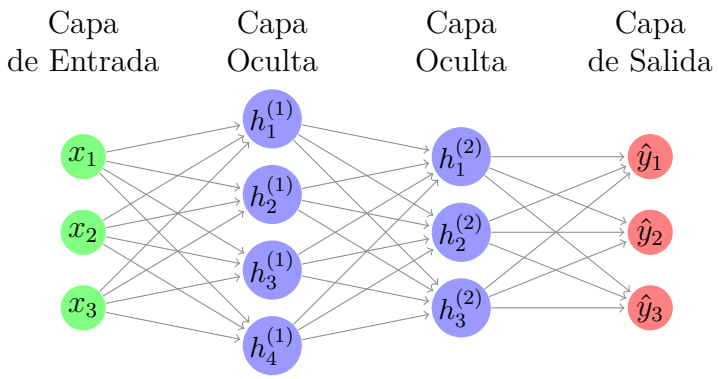


Figura 2.3: Red Feed Forward de dos capas ocultas.

Antes de revisar cómo funciona el entrenamiento de las MLP, a continuación. se define la representación matemáticamente de una red MLP:

$$NN_{MLP2} = g^2(g^1(\vec{x}W^1 + \vec{b^1})W^2 + \vec{b^2}) \tag{2.3}$$

Donde la salida esta dada por la aplicación de una función softmax sobre todas las operaciones realizadas en el "paso hacia adelante":

$$\hat{y} = Softmax(g^2(g^1(\vec{x}W^1 + \vec{b^1})W^2 + \vec{b^2})W^3 \tag{2.4}$$

Con lo anterior, como las redes neuronales son parte de un algoritmo basado en aprendizaje supervisado, se espera minimizar una función de error que entrega como valores el desempeño que posee la red neuronal para clasificar correctamente. Para realizar esta tarea, es utilizada la salida $\hat{y}$ y las etiquetas esperadas y, luego aplicando una función de error como la de *cross-entropy* 2.5, se obtienen los valores de desempeño de la red. Cabe señalar que existen múltiples funciones de perdida, para efectos de este marco teórico solo se ha considerada la más relevante del estado del arte actual, para más información sobre estas funciones se puede revisar material complementario en [37].

$$L_{cross-entropy}(\vec{\hat{y}}, \vec{y}) = -\sum_i \vec{y_i} log(\vec{y_i}) \tag{2.5}$$

Calculada la función de perdida es aplicado un proceso llamado Backpropagation donde son calculadas las gradientes de los parámetros W y b. Este proceso se realiza iterativamente y utilizando grafos computacionales, donde en cada iteración a través de las gradientes encontradas se actualizan los parámetros de entrenamiento a través de un algoritmo llamado descenso de gradiente estocástico (SGD). La principal característica de la técnica del descenso de gradiente radica en la búsqueda de los mínimos locales que minimizan los errores de la función de perdida 2.6.

$$\hat{\theta} = \text{argmin}\, L(\theta) \tag{2.6}$$

2.3. Word Embeddings

Los *word embedding* consisten en un conjunto de modelos que abstraen la representación semántica de una palabra a través un vector denso (sin ceros) de baja dimensionalidad. Debi-

do a que los vectores se basan en una hipótesis distribucional, estos representan información relevante como el contexto en el que ocurren, pudiendo ser útiles para identificar palabras que poseen un significado similar dentro del mismo contexto [36]. Si bien los modelos de *word embeddings* corresponden a la representación vectorial de palabras de un texto, estas pueden llegar a representar: oraciones, documentos y entre otros elementos lingüísticos a través de operaciones aplicadas sobre los vectores.

De forma general los *word embedding* son obtenidos a través del entrenamiento de redes neuronales en base a un corpus de texto. Donde los embeddings son tratados como un parámetro más de la red neuronal a ser entrenada. Debido al método de obtención de las caracteristicas, a diferencia de métodos clásicos, la dimensionalidad de los vectores obtenidos no es interpretable. Sin embargo, la capacidad de generar vectores densos y de baja dimensionalidad le entrega beneficios de generalización que no se lograban obtener con *sparse vectors* (vectores de alta dimensión), debido a la capacidad de generar representaciones vectoriales similares para palabras utilizadas en un mismo contexto [37].

2.3.1. Word2Vec

Como practicante de Microsoft, Thomas Mikolov, en el año 2012 descubre la forma de codificar la semántica de las palabras utilizando redes neuronales. Este descubrimiento se concretaría en el año 2013, donde junto al equipo de Google proponen los trabajos [9] y [10], proponiendo la creación de un software llamado *Word2Vec* capaz de generar una representación vectorial de las palabras en un corpus.

Word2vec es un software computacional que permite obtener *word embeddings* a través del entrenamiento de una red neuronal superficial. El software propone dos arquitecturas de redes neuronales para realizar la tarea de entrenamiento, una es *Continuous Bag of Words* (CBOW) y la segunda *Skip-gram*. Con esto, son utilizados dos algoritmos de optimización del modelo llamados *Negative Sampling* y *Hierarchical Softmax*.

Un modelo *skip-gram* es una red neuronal de aprendizaje no supervisado con solo una capa oculta (red neuronal superficial). Esta arquitectura tiene como objetivo predecir las palabras de contexto que rodean a una palabra central; para esto se utiliza una ventana de tamaño k que recorre todo el corpus de entrenamiento prediciendo las palabras que rodean a la palabra central. Del entrenamiento, el modelo es capaz de absorber del corpus las estadísticas de ocurrencia de las palabras del corpus, entregando como vector los pesos asociados al entrenamiento para cada una de las palabras que conforman el vocabulario.

De la figura 2.4 es posible observar una representación de la arquitectura de la red neuronal *skip-gram*. Con esto, es posible reconocer que la palabra central $w(t)$ es el parámetro de entrada de la red, mientras que las palabras de contexto que rodean a esta son las salidas a predecir por el modelo. Durante el entrenamiento de esta red cada una de las palabras del corpus se revisa por lo menos una vez, seleccionando en cada paso una palabra central nueva. Tal como en el proceso de entrenamiento de las MLP, las predicciones realizadas por la palabra central son comparadas con las palabras de contexto, de esta forma se calcula el error y se realiza el proceso de *backpropagation*. El proceso de entrenamiento se repite hasta que el

valor obtenido de la función de perdidas sea minimizado. Una vez finalizado el entrenamiento se obtienen los pesos de la red neuronal, quienes representan la representación vectorial de cada una de las palabras del corpus.

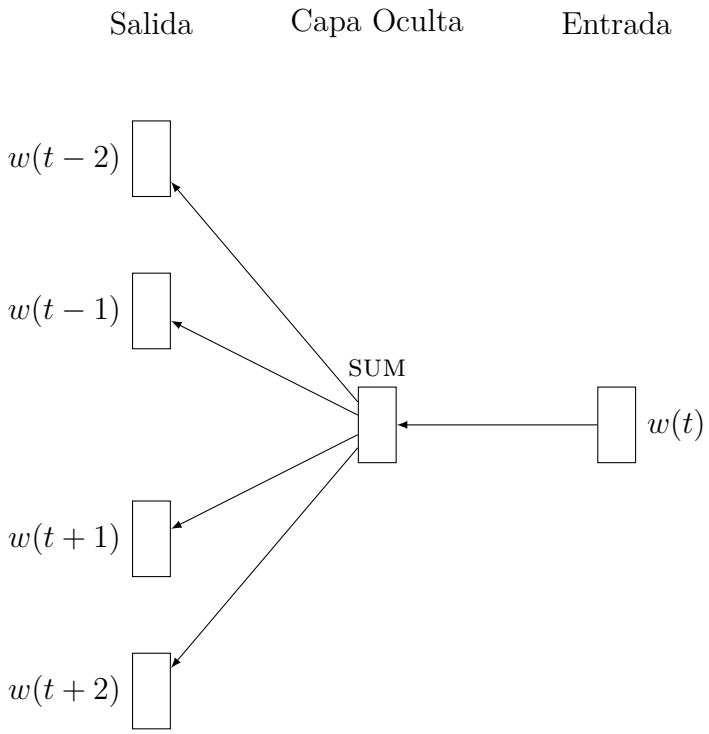


Figura 2.4: Representación de la red neuronal que da origen a *Skip-gram*

Para comprender mejor la teoría detrás del entrenamiento, supongamos que se tiene un corpus formado por la secuencia $w_1, w_2, w_3, \ldots, w_t$ y una ventana de contexto de tamaño k. Como ya se ha comentado, la palabra central es denotada por la letra c, mientras que las palabras contexto por $c_{1:k}$. Con lo anterior, la arquitectura de *skip-gram* tiene como objetivo maximizar la probabilidad logarítmica media de las palabras de contexto dadas por las palabras objetivo.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{c\in c_{1:k}} logP(c \mid w_t) \tag{2.7}$$

Sea C el conjunto de todas las palabras de contexto posibles (comúnmente es el mismo conjunto que el vocabulario). La probabilidad condicional de que una palabra de contexto c aparezca junto a la palabra central utilizando *softmax* queda dada por:

$$P(c \mid w) = \frac{e^{\vec{c}\vec{w}}}{\sum_{c\prime\in C} e^{\vec{c\prime}\vec{w}}} \tag{2.8}$$

Con lo anterior se deberá notar que $\vec{c}$ y $\vec{w}$ son los parámetros de entrenamiento del modelo. Donde al final del entrenamiento solo nos quedaremos con w, quien representa los pesos y *word embedding* de cada palabra.

Siendo D el conjunto de pares de palabra-contexto correctos (es decir, los pares de palabras que se observan en el corpus). El objetivo de optimización es maximizar la log-verosimilitud

condicional de los contextos c:

$$\arg\max_{\vec{C},\vec{w}} \sum_{(w,c)\in D} log\, P(c \mid w) = \sum_{(w,c)\in D} (log\, e^{\vec{c}\vec{w}} - log \sum_{c\prime\in C} e^{\vec{c\prime}\in\vec{w}}) \tag{2.9}$$

Se debe agregar que si asumimos el supuesto de que la maximización de la función 2.9 da lugar a buenas representaciones de *word embeddings*, las palabras similares tendrán una representación vectorial similar. Por otro lado, el término $P(c \mid w)$ posee un alto costo computacional debido a la suma $\sum_{c\prime\in C} e^{\vec{c\prime}\vec{w}}$ efectuada sobre todas las palabras de contexto $c\prime$. Este problema se suele solucionar con la sustitución de la softmax por una softmax jerárquica, la cual utiliza arboles de Huffman para reducir el número de salidas a evaluar.

Por otro lado, la arquitectura *Skipgram* con *Negative Sample* es una variación del problema anterior, donde es optimizada una nueva función [24] en busca de distinguir mejor el contexto desde el corpus. El nuevo objetivo de este modelo estará en maximizar la probabilidad que el par palabra-contexto (w, c), provenga del conjunto de pares palabra-contexto correctos D utilizando una función sigmoidea. En otras palabras, el modelo buscara distinguir los mejores par palabra-contexto del corpus. La formulación de esta tarea está dada por:

$$P(D = 1 \mid w, c_i) = \frac{1}{1 + e^{-\vec{w}\vec{c_i}}} \tag{2.10}$$

Luego, si asumimos que todas las palabras contexto c_i son independientes unas de otras, podremos tratar a cada par palabra-contexto como un ejemplo independiente para el entrenamiento. De esta forma la nueva función a optimizar, estará dada por:

$$P(D = 1 \mid w, c_{1:k}) = \prod_{i=1}^{k} P(D = 1 \mid w, c_i) = \prod_{i=1}^{k} \frac{1}{1 + e^{-\vec{w}\vec{c_i}}} \tag{2.11}$$

Conduciendo a la siguiente función objetivo:

$$\arg\max_{\vec{C},\vec{w}} log\, P(D = 1 \mid w, c_{1:k}) = \sum_{i=1}^{k} log \frac{1}{1 + e^{-\vec{w}\vec{c_i}}} \tag{2.12}$$

De la ecuación anterior es posible obtener una solución trivial no deseada, si establecemos que para todo el conjunto de palabra-contexto $P(D = 1 \mid w, c) = 1$, obteniendo en el entrenamiento que $\hat{w} = \hat{c}$. Para evitar este problema desestimamos algunas combinaciones palabra-contexto, presentando al modelo algunos (w, c) con una probabilidad $P(D = 1 \mid w, c)$ baja, por lo que se presentan pares que no posee el corpus. De esta forma, se consiguen muestras negativas del conjunto D, a través de un proceso en donde por cada par $(w, c) \in D$, se muestran m palabras $w_{1:m}$ y se añade cada una de las (w_i, c) como un ejemplo negativo de $\tilde{D}$ (palabras negativas). Con lo anterior, se obtiene la siguiente ecuación:

$$\arg\max_{\vec{C},\vec{w}} \sum_{(w,c)\in\tilde{D}} logP(D = 1 \mid w, c_{1:k}) + \sum_{(w,c)\in D} logP(D = 1 \mid w, c_{1:k}) \tag{2.13}$$

Finalmente, para entregar un peso relativo a las palabras menos frecuentes, las palabras negativas se suavizan por medio de la función *softmax* aplicada en las frecuencias del corpus:

$$\frac{\#(w)^{0.75}}{\sum_{w\prime}\#(w\prime)^{0.75}} \tag{2.14}$$

2.4. Limitaciones de los Word Embeddings

Como ya se ha comentado, los modelos de *Word Embdedding* nos ofrecen una plataforma que deriva las similitudes de las palabras de acuerdo con el contexto en el que fueron utilizadas. Sin embargo, esta plataforma al basarse en una hipótesis distributiva no es perfecta y pose múltiples limitaciones, Goldberg [28] en la sección 10.7 señala múltiples limitaciones derivadas de la hipótesis distributiva, las que son necesarias tener en cuenta al momento de trabajar con este tipo de representaciones.

2.4.1. Definición de Similitud

La similitud en los modelos distribucionales viene dada bajo el supuesto que las palabras son similares si estas son utilizadas en el mismo contexto. Esto difiere de la realidad, provocando que los modelos distribucionales proporcionen un nulo control sobre las similitudes que estas inducen.

Un ejemplo de esto puede darse con el conjunto de palabras *banana*, *apple* y *company*. Si deseamos encontrar la similitud respecto a frutas, *apple* debería ser más similar a *banana* debido a que ambas son frutas conocidas. Por otro lado, si buscamos la similitud de *apple* respecto a compañías, *apple* debería ser más similar a compañía al tratarse de una empresa ultra conocida. Si bien los humanos podemos interpretar y generar una correcta similitud dependiendo el caso, los modelos distribucionales carecen de un control en la similitud, pudiendo generar relaciones que carecen de sentido.

2.4.2. Ovejas Negras

Por eficiencia en la comunicación, las personas tienden a omitir la información conocida, expresando solo la información nueva si es necesaria. Esto produce que modelos computacionales se confundan e interpreten la nueva información como algo común dentro de un modelo distribucional. Un ejemplo de esto se expone en [28], donde se señala que cuando la gente habla de ovejas blancas, las personas tienden a asumir que el color de las ovejas y solo hablan de una oveja, mientras que para las ovejas negras es mucho más probable que señalen la información del color y digan oveja negra. Un modelo entrenado sólo estos datos de texto puede ser engañado y encontrar mayor similitud con términos que en la realidad no lo son.

2.4.3. Antónimos

Se sabe que los antónimos representan el significado opuesto de una palabra. Sin embargo, los sinónimos suelen utilizarse en los mismos contextos que los antónimos. Como consecuencia, los modelos basados en una hipótesis distributiva tenderán a tener obtener las mismas similitudes para palabras antónimas.

Este problema puede ser ilustrado con el siguiente ejemplo: supongamos que tenemos un texto conformado por las oraciones: *"he´s a bad boy" y "he´s a good boy"*, si bien el significado es completamente diferente, las palabras *bad* y *good* tendrán una alta relación al utilizarse en contextos similares.

2.4.4. Sesgo del Corpus

Los modelos distribucionales reflejan patrones extraídos desde el corpus de entrenamiento, esto es expresado en sesgos humanos del mundo real y obteniendo, para bien o para mal, estereotipos desde los datos. Con esto, múltiples estudios señalan que todos los textos poseen sesgos con los que se obtienen diferentes estereotipos.

Un ejemplo de esto se puede dar al analizar un corpus relacionado a trabajos de salud, donde el modelo reflejará que enfermeraøestá más relacionado a un trabajo femenino, mientras que doctor a uno masculino. Esto dependiendo de la tarea que se desee aplicar podría generar problemas, pero como en esta memoria nuestro objetivo es visualizar como se expresan lo periódicos de los emprendimientos, esta limitación puede dejar en claro carencias comunicacionales en los medios, que indirectamente podrían afectar a las personas.

2.4.5. Falta de Contexto

Debido a que los modelos distribucionales agregan el contexto en el cual los términos han sido utilizados en un corpus. Se obtienen como resultados representaciones aprendidas independientes del contexto. Esto implica que, cuando se utilizan, no pueden obtener información de su contexto. Esto es un problema importante porque muchas palabras tienden a variar su significado según el contexto que las rodea y representarlas únicamente como un vector podría generar problemas de significado.

Este problema lo podemos evidenciar cuando las palabras poseen múltiples significados (polisemias). Por ejemplo, la palabra *star* por un lado puede representar a un astro espacial. Mientras que, por otra parte, la misma palabra puede representar a algún artista famoso, una figura abstracta, un premio u otro concepto.

Como podemos ver existen múltiples limitaciones en el uso de *word embeddings*, si bien estas limitaciones pueden llevar a malas conclusiones del lenguaje como lo conocemos, estas las podemos aprovechas para abstraer información representativa de los textos estudiados. Con esto, como el trabajo en esta memoria es estudiar la realidad que expresan los textos noticiarios sobre los emprendimientos, queremos absorber esto a través de las limitaciones y observar cómo; la comunicación realizada por estos medios de comunicación podría inducir una interpretación diferente a la obviada por el lenguaje que conocemos.

2.5. Aplicaciones con Word Embedding

Si bien, a través de *Word2vec* se pueden obtener representaciones vectoriales de las palabras que componen un corpus, la información que expresan cada uno de estos vectores se encuentra codificada y no es entendible por el humano. Es por esto, que una vez obtenida la información vectorial forma parte crucial trabajar con los vectores, realizando diferentes aplicaciones para darles una interpretabilidad coherente al entendimiento humano.

En la correspondiente sección se darán a conocer de forma breve algunas aplicaciones con *word embeddings*, que resultan beneficiosas para la obtención de información descriptiva para este trabajo.

2.5.1. Similitud de Word-Embeddings

Bajo el supuesto que la generación de buenos *word embeddings* generan representaciones vectoriales similares, en palabras que fueron utilizadas en un mismo contexto. Desearemos calcular un valor escalar basado en dos vectores, para reflejar que tan similares o a que distancia se encuentran los vectores.

Para las definiciones que serán entregadas, supondremos dos vectores generados por un modelo Word2Vec dados por $u \in \mathbb{R}^d$ y $v \in \mathbb{R}^d$, donde d es la dimensión de los vectores.

2.5.2. Distancia Euclidiana

Una métrica para obtener la distancia a que se encuentran dos vectores es la distancia euclidiana, esta técnica queda definida por:

$$Dist_{euclideana}(u, v) = \sqrt{\sum_{i=1}^{d}(u_i - v_i)^2} = \|u - v\|_2 \tag{2.15}$$

De la ecuación, valores pequeños (cercanos a cero) señalan una mayor similitud entre los vectores y largas distancias una menor similitud.

2.5.3. Similitud de Coseno

Una técnica útil para obtener similitud entre las palabras es la aplicación de la similitud de coseno.

Debido a la naturaleza vectorial de las características obtenidas por *word2vec*, la similitud de coseno señala que dos palabras son similares entre sí mientras más cercano a 1 se encuentre el valor dado por la ecuación 2.16.

$$Sim_{cos}(u, v) = \frac{u \cdot v}{\|a\|_2 \cdot \|b\|_2} = \frac{\sum_i u_i \cdot v_i}{\sqrt{\sum_i (u_i)^2}\sqrt{\sum_i (v_i)^2}} \tag{2.16}$$

Dentro de los casos bordes que podemos obtener al aplicar la similitud de coseno entre dos vectores de palabras, tenemos:

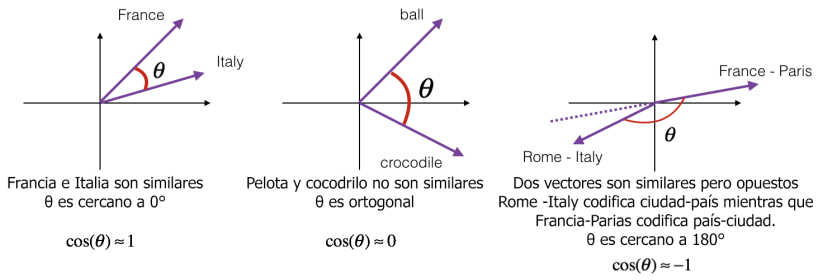


Figura 2.5: Similitud de coseno y casos borde al utilizar la ecuación.

2.5.4. Distancia Euclidiana v/s Similitud de Coseno

En la práctica posee mejores resultados aplicar la similitud de coseno para obtener la similitud entre dos palabras. El porqué de esto, es ilustrado en el siguiente ejemplo: supongamos que nuestro modelo ha generado un vector $a = 3b$, donde $a = [2, 4, -1]$ y $b = [6, 12, -3]$. Como podemos ver, los vectores poseen la misma dirección, pero diferente magnitud. Si aplicamos la similitud de coseno entre estos dos vectores, obtendremos que la similitud es igual a 1 y esto será producto que poseen las mismas proporciones relativas en cada una de sus componentes.

Luego, si calculamos la distancia euclidiana entre ambos vectores, obtendremos que es aproximadamente $9,14$. Bajo el valor obtenido, será fácil encontrar otro valor similar de algún vector c apuntando hacia otra dirección, donde a pesar de tener diferente semántica (dadas por la dirección), la función arrojará la misma similitud. Con esto, podemos notar que la distancia euclidiana se vuelve una métrica ambigua para visualizar la similitud entre dos palabras.

2.5.5. Clasificación

Como se revisó en la sección 2.1.1, existen aprendizajes supervisados que necesitan de datos etiquetados para el entrenamiento de un modelo. La clasificación es un problema que pertenece a este tipo de aprendizajes, donde la tarea principal es ajustar un modelo a las diferentes clases que posee un conjunto de datos para predecir la clase de futuras entradas.

Producto que nuestro objetivo es analizar la percepción sentimental asociada a las palabras del corpus, una buena aproximación es la aplicación de lexicones emocionales para clasificar la percepción temporal que se tiene de los *embeddings*. En específico, se reconoce como lexicón de emociones a un conjunto de palabras que son etiquetadas con el sentimiento percibido por un conjunto de etiquetadores. A continuación, definimos algunos clasificadores clásicos para realizar la tarea comentada:

2.5.5.1. Regresión Logística

La regresión logística es una técnica estadística supervisada que comúnmente se utiliza para estimar la probabilidad de pertenecía a una clase. La forma de clasificación es binaria, por lo que, si la probabilidad estimada es superior a 50 %, el modelo predice que la instancia pertenece a la clase positiva etiquetada como 1, de lo contrario la instancia pertenece a la clase negativa etiquetada con un 0.

Estimación de probabilidades

La regresión logística realiza estimaciones probabilísticas de forma muy similar a una regresión lineal, para esto el modelo calcula una suma pondera del vector de características de entrada (x) más un *bias* (b), pero en vez de entregar directamente la probabilidad aplica umbrales con los que determina la pertenecías a una clase u otra. La probabilidad estimada por el modelo queda por la ecuación 2.17.

$$\hat{p} = h_\theta(x) = \sigma(x^T\theta) \tag{2.17}$$

De 2.17 σ representa una función sigmoidal que posee un recorrido que va de 0 a 1 (notar que esta será la probabilidad de pertenencia a cada una de las clases), la ecuación que define a esta función viene dada por:

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{2.18}$$

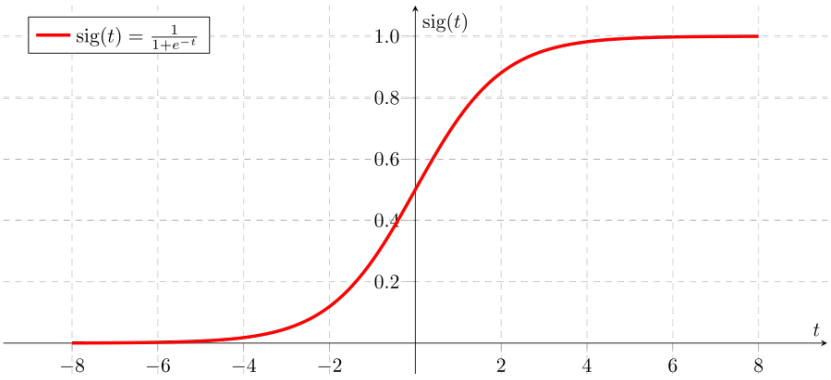


Figura 2.6: Representación gráfica de la función sigmoide.

Obtenidas las probabilidades, el modelo clasificará a una clase como positiva (1) o negativa (0) utilizando los umbrales señalados en 2.19.

$$\hat{y} = \begin{cases} 0 & \text{si } \hat{p} < 0.5 \\ 1 & \text{si } \hat{p} \geq 0.5 \end{cases} \tag{2.19}$$

Entrenamiento del Modelo

El objetivo del entrenamiento es buscar el conjunto de parámetros del vector θ con el que se estimen con una alta probabilidad las instancias positivas y con una baja probabilidad a las instancias negativas que entran al modelo.

La forma de realizar el entrenamiento es utilizando la función de costos *cross-entropy* vista para el entrenamiento de una red neuronal. La razón de utilizar esta función de costos es debido a que $-log(t)$ obtiene valores grandes cuando t es cercano a 0, castigando con altos valores la función de costos cuando se estima una probabilidad 0 para una clase positiva. De manera analoga, $-log(t)$ tendrá valores cercanos a 0 cuando t tiene valores cercanos a 1, de esta forma se obtendrán bajos valores en la función de costos debido a que una clase positiva estará asociada a altas probabilidades.

$$c(\theta) = \begin{cases} -log(\hat{p}) & \text{si } y = 1 \\ -log(1 - \hat{p}) & \text{si } y = 0 \end{cases} \tag{2.20}$$

La función de costos sobre todo el conjunto de entrenamiento es el promedio de los costos sobre todas las instancias de entrenamiento, escribiéndose como 2.21.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}log(\hat{p}^{(i)}) + (1-y^{(i)})log(1-\hat{p}^{(i)})] \tag{2.21}$$

Luego, al ser una función convexa la función de costos un algoritmo de optimización como el descenso del gradiente nos garantizaran encontrar un mínimo global con el que se minimicen las perdidas de la función de costos.

$$\frac{\partial}{\partial\theta_j}J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(\sigma(\theta^T x^{(i)}) - y^{(i)})x_j^{(i)} \tag{2.22}$$

2.5.5.2. Support Vector Machine

Un *Support Vector Machine* (SVM) es un clasificador discriminativo definido formalmente por un hiperplano de separación. En base a datos de entrenamiento etiquetados (aprendizaje supervisado), el algoritmo produce un hiperplano óptimo que categoriza los nuevos ejemplos, esto en base a la maximización de la distancia entre los diferentes vectores de soporte que conforman el conjunto de entrenamiento. Dicho de otras palabras, en un espacio de dos dimensiones y dos clases, el algoritmo generará un hiperplano (una línea en el caso de dos dimensiones) que dividirá el espacio de dos dimensiones diferenciando cada una de las clases.

2.5.5.3. Random Forest

Random Forest es un algoritmo de aprendizaje supervisado, el que se compone de diferentes árboles de decisión, cada uno de ellos con los mismos nodos, pero utilizando datos diferentes y generando diferentes asociaciones en sus nodos. Este algoritmo fusiona las decisiones de múltiples árboles de decisión, para encontrar una respuesta y predecir nuevos datos.

2.5.6. Métrica de Sesgo

Uno de los trabajos más relevante relacionados a la búsqueda de sesgos es realizado por por Garg et. al en [15]. En este trabajo se logra demostrar la diferencia de sesgo relacionado a estereotipos y actitudes hacia las mujeres y grupos étnicos a través de *word embeddings*. La métrica propuesta para realizar este estudio es la distancia normativa relativa (de sus siglas en ingles RND), la que se define como:

$$relative\,norm\,distance = \sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2 \tag{2.23}$$

De la ecuación se desprende la comparación de los *embeddings* de dos conjuntos de palabras de grupos sociales ($v_{1,2}$) frente a un único conjunto de palabras neutras v_m. Donde la diferencia respecto al conjunto de palabras neutra señala que tan sesgado se encuentran los grupos de estudios.

2.5.7. Reducción de Dimensionalidad

Debido a que al utilizar *Word2Vec* se obtienen vectores de una larga dimensionalidad, estos no son posibles de plasmar gráficamente en tres o dos dimensiones para visualizar las similitudes entre las palabras.

La reducción de dimensionalidad es un proceso en donde un conjunto de datos, quienes tienen una dimensión d, ven reducida su dimensión a una dimensión s, donde $s << d$, luego de aplicar un conjunto de operaciones matemáticas.

Es por esta razón que es necesario aplicar una reducción de dimensionalidad sobre los resultados obtenidos, de esta forma se podrán visualizar tanto en 3d o 2d los *embeddings*, haciendo posible una interpretación más rápida para el lector de los resultados.

2.5.7.1. UMAP: *Uniform Manifold Approximation and Projection*

UMAP es una técnica de aprendizaje manifold para la reducción de dimensionalidad propuesto por por McInnes et al. [25]. *UMAP* es construido basado en la geometría de Riemannian y topologías algebraicas. Obteniendo como resultado un practico algoritmo escalable que es aplicable a datos reales.

En su funcionamiento, *UMAP*, construye una representación de un grafo de alta dimensionalidad de los datos para optimizar y obtener un grafo de baja dimensionalidad con estructuras lo más similar posible a los datos originales. Para construir el grafo de alta dimensionalidad, *UMAP* comienza construyendo una estructura llamada "*Fuzzy simplicial complex*". La cual es una representación de un grafo ponderado, donde los bordes ponderados representan la likelihood que los dos puntos estén conectados. Para determinar la falta de conexión, UMAP extiende un radio exterior desde cada uno de los puntos, conectando los puntos cuyos radios exteriores se solapan.

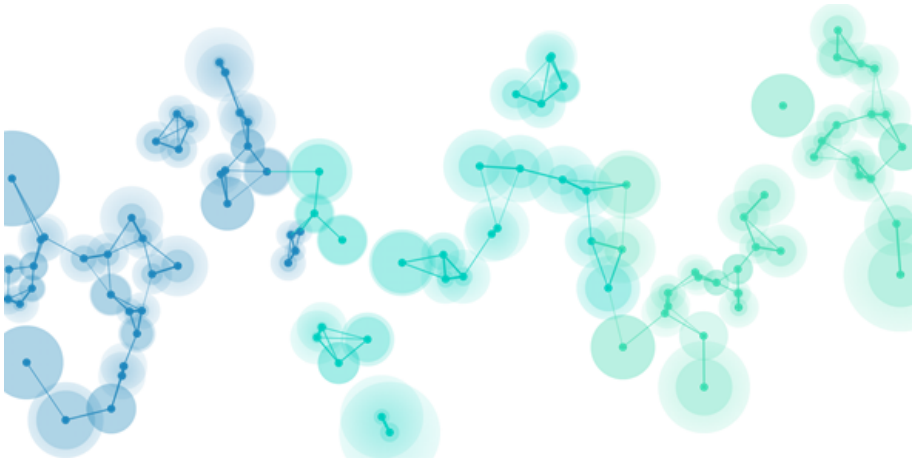


Figura 2.7: Representación de los grafos de alta dimensionalidad y los radios exteriores utilizados por el algoritmo.

Una vez construidas las representaciones de grafo en alta dimensionalidad, UMAP optimiza el diseño de un grafo de baja dimensionalidad, utilizando *cross-entropy* para ser lo más similar posible a la topología de alta dimensionalidad.

Alguna de las ventajas que posee UMAP sobre otros algoritmos de reducción de dimensionalidad es el fundamento matemático propuesto, lo que permite una mejor conservación de la estructura global de los datos y hace que tenga tiempos de ejecución más rápidos.

Capítulo 3

Metodología

Bajo la problemática de encontrar información clave para ayudar a la comprensión del desarrollo de la percepción asociada a conceptos relacionados a emprendimiento; a continuación, se presenta la metodología a seguir para desarrollar las diferentes problemáticas expuestas en el capítulo 1. Con esto, el capítulo comienza con la presentación formal del conjunto de datos utilizado en este proyecto, para luego dar paso a una exposición resumida de todas las actividades a realizar durante este trabajo de investigación, finalizando con una explicación más profunda de cada una de las propuestas metodológicas señaladas anteriormente.

3.1. Base de Datos

El conjunto de datos a utilizar en este trabajo es obtenido a través de *web scrapping*, la cual es una técnica para extraer texto desde páginas web. Del proceso, se obtienen y filtran noticias que tienen incluidas en su cuerpo palabras relacionadas a emprendimiento. Como resultado de la recopilación de datos, se obtienen múltiples *dataframes* para cada uno de los periódicos de los países: Estados Unidos, Reino Unido, Irlanda, Canadá, Australia y Nueva Zelanda. De cada una de las estructuras resaltan las siguientes columnas:

- TITLE: Titulo de la noticia publicada.
- BODY: Cuerpo de la noticia, en él podemos encontrar el desarrollo de la noticia.
- DATE: Fecha de publicación original de la noticia.
- LOAD DATE: Fecha en que se cargó la noticia a la base de datos.

Donde en total se tiene el siguiente número de noticias por países:

Tabla 3.1: Número de artículos noticiarios por países.

País	Artículos
Australia	104.603
United Kingdom	103.972
United States	75.670
Canada	51.780
Ireland	17.816
New Zealand	16.523

Es importante señalar que el país se identifica por los acrónimos de cada uno de los países estudiados, situándose este en cada uno de los archivos donde se han almacenado los corpus noticiarios obtenidos desde el *web scrapping* para cada periódico.

3.2. Resumen de Metodología Propuesta

La metodología a seguir en este trabajo es expuesta en la figura 3.1. Dentro de la metodología propuesta se pueden visualizar cuatro grandes pasos que son resumidos a continuación:

- Preprocesamiento y desambiguación de palabras: Se realiza un tratamiento en los datos para filtrar ruido del dataset y obtener el formato correspondiente al entrenamiento del modelo *Word2Vec*. Junto a esto se propone la desambiguación de palabras de interés, con el fin de obtener una referencia temporal y del país en el que fueron publicadas, para así obtener vectores representativos para cada país.
- Entrenamiento del modelo *Word2Vec*: Se obtiene una representación vectorial de los tokens correspondientes a los corpus noticiarios.
- Obtención de información desde los *Word Embeddings*: En base a los *Word Embeddings* obtenidos en el paso anterior, se propone la realización de cuatro tareas con el fin de visualizar relaciones tanto temporales como atemporales referidas a emprendimiento.
- Análisis de resultados: Son analizados los resultados, contrastando la información obtenida desde las tareas realizadas en los pasos anteriores.

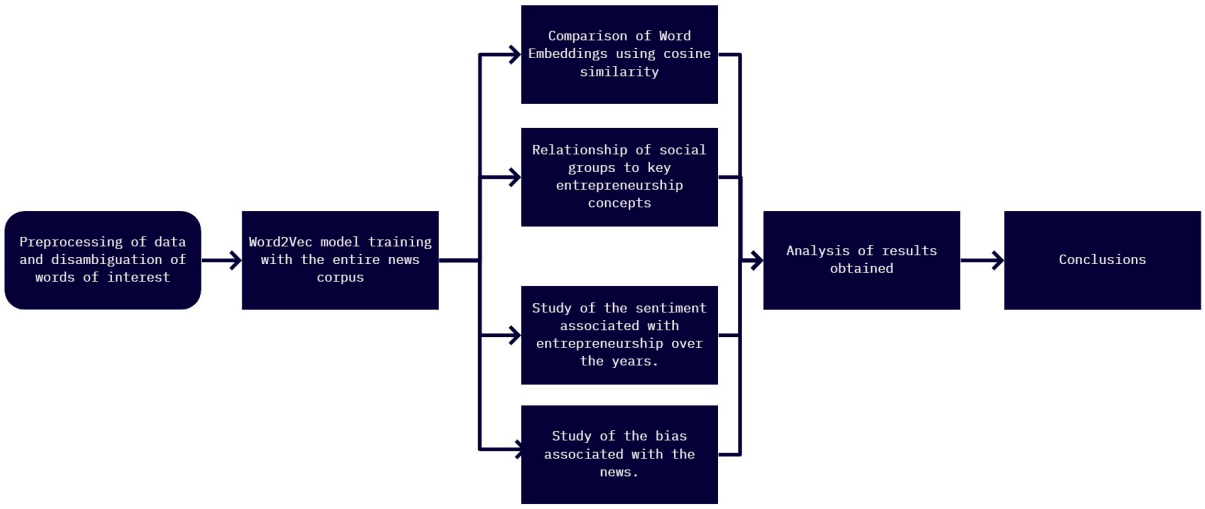


Figura 3.1: Esquema de trabajo propuesto. Fuente: Elaboración Propia.

Para cada una de las actividades propuestas se buscará obtener visiones que nos permitan identificar y comprender el significado semántico de emprendimiento para un conjunto de países de habla inglesa. Por lo señalado, el trabajo busca tener un impacto dentro del área de economía y *Data Science*, en un tema que hasta la fecha no ha sido estudiado con aplicaciones de *Machine Learning*.

3.3. Desarrollo de las Metodologías Propuesta

3.3.1. Pre-procesamiento de Datos

La primera etapa de este proyecto consiste en el pre-procesamiento de los datos. Esta etapa tiene como objetivo modificar y limpiar los datos para el entrenamiento del algoritmo *Word2vec*, obteniendo con esto una desambiguación de un conjunto de palabras de interés.

3.3.1.1. Exploración de los Datos

El primer paso es explorar de forma manual el conjunto de noticias, del que se visualizan un bajo número de datos faltantes en las variables BODY y DATE en alguno de los corpus a utilizar. En conocimiento de los datos faltantes, se propone eliminar las filas que poseen datos nulos para las variables BODY, ya que al no poseer un texto de noticia no entrega información relevante para el estudio. En segundo lugar, para los casos con datos faltantes en las columnas DATE, en caso de tener valores nulos se rellena este campo por el año de carga (LOAD DATE) para cada una de las noticias, esto producto que al estudiar los valores de las celdas DATE y LOAD DATE se presentan las mismas fechas. El objetivo de rellenar estas celdas con valores nulos es tener registro de la publicación de las noticias, de tal forma de realizar una futura desambiguación en base a las fechas y nacionalidad de las noticias.

3.3.1.2. Tokenización de los Documentos

Paso siguiente en esta etapa, es la tokenización de cada uno de los documentos que conforman el corpus. Esta tarea consiste en la transformación de cada uno de los textos en vectores de palabras, quienes están conformadas por las palabras que aparecen en cada uno de los documentos noticiarios. La tokenización, forma parte crucial para el entrenamiento del modelo de *Word2Vec*, ya que estos modelos poseen como entradas vectores de palabras que son recorridos para predecir las palabras colindantes. Dicho esto, a continuación, se ilustra un ejemplo de la tarea de tokenización para un documento compuesta por la oración "*have an entrepreneurship to prove*", en donde se obtienen los tokens:

Documento_tokenizado = [*"have", "an", "entrepreneurship", "to", "prove"*]

Cabe señalar que durante la tokenización de los documentos son eliminados los signos de puntuación y mayúsculas de los textos. Esto debido a que no le entregan mayor contenido semántico al análisis de noticias, y al momento de realizar un análisis sobre los *Word Embeddings* obtenidos podría provocar algún nivel de ruido en los resultados obtenidos.

3.3.1.3. Normalización de las Palabras

Según [38], la normalización de tokens es el proceso de canonización de tokens para que las coincidencias se produzcan a pesar de las diferencias superficiales en las secuencias de caracteres de los tokens. Una simple forma de generar una normalización de tokens es generando equivalencias entre un grupo de palabras y otro.

Dentro de este trabajo serán aplicadas dos tipos de normalización, la primera es enfocada en la **normalización de los tokens a un inglés estadounidense**, mientras que la segunda es enfocada en la **lematización** de los tokens.

Con relación a la normalización de los tokens a un inglés estadounidense. El objetivo de esta tarea es crear un vocabulario homogéneo entre los diferentes países que conforman el corpus, haciendo más uniformes la semántica de las palabras que conforman los documentos. Para realizar esta tarea, se estudia la frecuencia en que aparecen las diferentes palabras que componen al corpus, visualizando y buscando las palabras que difieren del inglés estadounidense. Hecho esto, se construye un diccionario que relaciona las palabras de un inglés estadounidense y no- estadounidense, luego, el diccionario es aplicado sobre el conjunto de tokens para normalizar las coincidencias.

Figura 3.2: Del ejemplo se tiene a la derecha un conjunto de palabras en un ingles británico, quienes al aplicarle el diccionario construido, son transformadas a un ingles estadounidense.

Para el caso de la lematización, este corresponde al proceso de transformar las diferentes formas flexionadas de una palabra a su forma base, conocida formalmente como lema, para que puedan ser analizadas como un único elemento. Este caso al igual que el anterior, es aplicado sobre todos los tokens y se de él se espera generar un vocabulario más similar entre las variaciones que pueden presentar las palabras.

Figura 3.3: Ejemplo del proceso de lematización. Al lado izquierdo se tiene un conjunto de palabras flexionadas, las cuales al ser procesadas con un algoritmo de lematización se obtiene la palabra base a la derecha.

3.3.1.4. Desambiguación de las palabras

Se propone la desambiguación de palabras de interés relacionadas a *entrepreneurship* , utilizando el año en que aparecen en el corpus y al país del corpus al que pertenecen. El objetivo de esta tarea es obtener palabras que nos permita obtener información relevante desde los *word embeddings*, donde podamos reconocer el año y concepto en los determinados años de búsqueda. A continuación, se expone la lista de palabras de interés a utilizar en este trabajo:

Palabras_de_interes = [“entrepreneurship”, “founders”, “entrepreneurialsm”, “entrepreneurial”, “entrepreneur”]

En base a las palabras señaladas será necesario recorrer y buscar en cada uno de los documentos la aparición de estos conceptos. En el caso de encontrar alguna de las palabras de interés, estas son modificadas anexando el año de la publicación de la noticia y el país en donde fue publicada al token encontrado. De esta forma se espera a través de la desambiguación obtener el siguiente formato en las palabras de interés:

XXXX_PAIS_PALABRA

Donde XXXX es el año en el que fue publicada la noticia, PAIS el país de donde procede el token noticiario y PALABRA alguna de las palabras de interés señaladas anteriormente.

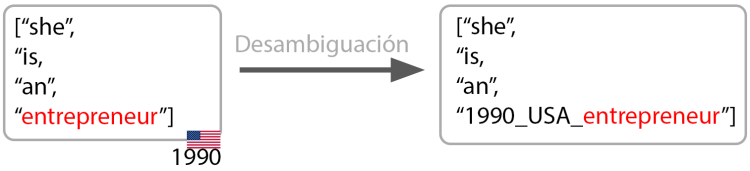


Figura 3.4: Ejemplo del proceso de desambiguación propuesto. A la izquierda se observa un documento estadounidense del año 1990 (tokenizado), mientras que a la izquierda se ve el resultado de aplicar la tokenización.

3.3.1.5. Unión de los Corpus

Finalizados cada uno de los procesos señalados anteriormente, son unificados cada uno de los tokens procesados en una gran lista, para luego, alimentar al modelo de *Word2Vec* y generar un espacio de vectores comparables entre sí.

3.3.2. Obtención de Word Embeddings

Una vez explorado y preprocesados los datos noticiarios, en esta etapa se busca obtener las representaciones vectoriales del corpus de noticias. Como se ha señalado, la obtención de características en texto a utilizar lleva por nombre word embeddings, por lo que se espera como resultados la obtención de representaciones vectoriales que logren capturar información semántica desde los textos noticiarios. De esta forma, los vectores obtenidos nos permiten, a través de técnicas complementarias, extraer información relevante para este estudio.

Para abordar este problema, se decide generar un modelo de *Word2Vec* a través de la librería *Gensim* de *Python*. Debido a que dentro de los objetivos del proyecto se encuentra visualizar las diferentes interpretaciones de emprendimiento para cada uno de los países y años que posee el corpus; se genera un modelo *Word2Vec* utilizando los corpus de noticias para todos los países. De esta forma, se obtendrán representaciones vectoriales comparables entre todos los países que conforman los datos.

Cabe señalar, que haber producido modelos *Word2Vec* para cada uno de los países de forma independiente, hubiese generado representaciones vectoriales incomparable unas con

otras. Esto se debe a que los vectores generados por la red neuronal dependen directamente de la información entregada en el proceso de entrenamiento, por lo que al trabajar con una diferente fuente de datos generaran un espacio vectorial con un determinado número de vocabulario, haciendo incompatibles la comparación de los vectores entre modelos al poseer una diferente significancia.

Con lo anterior, los parámetros escogidos para entrenar al modelo *Word2Vec* son los parámetros por defecto de la librería *gensim*. Esta elección se debe a que no se tiene registro, ni fundamentos de una configuración que presente un mejor desempeño para abstraer información en textos relacionados a emprendimientos. Es relevante señalar que la construcción del modelo considera una heurística genérica, donde se tienen como pasos la generación de bígamas, vocabulario y entrenamiento del modelo.

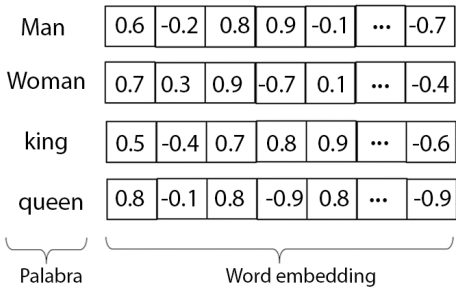


Figura 3.5: Ejemplo del tipo de representaciones vectoriales generadas por un modelo Word2vec.

3.3.3. Comparación de los Word Embeddings

Como primer paso en la interpretación de los *word embeddings*, se plantea observar los términos más cercanos que posee la palabra emprendimiento para los diferentes países. El objetivo de esta tarea es visualizar de forma cualitativa el comportamiento que posee la palabra emprendimiento, desambiguada, con diferentes términos del corpus. Así, se buscará observar diferencias en los conceptos que se relacionan a emprendimiento tanto para los países de estudio, como a lo largo de los años. Para lograr los objetivos señalados, esta parte es dividía en tres: la primera está enfocada en visualizar diferencias en la semántica de la palabra emprendimiento entre los diferentes países, la segunda tiene como objetivo visualizar la evolución de *entrepreneurship* a través del tiempo para cada uno de los países, mientras que la tercera conociste en construir una matriz de similitud entre para la palabra *entrepreneurship* y palabras anclas a través de los años para cada país.

3.3.3.1. Similitud de Semántica entre Países

Para lograr visualizar la similitud de conceptos relacionados a emprendimientos entre países, se propone la obtención de los 30 términos más similares a la palabra desambiguada *entrepreneurship* en el año más reciente del conjunto de datos por país (año 2019); la similitud señalada entre las palabras es obtenida a través de la similitud de coseno entre las representaciones vectoriales de las palabras.

Producto que es esperable una alta relación entre la palabra de búsqueda y las otras pa-

labras desambiguadas, se debe realizar un filtro que elimine las palabras desambiguadas de la consulta. Así, el resultado esperado son los conceptos no desambiguados más similares a la palabra de consulta, junto a los vectores asociados a cada una de estas palabras.

Finalmente, para lograr la visualización gráfica de los términos obtenidos, se aplica el reductor de dimensionalidad *UMAP* sobre los *embeddings* para obtener representaciones vectoriales en dos dimensiones. Dentro de los resultados esperados se espera visualizar diferencias entre las relaciones que se obtiene para la palabra *entrepreneurship* y las más similares para cada uno de los países. Asociándose diferentes conceptos para cada uno de los países.

3.3.3.2. Evolución de la Semántica para cada País

Esta tarea se desarrolla de forma similar a la mencionada anteriormente, pero a diferencia de consultar por solo el año más reciente del conjunto de datos, se hace una búsqueda de las 5 palabras más similares para cada uno de los años que compone a cada uno de los corpus por país. El objetivo de esto es visualizar una potencial evolución en la semántica del emprendimiento para cada uno de los países, esperando observar los conceptos con mayor relación que se tienen por años.

3.3.3.3. Matriz de Evolución Semántica

Con el objetivo de ahondar más en lo realizado en la subsección anterior, se propone la construcción de una matriz de relación a conceptos anclas a lo largo de los años. De esta forma, se podrá visualizar la variación en similitud de la palabra emprendimiento a lo largo de los años con conceptos relevantes en una sociedad. De esta forma, se construye una lista de palabras anclas que se consideran relevantes en temas de desarrollo social y emprendimiento, dando creación al siguiente conjunto de palabras:

palabras_anclas= [*"developer", "venture", "profit", "investor", "accelerator", "incubator", "networking", "copyright", "patent", "trademark", "risks", "launch", "pitch", "partner", "capital", "sustainability", "internet", "small_business", "startup", "science", "selfemployed", "energy"*]

Luego, a través de la aplicación de similitud de coseno, el conjunto de palabras es comparada con la palabra emprendimiento a través de los años, para visualizar como han evolucionada a lo largo de los años el concepto emprendimiento con estas palabras anclas.

3.3.3.4. Matriz de Evolución Semántica

Este ejercicio tiene como objetivo visualizar la variación que experimenta la palabra emprendimiento sobre su primera aparición a lo largo del registro. Al comparar respecto al primer registro de emprendimiento, se podrá visualizar la variación que presenta este concepto a lo largo de los años.

Para cada uno de los países son utilizadas las palabras desambiguadas que representan a emprendimiento. La similitud de coseno es utilizada entre el primer registro y los siguientes años de registro que se tienen en cada país. De este modo, cada comparación tendrá como referencia el primer registro que se tiene de emprendimiento.

3.3.4. Relación de Grupos Sociales a Conceptos Claves de Emprendimiento

Debido a que el conjunto de datos está relacionado principalmente con temas de emprendimiento, se propone estudiar a través de la similitud de coseno la relación que poseen palabras relacionadas a ciertos grupos sociales, respecto a conceptos claves en un emprendimiento. Esto tiene como objetivo identificar a través de la similitud, potenciales diferencias sociales para grupos antagónicos, expresando por medio de una brecha, que tan dispares son unos de otros respecto a los conceptos claves.

Para realizar la tarea señalada, se propone la utilización de conjuntos de palabras que identifican a grupos sociales. Dentro de los grupos se estudiarán aquellos que diferencien:

- Genero: Grupos conformados por los géneros masculinos y femeninos.
- Religión: Este grupo está conformado por términos que representan a algunas de las religiones con más adherentes en el mundo, entre ellas se encuentran: católicos, Judíos y Musulmanes.
- Ciudadanía: Grupo que conformado por conceptos relacionados a ciudadanos e inmigrantes.
- Razas: Para este grupo podemos encontrar las principales razas que conforman los países de estudio, entre ellos podemos encontrar: personas blancas, negras, latinos y asiáticos.

Un ejemplo de esto, son *"male"* y *"female"*, quienes representan grupos sociales de género. Donde, para cada uno de los grupos sociales, se produce una serie de palabras relacionados a estos, como por el ejemplo para el caso de *"male"* se esperaría encontrar palabras como: *"he"*, *"him"*, *"man"*, *"boy"*, etc. Para mayor información de los diferentes grupos sociales definidos para este proyecto y como estos están conformados, revisar en el apéndice 5 de este documento.

Luego, se debe definir el conjunto de palabras claves para un emprendimiento, para esto son escogidos términos que socialmente son relacionados a emprendimiento y que aparecen un número consistente de veces en el corpus, este conjunto de palabras viene dado por:

query_words= [*"developer", "venture", "profit", "investor", "accelerator", "incubator", "networking", "copyright", "patent", "trademark", "risks", "launch", "pitch", "partner", "outsourcing", "capital", "sustainability", "internet", "smallbusiness", "spinoff", "startup", "familybusiness", "selfemployed"*]

En base a lo definido, se utiliza la similitud de coseno entre los grupos antagonistas respecto a cada uno de los conceptos claves. Cabe señalar que, para obtener una similitud general para cada uno de los grupos, es obtenida la similitud de coseno para cada termino que conforman a los grupos sociales respecto a las palabras claves. Del resultado obtenido, se calcula la media del conjunto de similitudes para cada grupo, obteniendo de esta forma una representación general para cada grupo respecto a cada concepto clave.

3.3.5. Predicción del Sentimiento Asociado a Emprendimiento

Una de las problemáticas a resolver en esta investigación es observar la variación sentimental asociada a términos relacionados a emprendimiento. Para esto, se propone entrenar un modelo de regresión logística utilizando el lexicón de Bing Liu y los *word embeddings* obtenidos. De esta forma, se espera obtener un porcentaje que refleje la positividad que poseen las palabras desambiguadas a lo largo de los años.

De acuerdo con los porcentajes de positividad, se espera obtener valores bajos que permitan visualizar a través del texto la existencia de crisis económicas en ciertos periodos en los países. De esta forma, se podrá verificar que a través de los diarios se transfieren sentimientos que podrían influenciar a las masas.

Los resultados serán evaluados de forma cualitativa y dentro de los riesgos esta encontrar países en donde no se logren observar tendencias significativas por la diferencia de datos que se posee para cada uno de los países. En caso de visualizar estos comportamientos, serán estudiados los casos más relevantes.

3.3.6. Medición de Sesgo a Través de RND

Como ultima tarea se plantea obtener la variación de sesgo a través de los años utilizando la métrica RND señalada en el capítulo 2. El objetivo será visualizar la variación de sesgo que se tiene para los grupos sociales definidos en 3.3.1 a través de los años. Con esto, se espera demostrar la evolución que han experimentado los grupos sociales, a lo largo de los años con relación a la palabra emprendimiento para cada uno de los países. A priori, se espera que los resultados sean expuestos con un gráfico, el cual permita observar la variación del sesgo a través de los años.

Para calcular el sesgo a través de los años, se utilizan las listas antagonistas ya definidas de los grupos sociales. Con ellas, se utiliza como palabra de consulta las palabras desambiguadas para cada uno de los países, donde aplicando el algoritmo RND se obtiene el sesgo asociado para cada año.

Cabe señalar que para realizar esta tarea se hará uso de la librería WEFE de *Python*, la cual ofrece una forma amigable para obtener el sesgo entre diferentes grupos sociales respecto a una palabra de consulta.

Capítulo 4

Resultados y Análisis

En el presente capitulo son expuestos los resultados obtenidos al aplicar la metodología señalada en el capítulo 3 en base al corpus noticiario de habla inglesa. Los resultados expuestos a continuación, son comparados de forma cualitativa y cuantitativa frente a cada uno de los países, visualizando la evolución semántica y de sesgo que han experimentado cada uno de los países a largo de los años y entre ellos.

4.1. Interpretabilidad de Emprendimiento para Diferentes Países

Tras la obtención de los *Word Embeddings* desde los diferentes corpus de noticias de habla inglesa, son obtenidos vectores de una dimensionalidad 100 para cada una de las palabras que conforman el vocabulario de entrenamiento. Al presentar una alta dimensionalidad cada una de las palabras entrenadas a través del *Word2Vec*, se aplica un algoritmo de reducción de dimensionalidad *UMAP* para obtener representaciones vectoriales en dos dimensiones; de esta forma se busca obtener representaciones interpretativas que nos permitan asociar y observar las potenciales cercanías que tienen las palabras del corpus a las palabras de emprendimiento desambiguadas para cada uno de los países.

Al obtener las 30 palabras más similares para cada uno de los países en el último año que se tienen registros completos de los datos (año 2019), se obtiene un gráfico donde se representa la cercanía que poseen las palabras del corpus a la palabra emprendimiento para cada uno de los países estudiados.

Del gráfico 4.1, es posible observar varios puntos de interés: el primero son las palabras que distribuyen en torno a Nueva Zelanda, esto nos dice que los emprendimientos están más asociados a temas educacionales en comparación a otros países. En segundo lugar, se observa un segundo conjunto de países como lo son: Estados Unidos, Reino Unido y Australia; quienes presentan una mayor cercanía a emprendimientos ligados al área económica o generación de riquezas. Finalmente, se tiene un tercer grupo de países, quienes, conformados por Irlanda y Canadá, presentan una cercanía al desarrollo de emprendimientos ligados a la cultura y desarrollo sustentable.

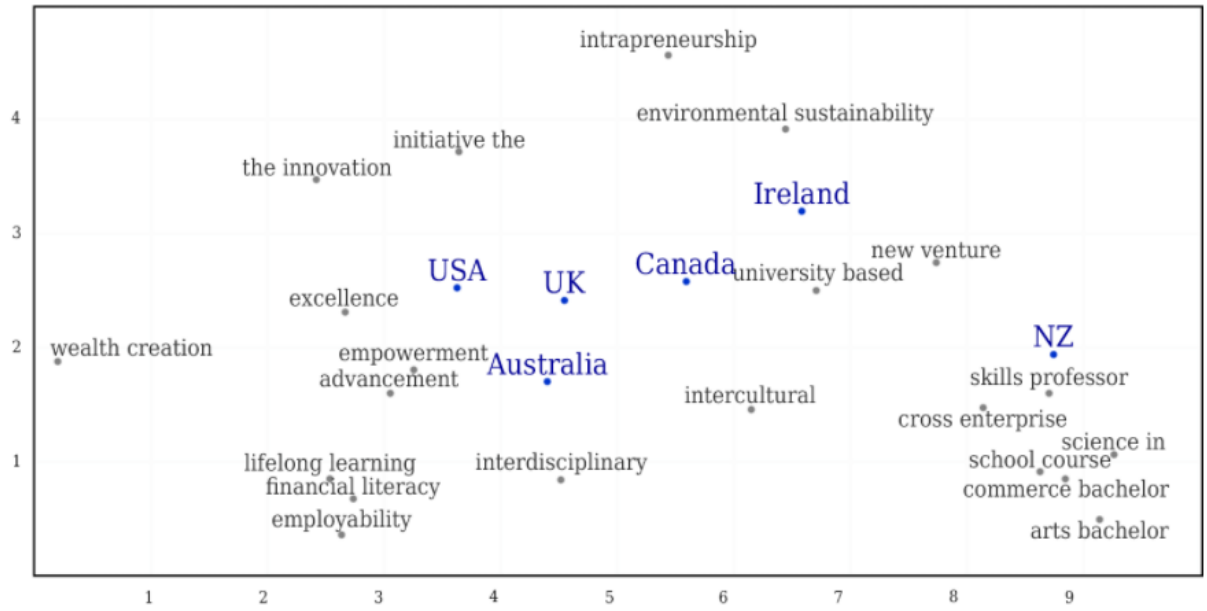


Figura 4.1: Similitud de *embeddings* para cada uno de los países del corpus.

4.2. Evolución de la Semántica para Cada uno de los Países

Para este ejercicio se proyectan las 5 palabras más similares respecto a conceptos relacionados a emprendimiento. Cabe señalar que parte de los resultados presentan un alto nivel de ruido, por lo que algunos de los resultados son filtrados escogiendo aquellos que poseen una interpretación human. Los resultados sin filtrar se encuentran en 5, donde son expuestas las tablas con las palabras más similares para cada uno de los países.

El objetivo de esta tarea es visualizar cómo ha evolucionado la semántica referida a emprendimientos a través de los años, buscando visualizar cambios en las interpretaciones u enfoques del desarrollo de emprendimientos para cada uno de los países. De esta forma, podrían visualizarse variaciones producto de temas sociales en boga para determinados años, o desviaciones respecto a los enfoques de emprendimientos que existen para cada uno de los países.

Los resultados son expuestos en una línea temporal creciente hacia la derecha para cada uno de los países. En la línea, cada uno de los puntos señala 5 palabras máximo por cada año de registro, dejando como cota superior el año 2019 debido a que hasta este año se tiene una totalidad de registros mensuales. La diferencia en los años de registro se debe netamente a la cantidad de datos presentes para cada uno de los países en sus respectivos corpus.

4.2.1. Evolución Semántica de Estados Unidos

A partir de 1980 es posible observar palabras con un alto nivel de ruido, en ellas se pueden encontrar palabras como: *finalistthe* y *ddf*, de las cuales no es posible obtener una interpretabilidad directa. Por otro lado, se presentan palabras referidas a localidad en estados unidos

como: *talawa* y *allagash*, quienes podrían referirse a lugares donde se podrían haber generado una alta cantidad de emprendimientos en el año estudiado. Finalmente se tiene la palabra *gemological*, quien hace referencia a la ciencia que estudia las piedras preciosas, lo que podría referirse a emprendimientos de joyería, lo cual no parece extraño en una época donde se vivía un boom de los diamantes en estados unidos.

Un comportamiento particular se da entre los años 1985 a 1995; donde los emprendimientos poseen una mayor similitud con la toma riesgos, innovación, individualismo y generación de riquezas. Con esto, un punto que se repite todos estos años es que se resalta la Autosuficiencia como posible característica de los emprendimientos. Con esto, este periodo señalaría una fase en donde los emprendimientos poseían como principal característica la creatividad, generación de riquezas y sobre todo la autosuficiencia, características que llevan a un desarrollo individualista de los emprendimientos.

En los años 2000 se comienza a observar un caso muy particular que es la aparición de *intrapreneurship*, los cual se refiere a la práctica de capacitar a los empleados de una determinada empresa para que innoven y actúen como propietarios de la empresa. O sea, comienzan a aparecer conceptos que dejan de lado la autosuficiencia e individualismo que deberían poseer los emprendimientos. Por otro lado, llama la atención la aparición de conceptos que motivan el aprendizaje referido a emprendimientos a través de conceptos como: *pedagogy* y *experimential learning*. Finalmente resalta la aparición de activismo político dentro de los emprendimientos, lo que podría estar asociado a la fuerte campaña que se estaba desarrollando con Bush y Al Gore en esos años en estados unidos.

Caso similar al anterior se ve en las similitudes obtenidas en el periodo del 2005 al 2019, en donde se observa una gran similitud de los emprendimientos con el aprendizaje al presentar similitudes con palabras como: *financial literacy*, *centennial collegue*, *experimental learning*, entre otras. A esto, se le suma nuevamente la aparición de conceptos como *intrapreneurship* o voluntariado, pudiendo asociar que los emprendimientos ya no son un concepto individualista o autosuficiente como lo eran en 1985 a 1995, sino que depende de aprendizaje y apoyo comunitario para la generación de riquezas.

Finalmente, se destaca la aparición en el 2015 de similitudes con tópicos sustentabilidad y medio ambiente, ya que hasta hace pocos años se comienza a tener una mayor conciencia en estos temas y esto confirma la aparición de emprendimientos sustentables con el medio ambiente.

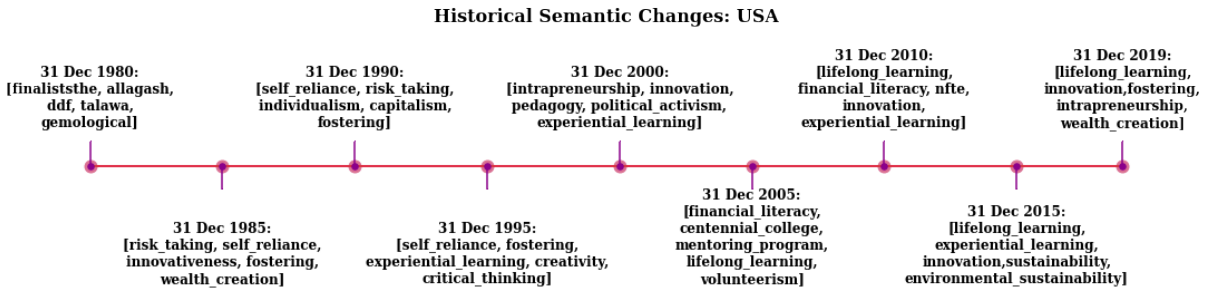


Figura 4.2: Evolución de la palabra *entrepreneurship* para el país Estados Unidos.

4.2.2. Evolución Semántica de Nueva Zelanda

Para el caso de Nueva Zelanda se logra observar un comportamiento más regular en las similitudes que se posee de registro. En ellas, es posible observar que para todo el registro se observa una gran cercanía hacia la educación, presentando todos los años de estudio tópicos relacionados a temas educativos como: *pedagogy*, *experimental learning*, *graduate certificate* o *excellence*. Esto nos dice que la educación es una característica esencial y que va de la mano para la generación de emprendimientos en ese país.

Por otro lado, otra de las características que destaca para este país es la similitud que posee con palabras sociales y del medio ambiente. Esto se observa desde el 2010 donde aparecen conceptos como: *civic engagement*, *environmental sustainability* y *organizational behavior*, palabras que nos señalan una preocupación tanto por el medio ambiente y por el desarrollo comunitario, lo cual se diferencia con los resultados obtenidos para estados unidos.

Finalmente, de los resultados se logra ver la lógica política que ha seguido el país de Nueva Zelanda y que lo hace tan característico a nivel mundial, donde destaca un pensamiento social demócrata y que se ve reflejado hasta en sus emprendimientos.

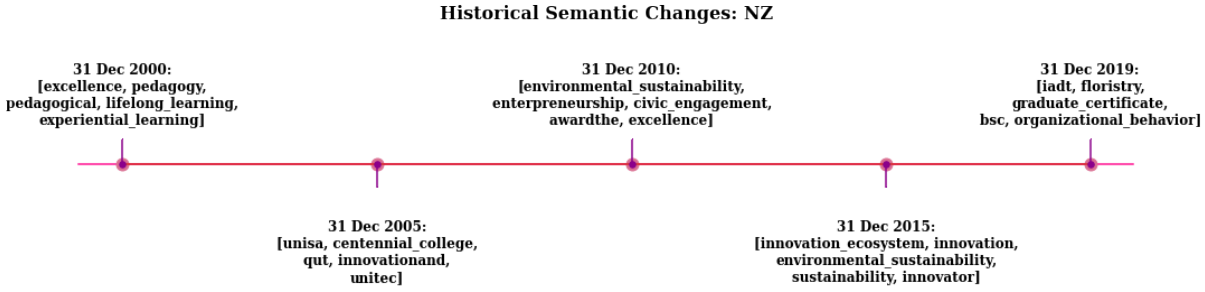


Figura 4.3: Evolución de la palabra “entrepreneurship” para el Nueva Zelanda.

4.2.3. Evolución Semántica de Reino Unido

De la línea de tiempo para Reino Unido es posible notar que para el periodo de 1985 a 1995 se tiene un gran número de palabras de las que no es posible obtener una interpre-

tación directa. Dentro de estas palabras podemos encontrar: *solecism*, *swadeshi*, *biohackers*, entre otras quienes no poseen una relación directa con el mundo del emprendimiento y no podríamos inferir como alguna característica del mundo del emprendimiento para este país. La razón de estos resultados podría deberse a un bajo número de documentos para los años señalados, donde no se logra abstraer una semántica más profunda de las palabras. Por otro lado, en los años señalados destaca la aparición de conceptos científicos. Si bien son interesantes destacar en una perspectiva de emprendimiento, debido al alto nivel de ruido que presentan estos años, no sería posible asociarlos a una característica de los emprendimientos para estos años.

Resultados interesantes se notan desde los años 2000 hacia adelante, donde es posible ver un comportamiento similar en el desarrollo de los años, enfocando los emprendimientos en la creatividad y desarrollo social. Dentro de los aspectos destacables, se observa una gran similitud hacia aspectos sociales como filantropía, voluntariado, sustentabilidad, medio ambiente y compromiso social; estos aspectos nos permiten inferir que los emprendimientos en Reino Unido han evolucionado con un enfoque en el desarrollo cívico y preocupados del medio ambiente desde muy temprano en los registros (2005).

Por lo señalado, para Reino Unido se logra ver un alto nivel de ruido en la información presentada en los primeros registros de evolución semántica. Estos no nos permiten inferir una muy clara perspectiva de emprendimiento en los primeros años de registro, pero esto cambia a partir de los 2000, donde es posible observar resultados más claros. A partir de los 2000 se observa un desarrollo similar a lo largo de los años, destacando la búsqueda de un desarrollo cívico en los emprendimientos.

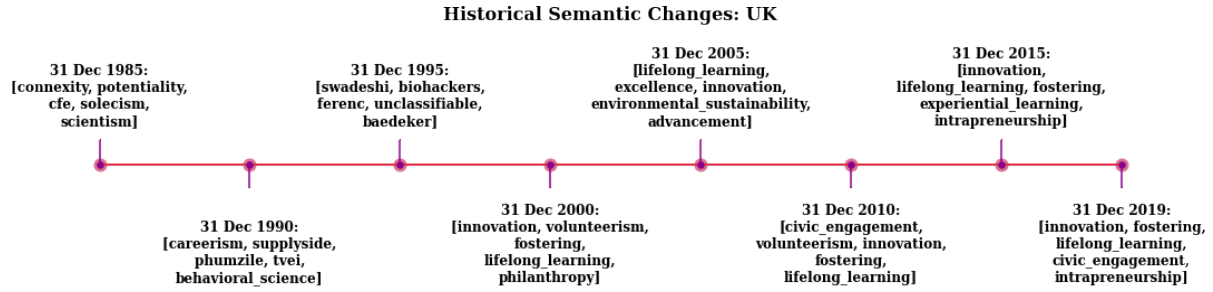


Figura 4.4: Evolución de la palabra *entrepreneurship* para Reino Unido.

4.2.4. Evolución Semántica de Irlanda

Tal como en otros países se comienza en los primeros dos años de registro con altos niveles de ruido, en donde es posible encontrar algunos casos de faltas ortográficas y gran cantidad de acrónimos para el país. La aparición de estos resultados podría deberse a una diferencia de formato en los textos más antiguos, lo que produciría errores al realizar el *web-scrapping*, obteniendo palabras incompletas o documentos con alto nivel de ruido. De igual forma, en el periodo de 1995 a 2000 se logran identificar la aparición de generación de riquezas y se mencionan un término de *start-up house*, lo que podría estar relacionado a la generación de emprendimientos.

Para los años siguientes (2005-2019), es posible observar un comportamiento similar, donde destaca el rol social de los emprendimientos, ya que podemos encontrar palabras como: *volunteerism*, *philanthropy*, *civic engagement*, entre otros. Siendo más detallados con los resultados obtenidos, podríamos observar que el comportamiento de irlanda sigue una evolución bastante similar a la vista en Reino Unido y generación de intrapreneurship en los últimos periodos de registro.

De los resultados, al igual que otros países se logra visualizar ruido en los primeros años de registro de evolución que se podrían deber a errores de formato de noticias antiguas que generan problemas al generar el *web-scraping*. Por otro lado, se logra visualizar a través de la similitud de los resultados con Reino Unido, que países geolocalizados cercanamente podrían presentan un desarrollo de emprendimiento bastante similar.

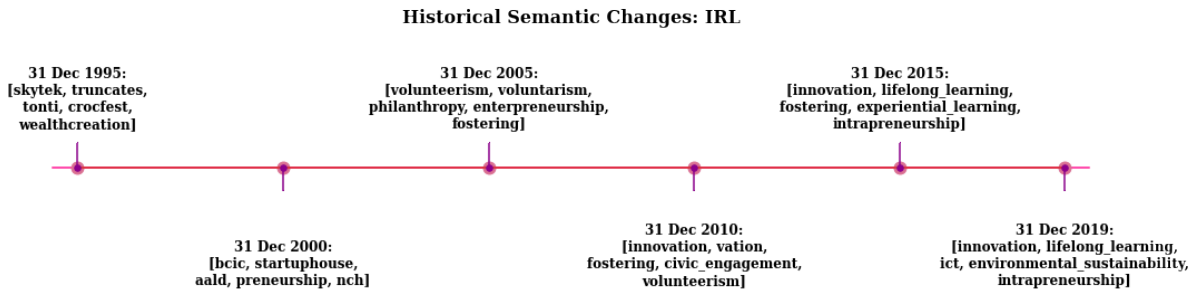


Figura 4.5: Evolución de la palabra *entrepreneurship* para Irlanda

4.2.5. Evolución Semántica de Canadá

Comenzando con un registro en el año 1985, las palabras más similares para Canadá presentan coherencia desde el principio. Esto se deberá a una mejor calidad de los datos extraídos desde sus sitios web de noticias, lo que nos permite obtener *embedding* coherentes desde un comienzo. Al analizar el primer año de registro para Canadá, podemos observar que este país no se separa muchos de estados unidos, quien en el mismo año presenta grandes similitudes con toma de riesgos.

Desde 1990 a 1995 se comienzan a observar una conducta más regular para Canadá, donde se observa una reiterativa similitud con tópicos educacionales como: *pedagody*, *experimential learning* y *lifelong learning*, lo que nos señala una importancia de aprendizaje al momento de realizar un emprendimiento. Por otro lado, destaca la aparición del término voluntariado durante estos años, lo que señala la aparición de emprendimientos con una motivación social.

Siguiendo con los años 2000 a 2005 resalta la aparición de la palabra *internationalization*. Esto es llamativo porque Canadá es el único país, de los estudiados, que posee una alta similitud con esta palabra, señalando que los emprendimientos buscan salir más allá del medio local en donde son creados. Por otro lado, durante estos años aparte de esta nueva palabra posee un comportamiento bastante similar a los otros países.

De 2010 al 2015, se mantiene la aparición de palabras ya visualizadas en otros países, llamando la atención nuevamente la aparición de los *intrapreneurship* en el año 2010. De forma

paralela, en particular para el año 2015 se visualizan similitudes con acrónimos relacionados a recintos educacionales de economía.

Finalmente, en el año 2019 podemos encontrar una nueva palabra que es *gender equality*, esto llama la atención puesto que Canadá es el único país que alude a temas de igualdad de género. La aparición de esta palabra podría referirse a que durante el 2019 se celebró en Canadá el "*Women Deliver 2019 Conference*", por lo que la discusión de igualdad de género estaría más en boga durante este año en los periódicos.

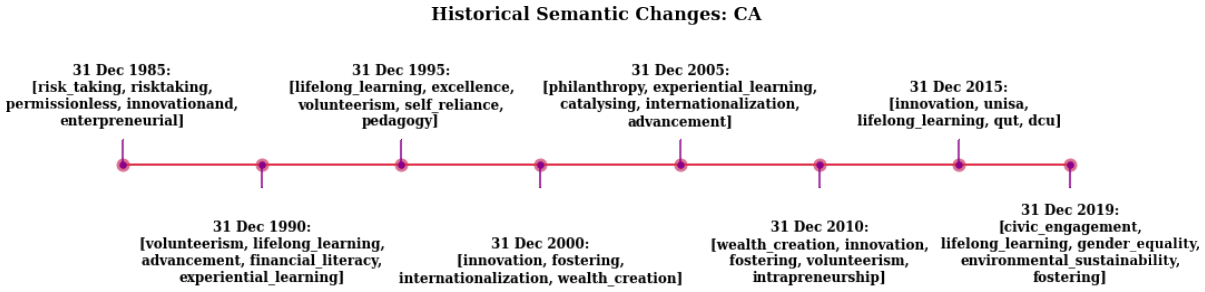


Figura 4.6: Evolución de la palabra *entrepreneurship* para Canadá.

4.2.6. Evolución Semántica de Australia

Comenzando con 1990, se puede visualizar que los emprendimientos tienen una mayor similitud con liderazgo al aparecer palabras como: *leadership* y *statesmanship*. Estas palabras rápidamente dejan de aparecer en años posteriores, por lo que se debería a un caso particular del corpus.

Para los años 1995 al 2019, se logra observar una conducta regular en las similitudes, ya que aparecen los mismos conceptos en cada uno de los años de registro. Entre las similitudes más destacadas podemos encontrar a *critical thinking*, *creativity* y *experimental learning*. En donde, si bien, no se ve una evolución diferenciada cada cinco años, el comportamiento señala con seguridad que los emprendimientos en Australia poseen como principal característica el pensamiento crítico, la creatividad y el aprendizaje en el desarrollo de los emprendimientos.

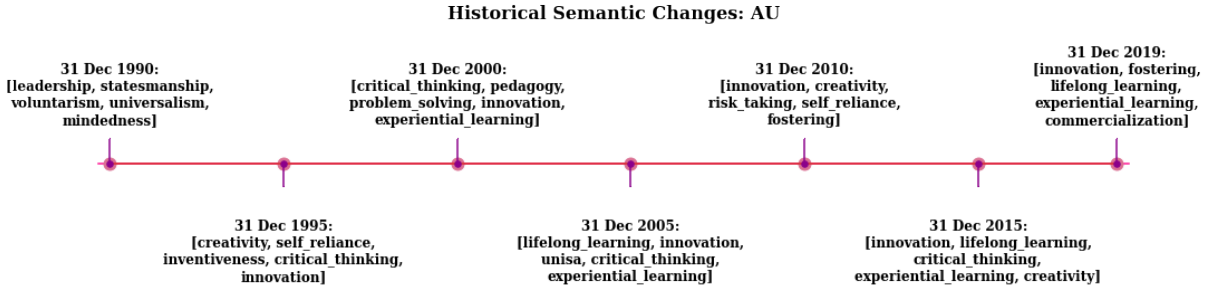


Figura 4.7: Evolución de la palabra *entrepreneurship* para Australia.

4.3. Evolución de Emprendimientos Utilizando Palabras Anclas

Considerando la metodología explicada en 3.3.3.3, se utilizan palabras anclas para visualizar la evolución de emprendimiento a lo largo de los años. Para esto, el conjunto de palabras anclas representan conceptos relacionados con el desarrollo social de una sociedad, por esto, se pueden encontrar palabras como *health*, *science*, entre otras.

El objetivo de este ejercicio es visualizar si el concepto de emprendimiento posee una evolución a través de los años frente a palabras que reconocemos como parte positiva en el desarrollo de una sociedad, comparando esta evolución frente a todos los países.

De los resultados, se espera a priori, visualizar aumentos en las similitudes respecto a ciertas palabras anclas a medida que se avanza en los años. Buscando visualizar potenciales diferencias en la evolución de la cercanía del concepto emprendimiento para cada uno de los países. Buscando visualizar la dirección a la que apunta cada uno de los países y aportando más información en la interpretabilidad que posee este trabajo sobre los emprendimientos.

Obtenidas las similitudes para los diferentes países, son graficados los resultados en un mapa de calor que representa a través de cambios en la intensidad de los colores, la existencia de mayores o menores similitudes entre las palabras. Para caso presentado en los gráficos presentados en las siguientes subsecciones, podemos observar que a medida que el color se hace más intentos (amarillo), mayor es la similitud de conceptos relacionados a emprendimientos frente a una determinada palabra ancla.

Comenzando con Estados Unidos, En 4.8 es posible observar que tanto *foreigner*, *family* y *internet*, son las palabras que menor evolución presentan a lo largo del registro, permaneciendo con valores cercanos a 0.2 a lo largo de los años. Esta relación señala que los emprendimientos en Estados Unidos carecen de características familiares o relacionadas con extranjeros, siendo este último factor relevante ya que podría representar bajas oportunidades a personas extranjeras para la generación de emprendimientos en este país. Por otro lado, la baja similitud con internet nos señala pocos emprendimientos de este tipo y esto podría deberse a la complejidad que podría significar realizar un emprendimiento enfocado solamente en internet.

En segundo lugar, se observa un conjunto de palabras que evolucionan positivamente a lo largo de los años, pero los valores no superan los 0.4 puntos de similitud. De estos resultados resalta el aumento de la similitud con *small business* y *startup*, lo que nos señala el incremento en la aparición de pequeños emprendimientos en los últimos años, quienes son expuestos concurridamente en los periódicos. Por otro lado, resultados similares se ven para *health* y *environmental*, lo que infiere emprendimientos enfocados en la salud y medio ambiente.

Finalmente, en el último grupo de palabras resalta la gran similitud que tienen las palabras anclas con emprendimiento, encontramos las palabras *education*, *science*, *innovation* y *sustainability*. Las palabras educación, ciencia e innovación tendrían una gran relación en el contexto de emprendimiento serian una característica primordial para los emprendimientos. Ahondando en esto, la educación y la ciencia permite entregan herramientas para generar mayor innovación en los emprendimientos a desarrollar. La sustentabilidad señalada, estará

relacionada a la necesidad de generar emprendimientos con un enfoque en la sustentabilidad del medio ambiente, tema que ha tomado bastante importancia por el calentamiento global a lo largo de los años.

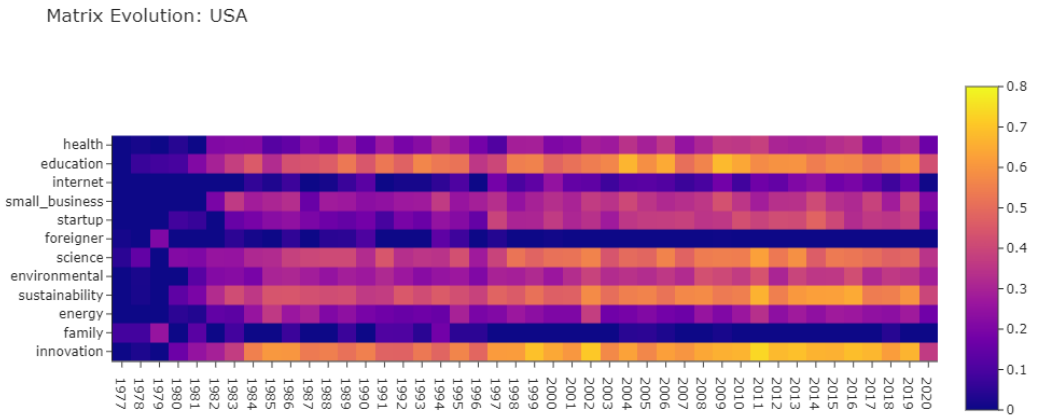


Figura 4.8: Matriz de similitud obtenida para Estados Unidos.

Estudiado los resultados para Estados Unidos, en 4.9 se logran ver resultados muy similares a los ya analizados, pero con una menor cantidad de registros. De los resultados, se observa una baja similitud en términos relacionados a extranjeros y familia, quienes presentan una nula relación al emprendimiento a lo largo de los años. Esto se podría interpretar como un bajo enfoque en los emprendimientos en estos tópicos para este país. Por otro lado, se mantienen los puntos que resaltan en la evolución son *education*, *science*, *innovation* y *sustainability*. Pero se observan valores menores en las palabras *small business* y *startup*, lo que significará una menor presencia de este tipo de emprendimientos en el país oceánico.

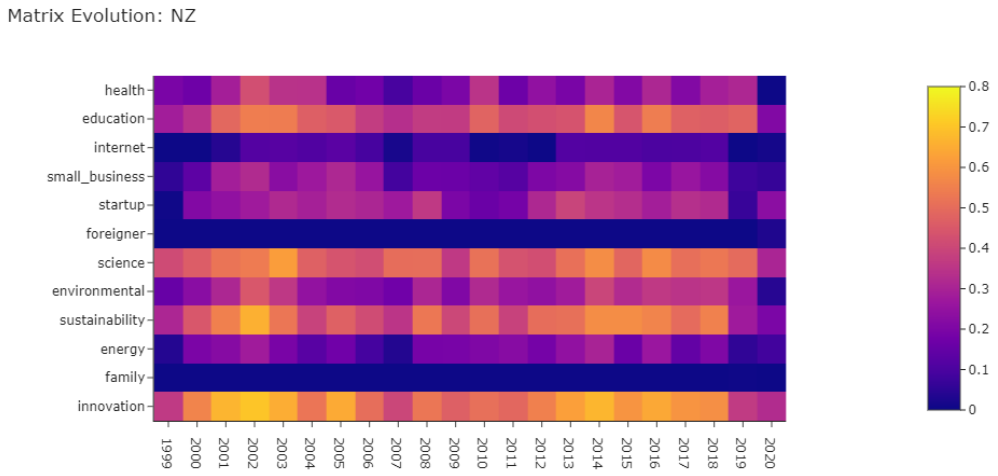


Figura 4.9: Matriz de similitud obtenida para Nueva Zelanda.

Para el caso de 4.10 y 4.11 se logra ver una evolución similar para ambos países, resaltando un comportamiento similar al visualizado para Estados Unidos con leves diferencias en las similitudes. El resultado nos señala que los países han evolucionado con enfoques similares en la evolución de los emprendimientos.

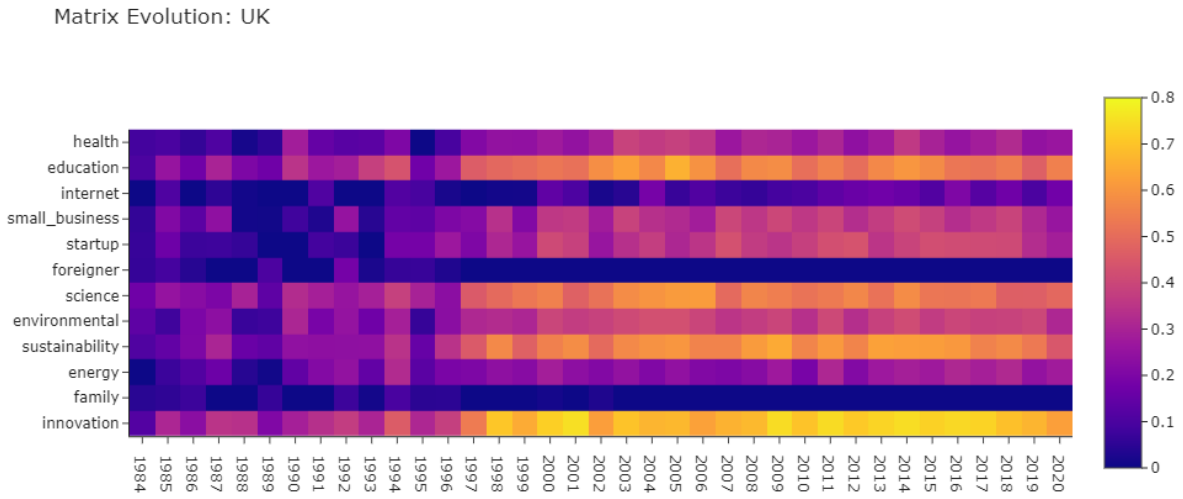


Figura 4.10: Matriz de similitud obtenida para Reino Unido.

Figura 4.11: Matriz de similitud obtenida para Irlanda.

Del resultado obtenido para Canadá en 4.12, se observa que las palabras con mayores similitudes se mantienen en valores similares desde el comienzo de los registros. Esto señalaría una mayor homogeneidad en el desarrollo de los emprendimientos, donde la importancia de

puntos como: ciencia, innovación, sustentabilidad y educación en todos los años son un pilar fundamental para este país.

Por otro lado, en 4.12 se mantienen valores bajos para *internet*, *foreigner* y *family*, repitiéndose las mismas características de falta de oportunidad en estas áreas de emprendimiento.

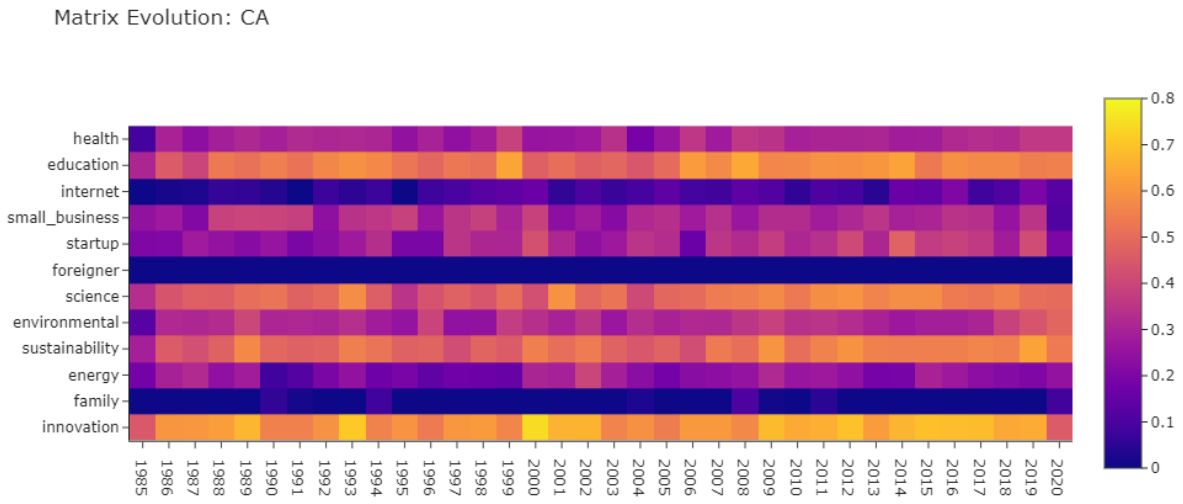


Figura 4.12: Matriz de similitud obtenida para Canadá.

Para el caso de Australia, los resultados señalan un comportamiento similar a los visualizados con Estados Unidos y los otros países, resaltando solamente la mayor similitud que se obtienen en la innovación en comparación a los otros países alcanzando valores cercanos a 0.8.

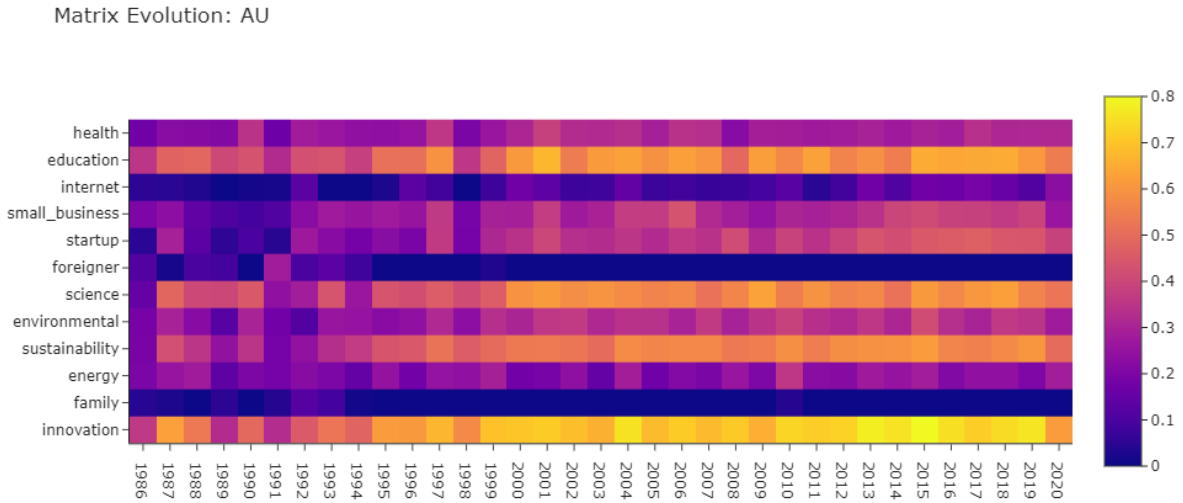


Figura 4.13: Matriz de similitud obtenida para Australia.

Analizada la evolución de las similitudes para cada uno de los países, se puede visualizar que los resultados siguen un patrón de forma general en el aumento de la similitud a través de los años. Con esto, se puede observar que los conceptos: *innovation*, *science*, *sustainability* y *education*, presentan una evolución positiva para cada uno de los países estudiados a medida que se avanza en los años. Del patrón señalado se podría comentar que los emprendimientos desarrollados en los países tienen en común la generación en estos tópicos, por ello existe una mayor relación semántica entre estas palabras y el concepto emprendimiento.

De los resultados analizados resalta la diferencia del concepto emprendimiento con las palabras *family* y *foreigner*, lo que nos señala que desde el conjunto de noticias estudiado no se realiza una mención frecuente de emprendimientos enfocados o desarrollados en la familia o extranjeros. Este resultado podría implicar que los países estudiados no han impulsado el desarrollo de emprendimientos enfocados en la familia y extranjeros.

4.3.1. Evolución de Emprendimiento

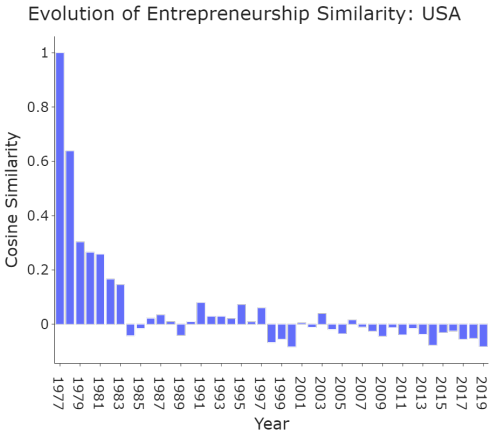
Como ultima tarea para estudiar la evolución de los emprendimientos a lo largo de los años, se gráfica la evolución de la similitud entre el primer registro de emprendimiento, y las siguientes palabras desambiguadas referidas a emprendimiento en el registro. Con esto, se obtienen los resultados expuestos en 4.14, donde los valores positivos señalan una similitud y los negativos un concepto opuesto entre las comparaciones.

El objetivo de esta tarea es confirmar la variación de lo que se interpreta como emprendimiento para cada uno de los países a través de los años. Buscando visualizar países donde se exponga una alta variación, y otros que no son tan susceptibles a las variaciones a lo largo de los años.

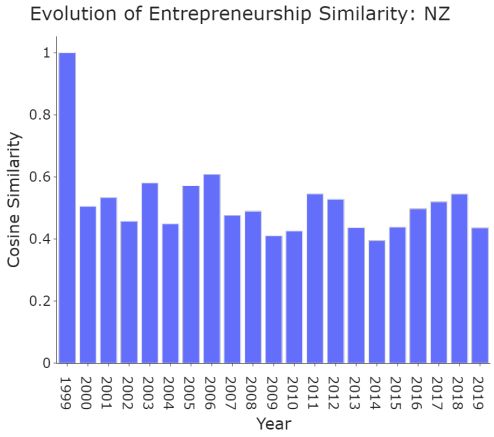
De los resultados se reconocen tres comportamientos: El primero, es la gran variación que han experimentado los países Estados Unidos e Irlanda a lo largo de los años, notando una gran diferencia en lo que representan los emprendimientos en el año 2019 en comparación al primer año de registro en ambos países. Esto señala que los países han tenido cambios consistentes a lo referido a emprendimientos, siendo totalmente diferentes a lo que representaban 2 décadas atrás.

En segundo lugar, se ve el caso de los países Reino Unido y Australia, quienes experimentan una alta variación entre el primer registro y los registros venideros, presentando a lo largo de los años pequeñas variaciones de alrededor 0.1 puntos de diferencia que señalan pequeñas variaciones a través de los años en lo que respecta a emprendimiento.

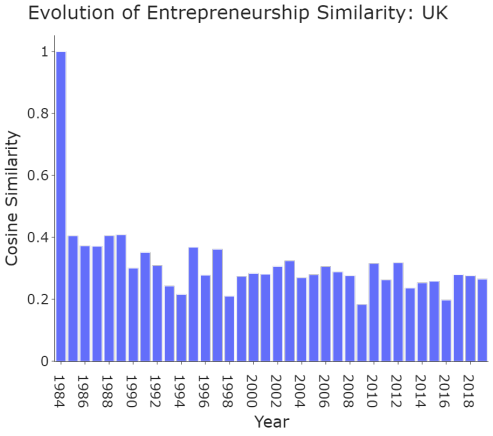
Finalmente se encuentra el caso de Canadá y Nueva Zelanda, quienes presentan similitudes superiores a 0.4 y con registros que no poseen altas variaciones a lo largo de los años. Este comportamiento viene a confirmar los resultados ya obtenidos, donde para el caso de Nueva Zelanda se han visto resultados que por persistentes en la relación de este país con temas educativos. Mientras que por otro lado para Canadá, se ve un comportamiento similar con los obtenidos en 4.12, donde demuestra ser un país con similitudes homogéneas con todas las palabras claves comparadas a lo largo de los registros.



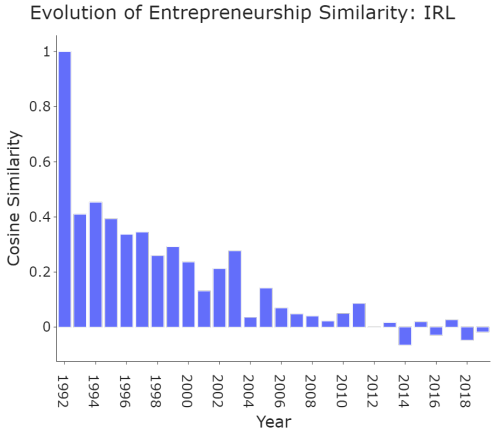
(a) USA



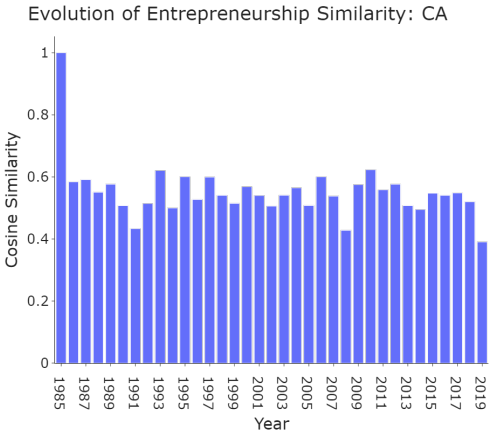
(b) NZ



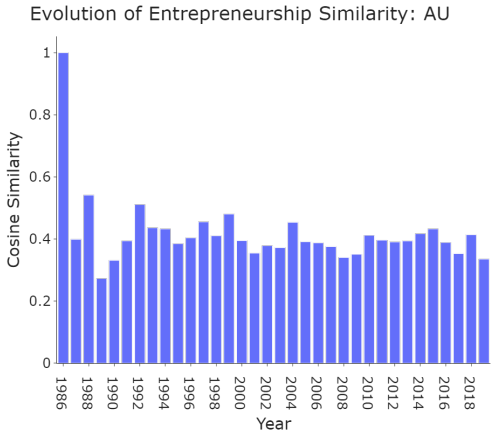
(c) UK



(d) IRL



(e) CA



(f) AU

Figura 4.14: Evolución de la similitud de emprendimiento respecto a la primera vez que se tiene registro de la palabra para los corpus de noticias en cada país.

4.4. Relación de grupos sociales a conceptos claves de emprendimiento

A continuación, son observadas y estudiadas las diferencias en la similitud de coseno que poseen los grupos sociales señalados en 5 respecto a múltiples palabras de interés que logran caracterizar a los emprendimientos. El objetivo de esta tarea es caracterizar a los grupos sociales a través de las similitudes que experimentan con las palabras de interés; contrastando las similitudes entre los diferentes grupos, de tal forma de generar una diferenciación de potenciales ventajas de un grupo frente al otro. Cabe señalar que este ejercicio se realiza utilizando la totalidad de los datos, buscando tener una visión general de la diferenciación esperada.

De los resultados obtenidos, el eje y representa el conjunto de palabras claves que son comparadas con los términos asociados a los grupos sociales. Por otro lado, en el eje x se tienen las similitudes de coseno obtenidas para cada una de las comparaciones realizadas. De las similitudes, mientras más negativos son los valores en el eje x, menos similitud existe entre los términos y las palabras claves. Por otra parte, mientras más positivos y cercanos a 1 sean los valores en x, mayor es la similitud de los términos a las palabras claves. Finalmente, para diferenciar los grupos antagonistas, estos son representados con círculos de colores señalando el grupo social que representan a través de una leyenda en el gráfico.

4.4.1. Similitud Observada para Términos de Genero

El primer grupo de términos a estudiar son los relacionados a hombres y mujeres. De los resultados expuestos en 4.15 podemos observar que gran parte de los valores son negativos, esto nos dice que en general las similitudes no se encuentran muy relacionadas a los términos comparados, ya que una buena relación de coseno debería ser superior a 0 y cercana a 1.

Dentro de las palabras que poseen una mayor similitud a los términos de genero se encuentran familia y negocio familiar, donde para familia se logra relacionar mayormente a términos femeninos, llamando la atención que esto se revierte para negocio familiar, donde obtiene una mayor similitud con términos relacionados a hombres. Este comportamiento podría deberse a una mayor cercanía a la generación de negocios por parte de conceptos masculinos, generando una mayor asociación a estos términos a pesar de que términos femeninos estén más relacionados a la familia.

Luego, al analizar las brechas que existen entre las similitudes obtenidas para cada palabra del eje y, se observa que las diferencias entre las similitudes no se escapan más allá de 0.1 puntos y por lo general son inferiores a 0.05. Entre las mayores brechas observadas, se logra relacionar a las mujeres con las palabras *networking*, *incubator*, *accelerator*, *small bussines*, *enviormental* y *sustainabulity*; sugiriéndonos que las mujeres poseen una mayor preocupación por el medio ambiente y tener características de precursoras al momento de realizar un emprendimiento. Por otro lado, entre los puntos donde más destaca una similitud hacia los hombres, nos encontramos con las palabras: *profit*, *failed*, *investor* y *developer*; lo que nos señala que los términos masculinos están más relacionado a la inversión y desarrollado de los emprendimientos, donde a pesar de tener una mayor relación con fracasos, logran estar más relacionados a la generación de ganancias al estar relacionados con *profit*.

Si bien los resultados nos exponen bajas similitudes con las palabras de interés y pequeñas brechas entre los términos comparados. Estos nos sugieren potenciales caracterizaciones contrastables con la realidad, verificando aspectos desfavorables como mayor asociación hacia la generación de ingresos que poseen los términos masculinos.

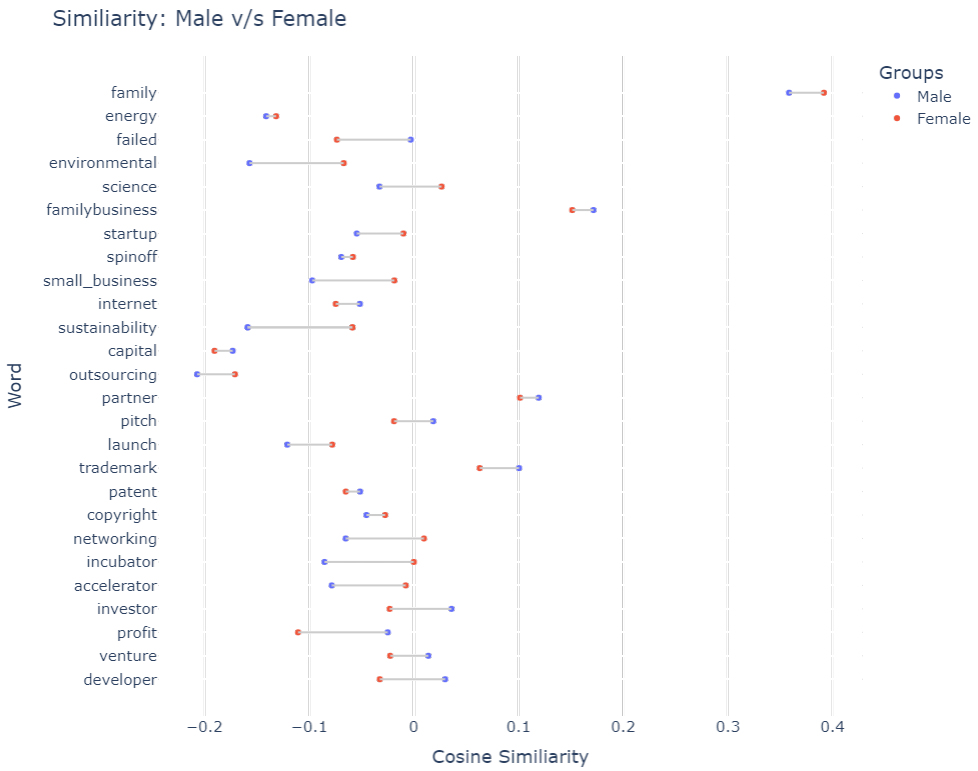


Figura 4.15: Similitud de coseno obtenidas entre las palabras anclas y términos masculinos y femeninos de la tabla 5.1. La positividad en los resultados señala una mayor similitud a los conceptos señalados en el eje y.

4.4.2. Similitud Observada para Términos de Ciudadanía

De los resultados obtenidos al comparar términos relacionados a Ciudadanos e Inmigrantes se visualiza una mayor cantidad de casos con similitudes superiores a 0. Esto nos señala una mayor relación en los términos analizados con las palabras estudiadas, destacando que gran parte de las palabras positivas están asociadas a los términos de ciudadano, mientras que gran parte de las palabras que poseen una similitud negativa son los términos relacionados a inmigrantes. Esta gran diferencia en la positividad de los resultados nos señala una clara diferencia de oportunidades entre ambos grupos, la que nos señalaría que los emprendimientos no estarían muy asociados a extranjeros, teniendo una baja participación términos de inmigración en el corpus.

Otro de los aspectos que resalta de los resultados son las grandes diferencias observadas en las similitudes para cada uno de los casos. Estos resultados en comparación a los obtenidos para los términos de género poseen una gran brecha, alcanzando valores cercanos a 0.2 puntos de similitud en los casos más distanciados. Dentro de los puntos que mayor diferencia

de similitud poseen, se encuentran los términos de sustentabilidad, ambiental y energía, lo que nos llevara asociar que los emprendimientos de extranjeros no poseen una conciencia del medio ambiente.

Por otro lado, al revisar cada uno de los resultados obtenidos, podemos percatarnos que solo en 4 de 26 puntos los inmigrantes logran obtener mayores similitudes a las palabras de interés. Destacando en estos 4 puntos el rol de inversor, ya que es uno de los puntos que más similitud alcanza en el conjunto de palabras estudiados. Esto nos señala que, si bien los inmigrantes no son reconocidos como precursores de emprendimientos, estos suelen estar asociados a la inversión de estos proyectos.

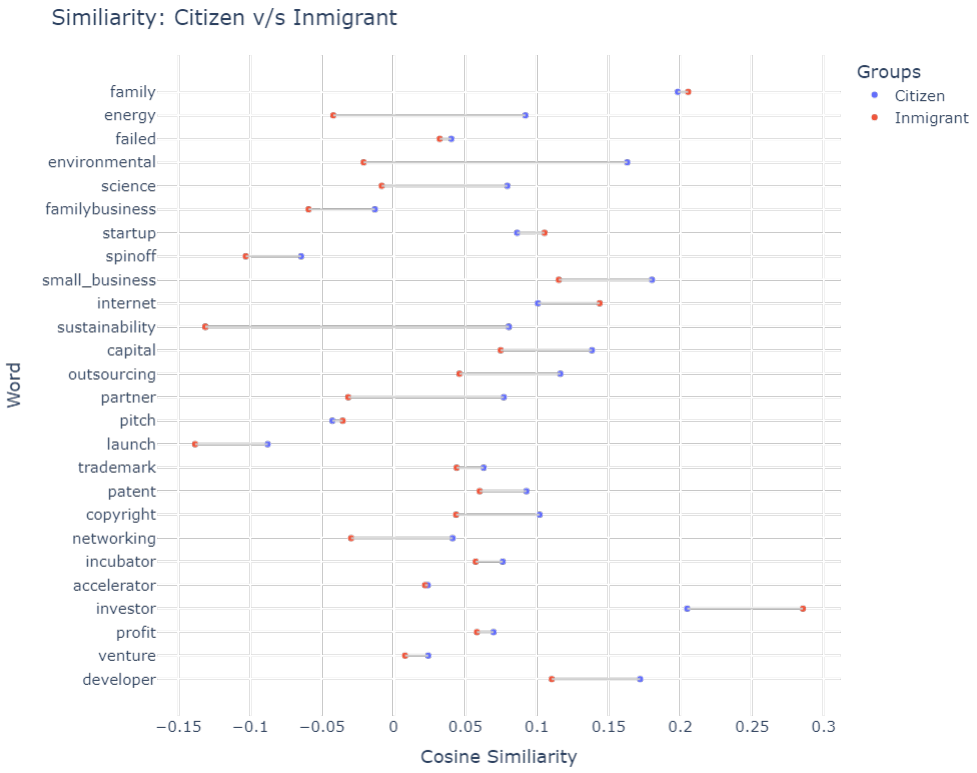


Figura 4.16: Similitud de coseno obtenidas entre las palabras anclas y términos relacionados a ciudadanos e inmigrantes de la tabla 5.1. La positividad en los resultados señala una mayor similitud a los conceptos señalados en el eje y.

4.4.3. Similitud Observada para Religiones

Comenzando con los resultados obtenidos en 4.17, se logran visualizar que gran parte de las similitudes están más relacionadas a los términos de judaísmo. Donde los términos judíos destacan en palabras que representan el desarrollo e inversión de un emprendimiento. Por otro lado, los términos referidos a cristianismo destacan en tres tópicos: ciencia, sustentabilidad y generación de ingresos (*profit*). Esto nos señala que los emprendimientos de personas cristianas poseen un enfoque más científico, con sustentabilidad y se podría asumir que estos generan mayores ingresos.

En segundo lugar al revisar los resultados de 4.18, notamos que gran parte de las similitudes están más relacionadas a cristianismo en vez de los términos islámicos. Esto nos dirá que los cristianos tienen mayores ventajas que los islámicos ante la generación de negocios. Sin embargo, es importante señalar que a pesar llama la atención que los puntos donde mayor similitud poseen los términos islámicos son fallar, inversor y capital. Estas similitudes nos permiten caracterizar a los islámicos como inversores, pero con más cercanía a fallar en sus negocios en comparación a los cristianos.

De los resultados se obtiene que los términos judíos son los que más relación a palabras precursores de emprendimiento, señalando ventajas en gran parte de los puntos frente a las otras religiones. Comparando estos resultados con la realidad es algo esperable, ya que las comunidades judías se caracterizan por poseer una conocida participación en el mundo de los negocios. Un ejemplo de esto es Estados Unidos, país donde han nacido grandes emprendimientos desarrollados por personas judías, algunos ejemplos de estos son: Oracle o Google.

Por otro lado, la baja relación de personas islámicas con los términos de interés podría deberse principalmente a una visión sesgada que poseen los países estudiados frente a esta religión. Esto producto de los atentados a Estados Unidos o Europa, las cuales son localidades de las que se posee una mayor cantidad de datos y afectarían en esta vista global de las religiones.

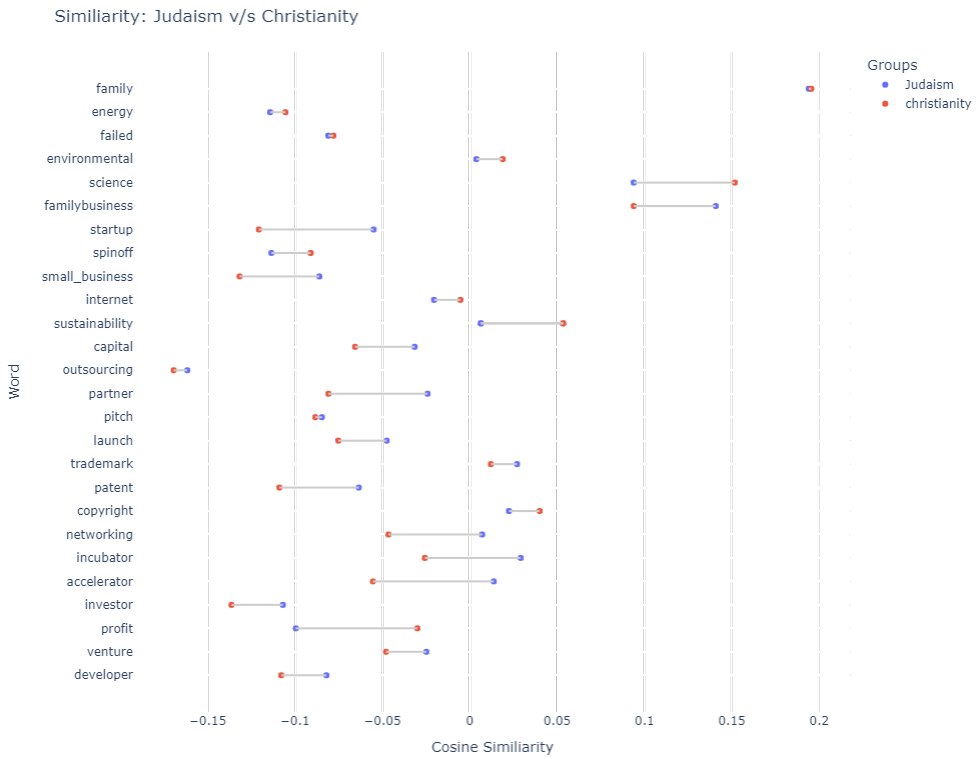


Figura 4.17: Similitud de coseno obtenidas entre diferentes términos que representan a las religiones cristianas, islámica y judía extraídos de la tabla 5.1.

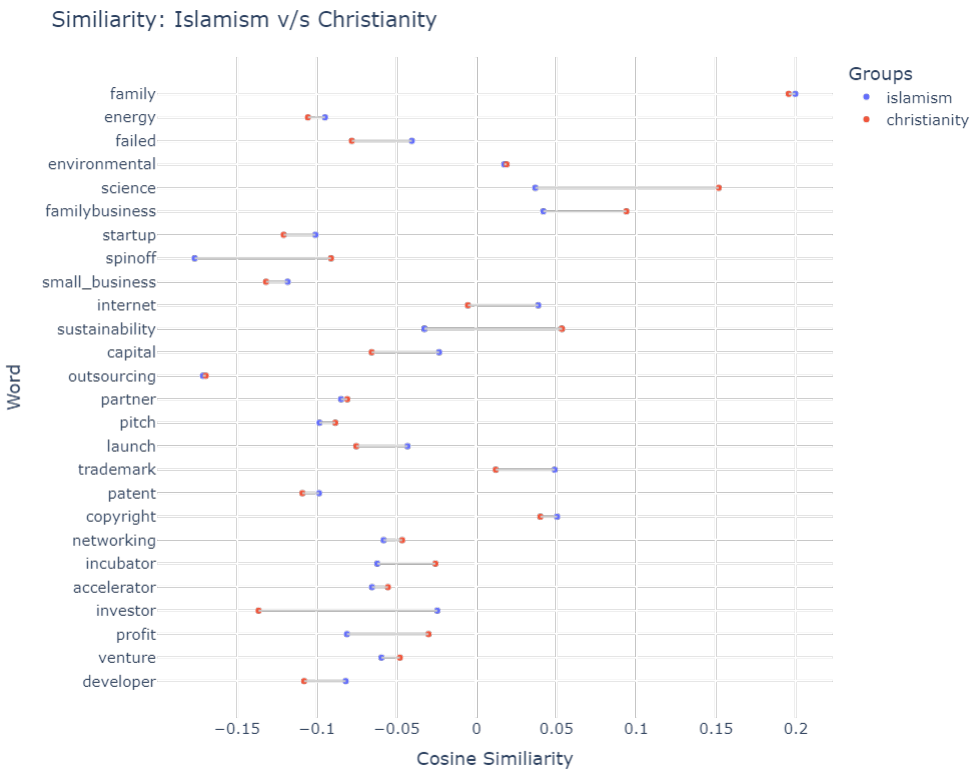


Figura 4.18: Similitud de coseno obtenidas entre diferentes términos que representan a las religiones cristianas, islámica y judía extraídos de la tabla 5.1.

4.4.4. Similitud Observada para Nombres Relacionados a Diferentes Etnias

A continuación, son comparados apellidos asociados a personas de raza blanca occidentales y apellidos de raza hispana y rusa respecto a las diferentes palabras de interés. De los resultados en 4.19 y 4.20 se observa que gran parte de las similitudes obtenidas tanto para los hispanos y rusos poseen una mayor relación a los apellidos de personas occidentales de raza blanca. Esta característica nos señala que la generación de emprendimientos en los países estudiados se da principalmente por personas de raza blanca occidental, lo que sería esperable debido a que la población de los países estudiados principalmente está conformada por personas de este tipo

En segundo lugar, llama la atención que tantos los apellidos rusos como hispanos poseen las mismas similitudes donde resalta una mayor similitud frente al grupo que son comparados. Con esto se observa que los únicos puntos donde mayor diferencia obtienen son la familia y capital. Esto llama la atención, ya que, a pesar de presentar una mayor similitud hacia familia y capital, los negocios familiares no se vean más relacionados a los apellidos rusos e hispanos.

En general se logran visualizar resultados bastantes similares en interpretabilidad con los obtenidos para ciudadano versus inmigrante, donde en cada una de las palabras estudiadas

destaca una mayor similitud a favor de los apellidos de hombres de raza blanca occidental, quienes representarían a ciudadanos en los países estudiados.

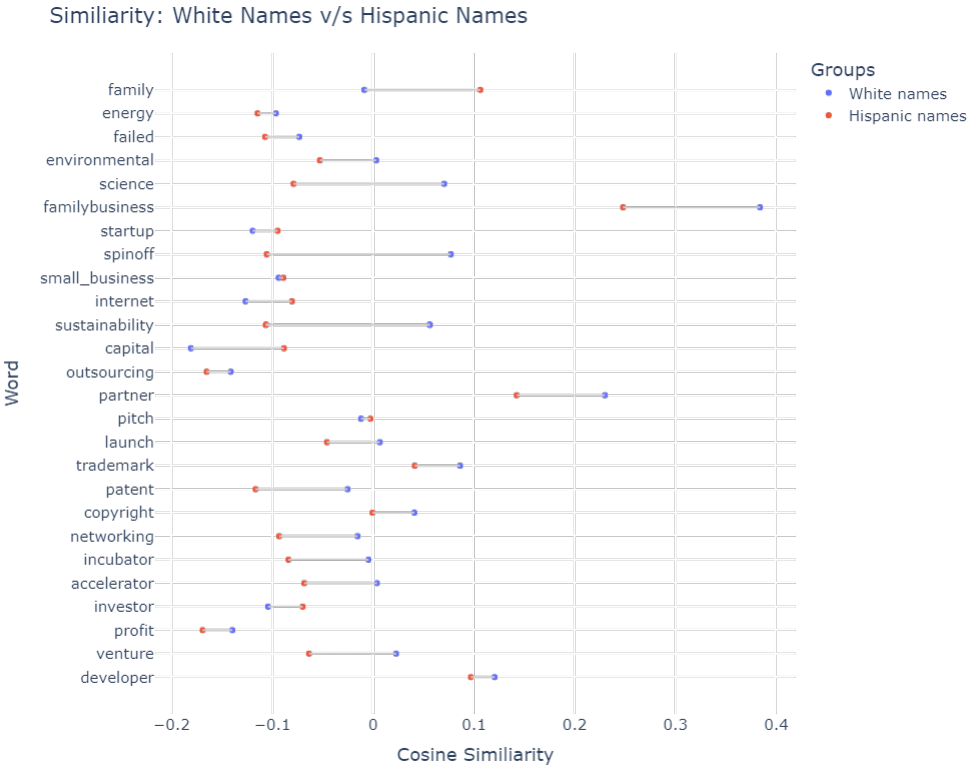


Figura 4.19: Similitud de coseno obtenidas entre diferentes términos que representan a las religiones cristianas, islámica y judía extraídos de la tabla 5.1.

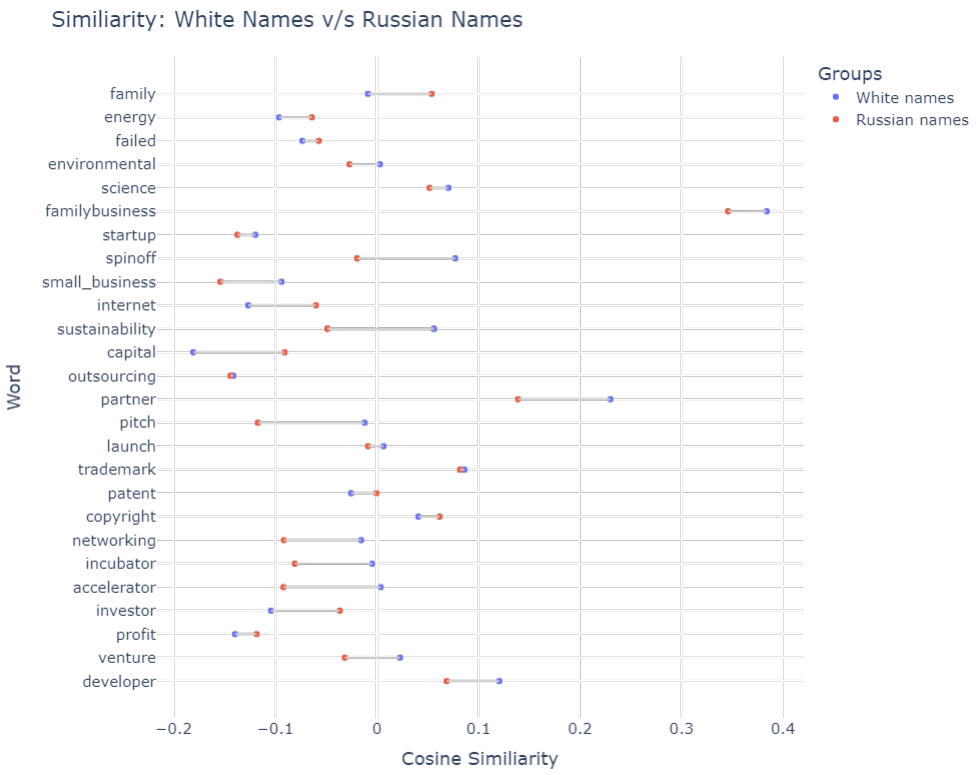


Figura 4.20: Similitud de coseno obtenidas entre las palabras de interés y apellidos relacionados a personas de raza blanca occidental y apellidos hispánicos y/o rusos extraídos de la tabla 5.1.

Si bien de los resultados que se logran visualizar en esta serie de experimentos se exponen cercanías que poseen ciertos grupos sociales a conceptos como inversores, dinero u otros términos. Gran parte de los resultados no presenta una gran diferencia que nos permita visualizar o concluir fehacientemente la tendencia de sesgo. Por esta razón se hace necesaria la aplicación de otras técnicas que nos permitan absorber de forma más profunda el sesgo que experimentan los grupos sociales señalados.

4.5. Medición de Sesgo a Través de RND

Como apoyo al estudio de la variación de sesgo presente en algunos grupos sociales. Se presentan los resultados obtenidos al aplicar la métrica de sesgo RND sobre los embeddings obtenidos. El objetivo de este ejercicio es ver a través de un puntaje global como se relacionan conceptos relacionados a emprendimientos a diferentes grupos sociales; de esta forma, a través de un gráfico podremos ver la evolución que ha experimentado el sesgo a través de los años.

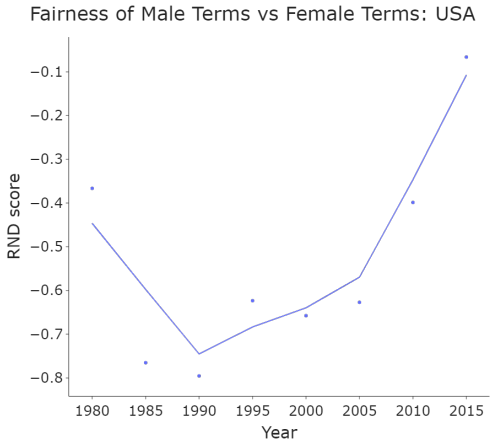
Los resultados son expuestos en 4.21 y 4.22. De estos, se pueden visualizar múltiples gráficos para cada uno de los países estudiados, donde la temática central de estos estudios es enfocada en las temáticas de género y nacionalidad. En el eje y de cada uno de los gráficos se expone el puntaje obtenido en la métrica RND para cada año de registro, en ellos los valores negativos se referirá a una mayor cercanía al primer grupo señalado en cada uno de los gráficos. Por ejemplo, para el caso de sesgo de genero se señala Términos de hombre vs términos de mujer; en esta lógica los valores negativos estarán asociados a términos de hombre, mientras que los positivos a términos que se refieren a mujeres. El caso ideal de estas comparaciones es obtener un puntaje 0 en cada uno de los años, lo que señalaría una nula diferencia entre los términos comparados.

Para el primer grupo de resultados compuestos por la comparación de términos relacionados a los genero masculino y femenino; se observa que en su totalidad todos los países obtienen valores inferiores a 0 para cada uno de los años de registro, esto nos señala que existe una mayor relación entre los conceptos de emprendimiento con el género masculino. Por otro lado, a pesar del resultado negativo hacia las mujeres, se observa una constante baja en la relación hacia los hombres a través de los años.

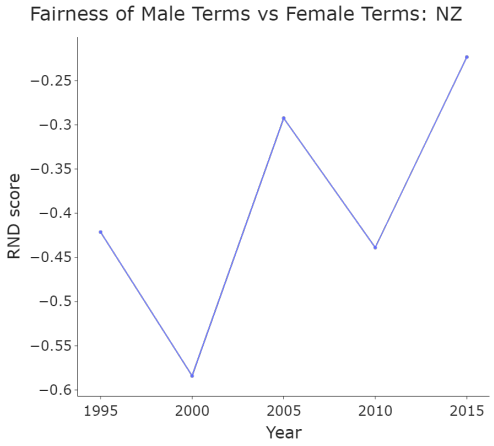
Con esto, destaca entre los países estudiados los puntajes obtenidos para Canadá, quien es el país que posee una evolución más marcada hacia las mujeres a través del tiempo, y la que obtiene valores más altos. Estos resultados concuerdan con lo visto anteriormente en la evolución semántica de Canadá, quien era el único país que señalaba igualdad de género entre las palabras más similares a emprendimiento.

Para el caso de la comparación con los términos relacionados a ciudadanos e inmigrantes, nuevamente se ven resultados cargados hacia un grupo. Donde, para cada uno de los países se observa una mayor relación de los emprendimientos con ciudadanos, alejándose más puntos que los vistos en la comparación de género.

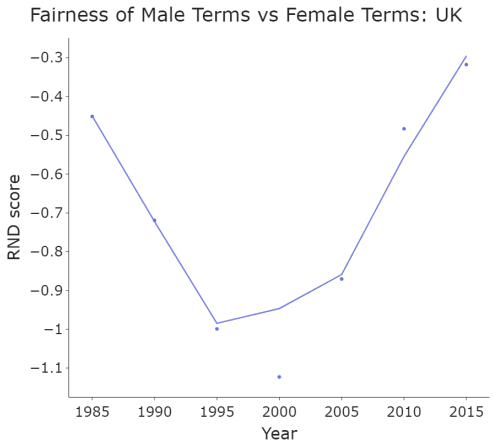
En segundo lugar, al observar la evolución en cada uno de los países se logran observar dos tendencias: la primera son un grupo de países donde se ve una variación positiva hacia los inmigrantes, pero que de igual forma al llegar al último registro se observan valores alejados de 0, y con una alta diferencia obtenida para la comparación realizada para los géneros. En segundo lugar, se observan otra mitad de países que poseen una variación positiva hacia a emprendimientos con inmigrantes.



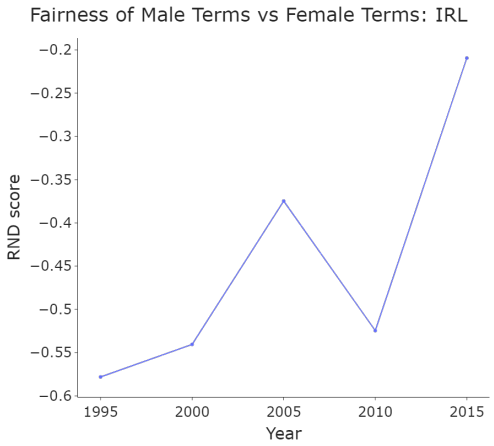
(a) USA



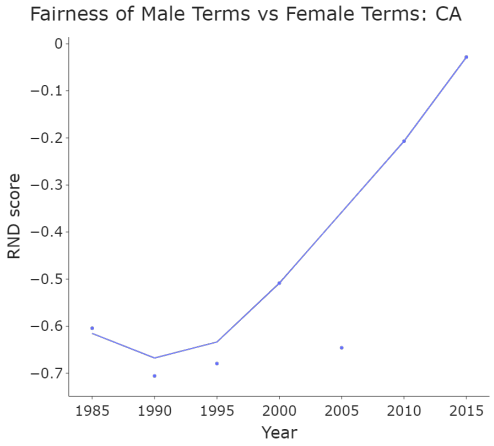
(b) NZ



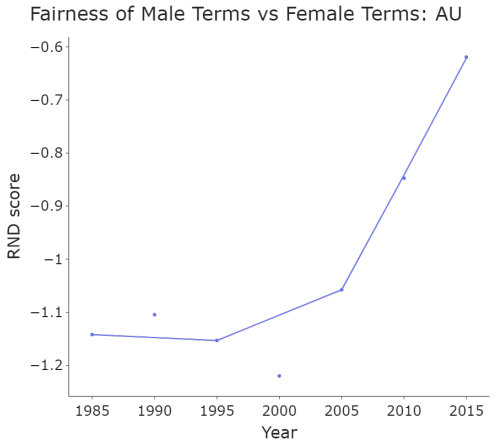
(c) UK



(d) IRL

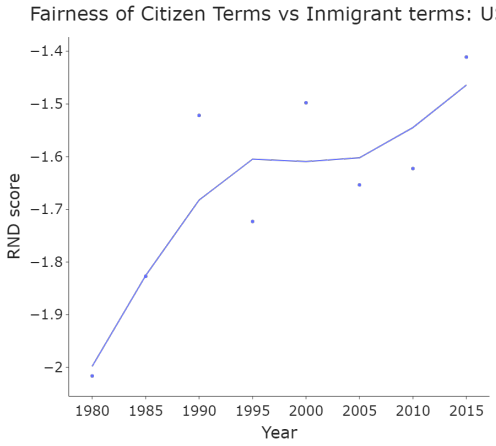


(e) CA

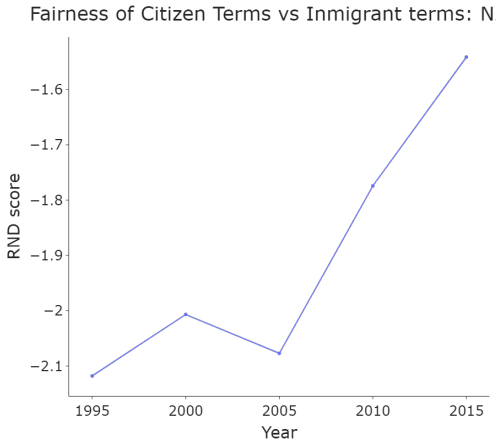


(f) AU

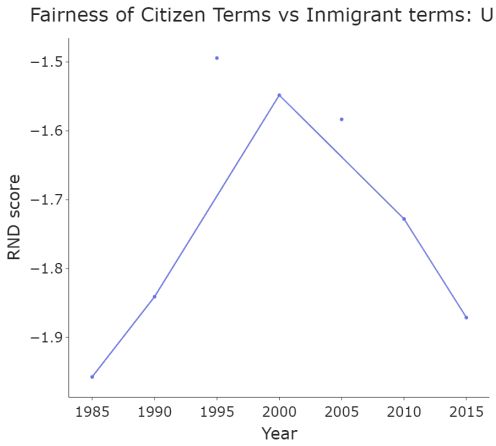
Figura 4.21: Variación del sesgo obtenido bajo la métrica RND, entre los grupos conformado por términos de Hombre y Mujer. De los gráficos, mientras más negativo sean los valores obtenidos por la métrica RND, mayor es la relación que tienen los términos masculinos con emprendimiento.



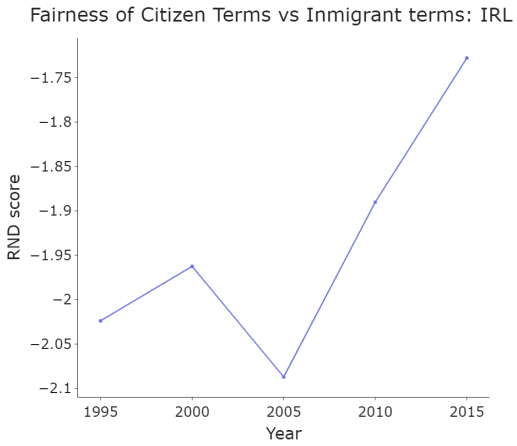
(a) USA



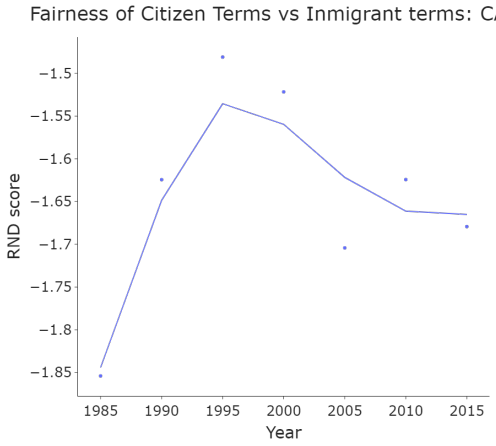
(b) NZ



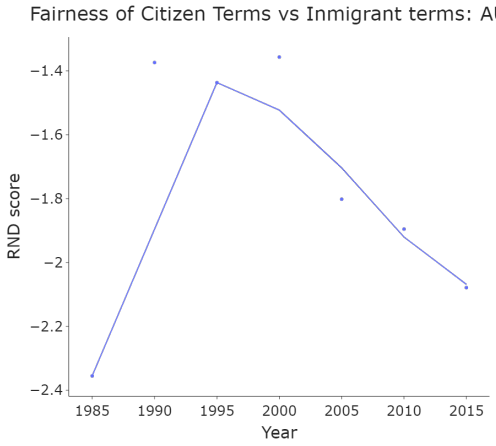
(c) UK



(d) IRL



(e) CA



(f) AU

Figura 4.22: Variación del sesgo obtenido bajo la métrica RND, entre los grupos conformado por términos de Ciudadano y extranjero. De los gráficos, mientras más negativo sean los valores obtenidos por la métrica RND, mayor es la relación que tienen los términos de ciudadanos con emprendimiento.

4.6. Predicción del Sentimiento Asociado a Emprendimiento a lo Largo de los Años

Como ultimo resultado se presenta la predicción de los sentimientos asociados a emprendimientos a lo largo de los años utilizando un lexicón marcado con sentimientos positivos y negativos. Este experimento tiene como objetivo visualizar potenciales cambios en el sentimiento asociado a emprendimiento a lo largo de los años, donde se espera visualizar bajas en la positividad para periodos económicos donde existen recesiones. El resultado esperado, nos podría exponer que tan susceptibles son los medios de comunicación ante las recesiones para exponer información relacionada a emprendimientos, verificando si existe o no un incentivo a no generar emprendimientos en estos periodos.

Los resultados son expuestos en el conjunto de gráficos en 4.23, donde en cada uno de ellos el eje y representa el porcentaje de positividad asociado a emprendimiento predichos por el modelo de regresión logística; mientras que en el eje x se presentan los años de registros para cada país. Cabe señalar que, de los resultados, mientras mayores sean los valores obtenidos en el eje y, mayor positividad tendrán los conceptos de emprendimientos en un determinado año.

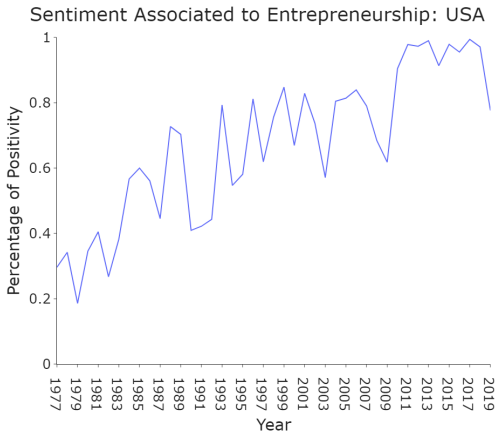
De los gráficos obtenidos, es posible observar dos comportamientos bien marcados, el primero es un conjunto de países conformados por: Estados Unidos, Reino Unido, Irlanda y Australia, quienes poseen una tendencia positiva hacia el concepto de emprendimiento a medida que pasan los años. Por otro lado, se observa un segundo comportamiento para los países Nueva Zelanda y Canadá, quienes obtienen una predicción plana del porcentaje de positividad, donde se observa en cada año una escasa tendencia a medida que se avanza en los registros.

Dentro del caso de los países que evolucionan positivamente a lo largo de los años, se observa que los decaimientos en la positividad se dan en periodos donde han existido recesiones económicas mundiales, o eventos locales que han afectado la economía en dichos países. Tal es el caso por ejemplo de Estados Unidos, donde se logran ver considerables bajan en grandes recesiones como lo fueron la de 1982, 1991, 2009 e incluso se comienza a visualizar una baja con la reciente recesión producida por la pandemia [39].

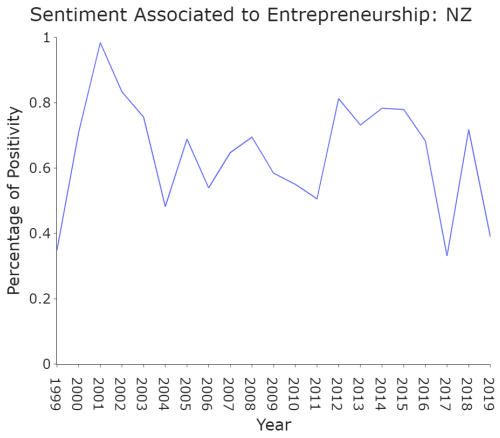
Para el segundo caso de los países que no poseen una tendencia clara en la positividad, para el caso de Nueva Zelanda puede deberse a la poca cantidad de registros que se poseen, exponiendo solo los últimos resultados en comparaciones a los otros países que muestran un registro más amplio de años. Sin embargo, es un aspecto llamativo la nula tendencia que presenta el país oceánico. Por otro lado, para el caso de Canadá llama la atención que gran parte de las predicciones poseen valores elevados en comparación a los otros países, presentando una nula tendencia producto a que este país ha demostrado, a través de los experimentos, tener una homogeneidad en sus características de emprendimiento a lo largo de los años.

Finalmente, de los resultados se logra visualizar un grupo de países en donde los periódicos muestran ser susceptibles a las recesiones económicas, donde, de una u otra forma tienden a exponer una negatividad en su lenguaje al pasar por inestabilidades. Por otro lado, se tienen medios de comunicación neutrales para el caso de Nueva Zelanda y Canadá, quienes no exponen un lenguaje muy susceptible a las recesiones, o las características que representan

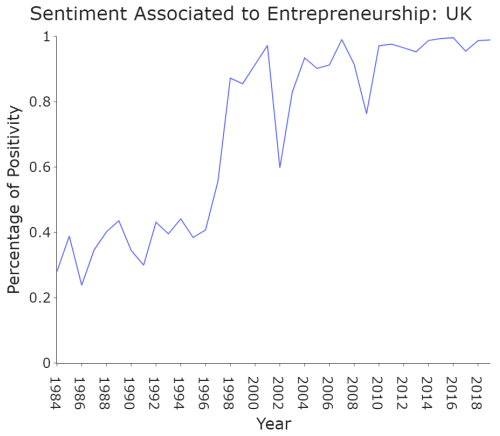
sus emprendimientos son homogéneas a lo largo de los registros que se poseen.



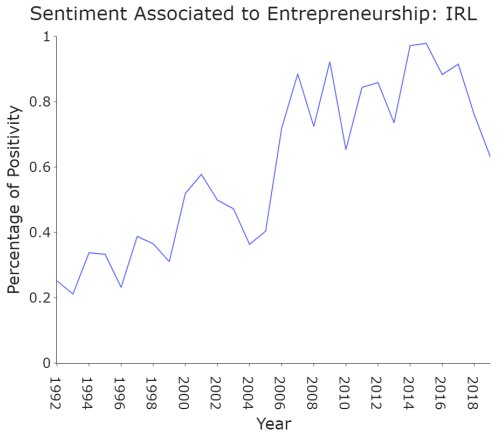
(a) USA



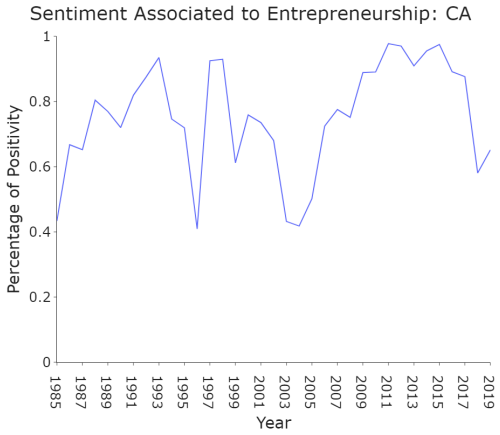
(b) NZ



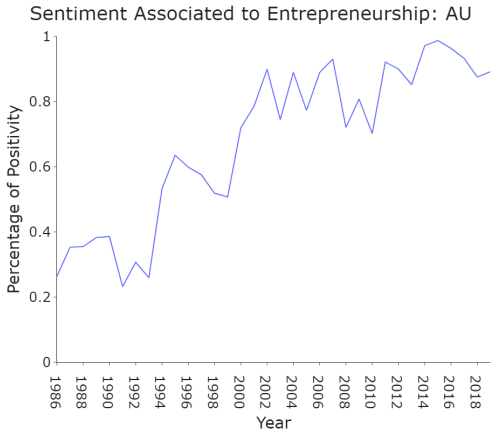
(c) UK



(d) IRL



(e) CA



(f) AU

Figura 4.23: Evolución del sentimiento asociado a "entrepreneurship" a lo largo de los años para cada uno de los países estudiados. En el gráfico, el eje y representa la positividad predicha por un modelo de regresión logística en base a conceptos relacionados a emprendimiento.

4.7. Resumen y Caracterización de los Países

En vista de los resultados analizados en este capítulo, a continuación, se expone un resumen de los aspectos más relevantes encontrados para cada país en este estudio.

4.7.1. Caracterización de Estados Unidos

- País con un enfoque más relacionado a la generación de riquezas en temas de emprendimiento.
- Gran variación por lo que se comprende con emprendimiento a lo largo de los años. El país experimenta variaciones dependiendo de eventos locales como políticos u económicos.
- Presenta alto sesgo hacia las mujeres e inmigrantes. Sin embargo, ha experimentado un cambio positivo en relación con la igualdad de género.
- El país es susceptible a las recesiones, mostrando variaciones negativas en la positividad que se ven los emprendimientos. Por otro lado, el país ha evolucionado hacia un sentimiento positivo hacia los emprendimientos a lo largo de los años.

4.7.2. Caracterización de Nueva Zelanda

- Los emprendimientos tienen como principal característica la educación.
- El país presenta una baja variación en los conceptos relacionados a emprendimiento a lo largo de los años, manteniendo en todo el registro una gran similitud con palabras relacionadas a la educación.
- Presenta alto sesgo hacia las mujeres e inmigrantes. Sin embargo, ha experimentado un cambio positivo en relación con la igualdad de género.
- Debido al bajo número de registros que se posee para este país no se observa una tendencia al crecimiento de la positividad en el país. Sin embargo, se los puntos negativos señalan puntos de recesiones mundiales.

4.7.3. Caracterización de Reino Unido

- A pesar de observar en unos primeros resultados que el país poseía una mayor relación a la generación de riquezas. Al estudiar la evolución de la semántica se observa que los emprendimientos en este país poseen un rol social y a lo largos de la evolución siempre se les puede asociar a tópicos sociales.
- En términos de evolución, los emprendimientos poseen pequeñas variaciones a lo largo de los años respecto a emprendimientos.
- Presenta alto sesgo hacia las mujeres e inmigrantes. Se experimenta un cambio positivo en relación con la igualdad de género, pero un estancamiento en el sesgo asociado hacia los inmigrantes.

- Respecto a la evolución de los sentimientos, el país tiende a percibir a los emprendimientos como positivos. Al igual que los otros países presentan bajas en la positividad en las recesiones

4.7.4. Caracterización de Irlanda

- País con similitudes similares a Reino Unido, donde destaca el rol social que posee y la sustentabilidad del medio ambiente.
- Al estudiar la variación que ha experimentado el concepto emprendimiento a lo largo de los años, es uno de los países que más cambios presenta. Esto puede deberse al gran nivel de ruido que presenta en los primeros años, ya que al estudiar la variación en años posteriores al 2000, se observa una característica social.
- El país presenta preferencia hacia los hombres y a los ciudadanos. Si bien se presentan sesgos negativos en temas de genero e inmigración, se ve una tendencia a mejorar con los años.
- El país tiende a percibir a los emprendimientos como positivos al avanzar en los años.

4.7.5. Caracterización de Canadá

- Destaca la similitud que poseen los emprendimientos con la internacionalización e igualdad de género, siento el único país que señala una similitud considerable con igualdad de género.
- A lo largo de los años posee una baja variación en lo que se interpreta como emprendimiento, teniendo las mayores similitudes respecto a palabras relacionadas a emprendimiento.
- Si bien presenta un sesgo hacia la generación de emprendimientos por migrantes, posee los resultados más igualitarios respecto a género.
- El sentimiento asociado a través de los años no presenta una tendencia clara, manteniéndose plano exceptuando en puntos donde aparecen recesiones. Es por esto que en general para este país se considera a los emprendimientos como algo positivo.

4.7.6. Caracterización de Australia

- Baja variación en los conceptos asociados a emprendimiento a lo largo de los años, destacando el pensamiento crítico y creatividad como características de los emprendimientos para este país.
- Presenta sesgo negativo hacia la mujer, pero con una tendencia positiva a lo largo de los años. Por otro lado, se observa un estancamiento negativo en el sesgo asociado hacia los inmigrantes.
- El sentimiento asociado a través de los años presenta una tendencia positiva clara, repitiéndose el factor de decaimientos en puntos recesivos.

Capítulo 5

Conclusiones

En el desarrollo del presente trabajo, se han propuesto y desarrollado múltiples experimentos para la extracción de información descriptiva para la comprensión del concepto emprendimiento. Para esto, se ha estudiado la evolución de la semántica y sesgo asociados a emprendimientos logrando el objetivo principal de esta memoria señalado en 1.3; ya que a partir de texto de periódicos se ha logrado extraer *Word Embeddings* que nos permiten caracterizar y diferenciar cómo se perciben los emprendimientos en seis países de habla inglesa.

Por medio de los *Word Embeddings* se obtienen similitudes entre las diferentes palabras que posee el vocabulario de entrenamiento, notando claras diferencias en las similitudes que poseen las palabras desambiguadas de cada país. Este ejercicio nos señala una variación respecto a la visión de emprendimiento que poseen los países estudiados, visualizando a Nueva Zelanda como un país con emprendimientos enfocados en educación, mientras que Estados Unidos en la creación de riquezas.

Al estudiar la evolución semántica de cada uno de los países, se observan variaciones temporales que nos señalan que los emprendimientos no son un concepto constante en el tiempo, si no que es un concepto susceptible a los cambios que se generan en determinados periodos y mutan de acuerdo con la evolución que experimentan las sociedades.

Utilizando un clasificador de regresión logística se logra predecir la positividad asociada a los emprendimientos para el conjunto de años que se tiene registro para cada país. De la predicción se observa que gran parte de los países poseen un aumento sustancial de la positividad hacia los emprendimientos, teniendo bajas en periodos de recesión mundial y/o locales de cada país. Por otro lado, para los países Nueva Zelanda y Canadá se ven comportamientos más planos en donde la positividad no posee una tendencia clara.

Respecto a los experimentos de sesgo, a través de la similitud de coseno y la métrica de sesgo RND se logra visualizar que gran parte de los países estudiados poseen un comportamiento similar. Resaltando el sesgo que poseen hacia las mujeres e inmigrantes, donde se observan menores oportunidades por los constantes resultados negativos en los experimentos. Sin embargo, de todo el conjunto de países estudiados, resalta Canadá quien es el único país que en múltiples experimentos señala variaciones positivas en la igualdad de género.

En general los experimentos realizados en esta memoria permiten obtener una visión

resumida de cómo son interpretados los emprendimientos por los periódicos. Extrayendo información y evidencia de la visión desarrollada por diferentes países respecto a conceptos relacionados a emprendimiento, donde se logra caracterizar y exponer los sesgos sociales que poseen las sociedades estudiadas.

Trabajo Futuro

Si bien en esta memoria son propuestos múltiples experimentos que desarrollan un conocimiento para el área de emprendimiento, estos podrían ser desarrollado más afondo mitigando algunos aspectos visualizados en el desarrollo de este trabajo y/o aplicando nuevas tecnologías en el análisis realizado.

Uno de los principales puntos que deberían ser trabajados es la calidad de los datos trabajados. Esto producto que en algunos experimentos se visualizó presencia de ruido en los primeros años de registro y una baja cantidad de datos para registros actuales. Por estas razones, en próximas investigaciones se debería considerar la realización de *Web-Scrapping* más robustos, capaces de mitigar el ruido que poseen noticias los primeros años de registro para cada país. Por otro lado, se deberá aumentar los registros de noticias, de tal forma de poseer datos más actualizados y observar los potenciales efectos del COVID-19 en los emprendimientos.

En segundo lugar, deberían aplicarse técnicas de procesamiento del lenguaje más modernas para la generación de *Word Embeddings*. Una buena aproximación seria la utilización de BERT, el cual es un modelo de representación del lenguaje con el que se podría absorber de forma más robusta el contexto del texto analizado en este trabajo.

Finalmente reuniendo el conocimiento generado en este trabajo se podría utilizar para la generación de un sistema inteligente capaz de generar estimaciones de periodos en los que son más convenientes la generación de emprendimientos.

Bibliografía

[1] Mykhailyuk, O.Yu & Pohlod, H.Ya. (2015). The Languages We Speak Affect Our Perceptions of the World. Journal of Vasyl Stefanyk Precarpathian National University. 2. 10.15330/jpnu.2.2-3.36-41.

[2] Simundic, Ana-Maria. (2013). Bias in research. Biochemia Medica. 23. 12-15. 10.11613/BM.2013.003.

[3] U.S. Small Business Administration. "2020 Small Business Profile." [Accessed: 23- May- 2021].

[4] Shinnar, Rachel & Giacomin, Olivier & Janssen, Frank. (2012). Entrepreneurial Perceptions and Intentions: The Role of Gender and Culture. Entrepreneurship Theory and Practice. 36. 465-494. 10.1111/j.1540-6520.2012.00509.x.

[5] Reis, Julio & Benevenuto, Fabrício & Vaz de Melo, Pedro & Prates, Raquel & Kwak, Haewoon & An, Jisun. (2015). Breaking the News: First Impressions Matter on Online News.

[6] Hopkins, Daniel & Kim, Eunji & Kim, Soojong. (2017). Does newspaper coverage influence or reflect public perceptions of the economy?. Research & Politics. 4. 205316801773790. 10.1177/2053168017737900.

[7] Renko, Maija & Shrader, Rodney & Simon, Mark. (2012). Perception of entrepreneurial opportunity: A general framework. Management Decision. 50. 1233-1251. 10.1108/00251741211246987.

[8] Nigam, K., Mccallum, A.K., Thrun, S. et al. "Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning" 39, 103–134 (2000). https://doi.org/10.1023/A:1007692713085 [Accessed: 24- May- 2021].

[9] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1301.3781. [Accessed: 24- May- 2021].

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In Advances in neural information processing systems, 3111–3119, 2013.

[11] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors with Subword Information", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1607.04606. [Accessed: 24- May- 2021].

[12] Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv.org, 2021. [Online]. Available:

https://arxiv.org/abs/1810.04805. [Accessed: 24- May- 2021].

[13] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1907.11692. [Accessed: 24- May- 2021].

[14] Hamilton, William & Leskovec, Jure & Jurafsky, Dan. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. 1489-1501. 10.18653/v1/P16-1141.

[15] Garg, Nikhil & Schiebinger, Londa & Jurafsky, Dan & Zou, James. (2017). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. Proceedings of the National Academy of Sciences. 115. 10.1073/pnas.1720347115.

[16] Caliskan, Aylin & Bryson, Joanna & Narayanan, Arvind. (2017). Semantics derived automatically from language corpora contain human-like biases. Science. 356. 183-186.

[17] Kolchyna, Olga & Souza, Thársis & Treleaven, Philip & Aste, Tomaso. (2015). Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination.

[18] Liu, Monica & Chen, Jiun-Hung. (2015). A multi-label classification based approach for sentiment classification. Expert Systems with Applications. 42. 1083–1093. 10.1016/j.eswa.2014.08.036.

[19] Taj, Soonh & Meghji, Areej & Shaikh, Baby. (2019). Sentiment Analysis of News Articles: A Lexicon based Approach. 10.1109/ICOMET.2019.8673428.

[20] Jatowt, Adam & Duh, Kevin. (2014). A framework for analyzing semantic change of words across time. 229-238. 10.1109/JCDL.2014.6970173.

[21] P. Badilla, F. Bravo-Marquez, and J. Pérez WEFE: The Word Embeddings Fairness Evaluation Framework In Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI 2020), Yokohama, Japan.

[22] McCulloch, W.S., Pitts, W. "A logical calculus of the ideas immanent in nervous activity". Bulletin of Mathematical Biophysics 5, 115–133, 1943. Available: https://doi.org/10.1007/BF02478259

[23] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. "Learning representations by back-propagating errors". Nature, 323, 533–536, 1986. Available http://www.cs.toronto.edu/ hinton/absps/naturebp.pdf

[24] . Goldberg and O. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1402.3722. [Accessed: 31- May- 2021].

[25] McInnes, L., & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv, 2018, Available: https://arxiv.org/pdf/1802.03426.pdf.

[26] Steven Bird, Ewan Klein, and Edward Loper. (2009). Natural language processing with Python. O'Reilly Media, Inc.

[27] Chris Manning and Hinrich Schutze. (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

[28] Yoav Goldberg, Neural Network Methods in Natural Language Processing , Morgan & Claypool, 2017.

[29] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word-emotion association lexicon.” Computational Intelligence, vol. 29, no. 3, pp. 436– 465, 2013.

[30] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews.". Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems 26, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., (2013), pp. 3111–3119.

[32] “The Role of Gender Stereotypes in Perceptions of Entrepreneurs and Intentions to Become an Entrepreneur - Vishal K. Gupta, Daniel B. Turban, S. Arzu Wasti, Arijit Sikdar, 2009”, SAGE Journals, 2021. [Online]. Available: https://journals.sagepub.com/doi/10.1111/j.1540-6520.2009.00296.x. [Accessed: 20- May- 2021].

[33] “Research on Women Entrepreneurs: Challenges to (and from) the Broader Entrepreneurship Literature? | Academy of Management Annals”, Journals.aom.org, 2021. [Online]. Available: https://journals.aom.org/doi/10.5465/19416520.2013.782190. [Accessed: 21- May- 2021].

[34] “Inside the Black Box of Organizational Life: The Gendered Language of Performance Assessment - Shelley J. Correll, Katherine R. Weisshaar, Alison T. Wynn, JoAnne Delfino Wehner, 2020”, SAGE Journals, 2021. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0003122420962080. [Accessed: 25- May- 2021].

[35] Stinchcombe and White, “Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions,” International 1989 Joint Conference on Neural Networks, 1989, pp. 613-617 vol.1, doi: 10.1109/IJCNN.1989.118640.

[36] Zellig S. Harris “Distributional Structure, WORD”, 10:2-3, 146-162, 1954. Available: 10.1080/00437956.1954.11659520

[37] Yoav Goldberg; Graeme Hirst, Neural Network Methods in Natural Language Processing, Morgan & Claypool, 2017.

[38] Stanford, Normalization (equivalence classing of terms), 07-Apr-2009. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/normalization-equivalence-classing-of-terms-1.html.

[39] BBC News Mundo. 2021. Las 14 recesiones de los últimos 150 años (y por qué la del coronavirus sería la cuarta peor) - BBC News Mundo. [online] Available at: <https://www.bbc.com/mundo/noticias-53303499> [Accessed 24 December 2021].

Anexos

Anexo A: Grupos Sociales Utilizados

A continuación, son señalados las palabras que conforman a los diferentes grupos sociales propuestos en este trabajo:

Tabla 5.1: Términos para los grupos sociales utilizados en los experimentos.

Grupo	Palabras
islamic_words	['allah', 'ramadan', 'turban', 'emir', 'salaam', 'sunni', 'koran', 'imam', 'sultan', 'prophet', 'veil', 'ayatollah', 'shiite','mosque', 'islam', 'sheik', 'muslim', 'muhammad']
christianity_words	['baptism', 'messiah', 'catholicism', 'resurrection', 'christianity', 'salvation', 'protestant', 'gospel', 'trinity','jesus', 'christ', 'christian', 'cross', 'catholic', 'church']
female_words	['she', 'daughter', 'her', 'mother', 'woman', 'girl', 'herself', 'female', 'sister', 'women', 'femen','aunt', 'niece']
male_words	['he', 'son', 'his', 'him', 'father', 'man', 'boy', 'himself', 'male', 'brother', 'nephew']
inmigrant_words	['immigration','foreigner','outsider','stranger','alien','immigrant','foreign']
citizen_words	['citizenship','citizen','domestic','native','insider','local','resident','Familiar']
asian_last_names	['cho', 'wong', 'tang', 'huang', 'chu', 'chung', 'ng', 'wu', 'liu', 'chen', 'lin', 'yang', 'kim', 'chang', 'shah', 'wang', 'li', 'khan', 'singh', 'hong']
white_last_names	['harris', 'nelson', 'robinson', 'thompson', 'moore', 'wright', 'anderson', 'clark', 'jackson', 'taylor', 'scott', 'davis','allen', 'adams', 'lewis', 'williams', 'jones', 'wilson', 'martin', 'johnson']
chinese_last_names	['chung', 'liu', 'wong', 'huang', 'ng', 'hu', 'chu', 'chen', 'lin', 'liang', 'wang', 'wu', 'yang', 'tang', 'chang', 'hong', 'li']
hispanic_last_names	['ruiz', 'alvarez', 'vargas', 'castillo', 'gomez', 'soto', 'gonzalez', 'sanchez', 'rivera', 'mendoza', 'martinez', 'torres', 'rodriguez', 'perez', 'lopez', 'medina', 'diaz', 'garcia', 'castro', 'cruz']
russian_last_names	['gurin', 'minsky', 'sokolov', 'markov', 'maslow', 'novikoff', 'mishkin', 'smirnov', 'orloff', 'ivanov', 'sokoloff', 'davidoff', 'savin', 'romanoff', 'babinski', 'sorokin', 'levin', 'pavlov', 'rodin', 'agin']

Anexo B: Resultados Complementarios

A continuación son expuestas las palabras de las palabras mas similares a *entrepreneurship* cada cinco años para cada uno de los países. Al comienzo de cada año se encuentra la palabra desambiaguada y abajo de ellas el conjunto de a lo mas 10 palabras mas similares en ese periodo:

Tabla 5.2: Top 10 de las palabras mas similares de Estados Unidos a lo largo de los registros.

words	Similitud de Coseno
1980_USA_entrepreneurship	
finaliststhe	0.510089
allagash	0.506752
ddf	0.499119
talawa	0.498963
gemological	0.494177
synetic	0.488017
belier	0.488011
partech	0.478147
seona	0.477058
hakko	0.474960

1985_USA_entrepreneurship	
risk_taking	0.714487
risktaking	0.704449
self_reliance	0.664213
innovativeness	0.646363
fostering	0.629115
wealth_creation	0.627289
dynamism	0.620198
1990_USA_entrepreneurship	
self_reliance	0.654653
risk_taking	0.650787
individualism	0.609477
capitalism	0.608732
fostering	0.608230
wealth_creation	0.606367
enterprise	0.580802
conformity	0.579468
capitalist	0.577233
collectivism	0.575504
1995_USA_entrepreneurship	
self_reliance	0.676844
fostering	0.615587
experiential_learning	0.604141
creativity	0.587163
critical_thinking	0.583547
risk_taking	0.581351
pedagogy	0.581268
volunteerism	0.580111
dynamism	0.579816
wealth_creation	0.579375
2000_USA_entrepreneurship	
intrapreneurship	0.656244
innovation	0.640912
pedagogy	0.638947
political_activism	0.628883
experiential_learning	0.627902
volunteerism	0.627718
philanthropy	0.623204
technological_innovation	0.621500
wealth_creation	0.619303
lifelong_learning	0.619295

2005_USA_entrepreneurship	
financial_literacy	0.697727
centennial_college	0.670166
mentoring_program	0.665560
lifelong_learning	0.656629
volunteerism	0.655456
outreach_program	0.630535
environmental_sustainability	0.630338
experiential_learning	0.625650
philanthropy	0.619589
social_enterprise	0.618088
2010_USA_entrepreneurship	
lifelong_learning	0.714886
financial_literacy	0.705257
nfte	0.676224
innovation	0.665720
experiential_learning	0.644975
education	0.639061
pedagogy	0.635365
volunteerism	0.634153
environmental_sustainability	0.631932
unisa	0.631373
2015_USA_entrepreneurship	
lifelong_learning	0.687743
experiential_learning	0.670823
innovation	0.659713
environmental_sustainability	0.650328
sustainability	0.623600
ashoka	0.623285
philanthropy	0.620524
volunteerism	0.614151
fostering	0.612240
civic_engagement	0.608486
2019_USA_entrepreneurship	
lifelong_learning	0.705233
innovation	0.666457
fostering	0.661017
intrapreneurship	0.652214
wealth_creation	0.650335
employability	0.646357
volunteerism	0.644581

experiential_learning	0.644357
self_reliance	0.640975
financial_literacy	0.637624

Tabla 5.3: Top 10 de las palabras mas similares de Nueva Zelanda a lo largo de los registros.

words	Similitud de Coseno
2000_NZ_entrepreneurship	
excellence	0.620451
pedagogy	0.590702
pedagogical	0.584217
lifelong_learning	0.569818
experiential_learning	0.562875
innovation	0.559786
creativity	0.549085
unitec	0.546680
innovativeness	0.546553
critical_thinking	0.540638
2005_NZ_entrepreneurship	
unisa	0.672875
centennial_college	0.672682
qut	0.661655
innovationand	0.660726
unitec	0.660640
uvic	0.658056
nzte	0.656560
innovation	0.644856
dcu	0.643714
massey_university	0.636707
2010_NZ_entrepreneurship	
environmental_sustainability	0.628873
enterpreneurship	0.627324
civic_engagement	0.584405
awardthe	0.569028
excellence	0.567131
innovationand	0.562839
advancement	0.551032
atamira	0.544424
lifelong_learning	0.539000
2015_NZ_entrepreneurship	

innovation_ecosystem	0.612008
innovation	0.595785
environmental_sustainability	0.580581
sustainability	0.580402
innovator	0.579129
ashoka	0.568841
innovation_hub	0.559682
civic_engagement	0.554511
echallenge	0.550031
lifelong_learning	0.548831
2019_NZ_entrepreneurship	
iadt	0.590319
floristry	0.563343
graduate_certificate	0.549434
bsc	0.546562
organizational_behavior	0.543972
econ	0.542970
graduate_diploma	0.539140
qut	0.537809
ucd	0.537054
aut	0.535575

Tabla 5.4: Top 10 de las palabras mas similares de Reino Unido a lo largo de los registros.

words	Similitud de Coseno
1985_UK_entrepreneurship	
connexity	0.534004
potentiality	0.531047
cfe	0.530862
solecism	0.526629
scientism	0.512859
womxn	0.509613
compartmentalisation	0.506589
opportu	0.505578
losartan	0.505088
historiography	0.502617
1990_UK_entrepreneurship	
careerism	0.485442
supplyside	0.479727
phumzile	0.475582
tvei	0.473093

behavioral_science	0.469185
humanitarianism	0.465495
wealthcreation	0.464398
innovationthe	0.462193
critical_thinking	0.461481
institutionalising	0.460866
1995_UK_entrepreneurship	
swadeshi	0.548976
biohackers	0.546089
ferenc	0.544870
unclassifiable	0.544694
baedeker	0.539263
ouroussoff	0.538829
kounkuey	0.537035
longestablished	0.535927
rebarbative	0.530967
enterpise	0.529594
2000_UK_entrepreneurship	
innovation	0.721284
volunteerism	0.700816
fostering	0.661807
lifelong_learning	0.649746
philanthropy	0.640935
intrapreneurship	0.636185
creativity	0.635656
environmental_sustainability	0.625093
interdisciplinary	0.624720
wealth_creation	0.621377
2005_UK_entrepreneurship	
lifelong_learning	0.701156
excellence	0.679286
innovation	0.674894
environmental_sustainability	0.669723
advancement	0.662882
behavioral_science	0.661416
education	0.659690
volunteerism	0.656685
tertiary_education	0.654366
philanthropy	0.639489
2010_UK_entrepreneurship	
civic_engagement	0.707893

volunteerism	0.702263
innovation	0.695769
fostering	0.690808
lifelong_learning	0.672350
creativity	0.639767
intrapreneurship	0.638823
excellence	0.637213
self_reliance	0.635220
wealth_creation	0.626695
2015_UK_entrepreneurship	
innovation	0.723879
lifelong_learning	0.719843
fostering	0.705258
experiential_learning	0.676016
intrapreneurship	0.671718
social_enterprise	0.670949
volunteerism	0.670158
civic_engagement	0.666191
environmental_sustainability	0.652507
poverty_reduction	0.647212
2019_UK_entrepreneurship	
innovation	0.667545
fostering	0.654541
lifelong_learning	0.629283
civic_engagement	0.622392
intrapreneurship	0.616461
advancement	0.606089
excellence	0.590389
volunteerism	0.589115
environmental_sustainability	0.582734
commercialization	0.581769

Tabla 5.5: Top 10 de las palabras mas similares de Irlanda a lo largo de los registros.

words	Similitud de Coseno
1995_IRL_entrepreneurship	
skytek	0.621638
truncates	0.596364
tonti	0.594500
crocfest	0.594414
wealthcreation	0.594069

edisonian	0.592701
szeemann	0.589332
tiveness	0.588375
sbusiness	0.587569
korero	0.585264
2000_IRL_entrepreneurship	
bcic	0.605780
startuphouse	0.601227
aald	0.601158
preneurship	0.588529
nch	0.584046
dih	0.582837
eile	0.579137
shahrzad	0.576550
capitalvision	0.574267
testtown	0.573467
2005_IRL_entrepreneurship	
volunteerism	0.606229
voluntarism	0.605360
philanthropy	0.577740
enterpreneurship	0.573704
fostering	0.562322
advancement	0.543439
civic_engagement	0.539937
ashoka	0.536762
kauffman_foundation	0.521672
aspen_institute	0.516145
2010_IRL_entrepreneurship	
innovation	0.696667
vation	0.657792
fostering	0.647833
civic_engagement	0.636274
volunteerism	0.630289
lifelong_learning	0.628576
excellence	0.613804
intrapreneurship	0.606626
risk_taking	0.604977
social_enterprise	0.603701
2015_IRL_entrepreneurship	
innovation	0.732013
lifelong_learning	0.716511

fostering	0.703210
experiential_learning	0.697941
intrapreneurship	0.667063
employability	0.661717
environmental_sustainability	0.658970
volunteerism	0.650670
critical_thinking	0.646529
education	0.640549
2019_IRL_entrepreneurship	
innovation	0.681073
lifelong_learning	0.663590
ict	0.663479
environmental_sustainability	0.629902
intrapreneurship	0.627715
volunteerism	0.627282
fostering	0.623036
inclusion	0.619425
innov	0.618779
financial_literacy	0.608496

Tabla 5.6: Top 10 de las palabras mas similares de Canadá a lo largo de los registros.

words	Similitud de Coseno
1985_CA_entrepreneurship	
risk_taking	0.543389
risktaking	0.535262
permissionless	0.534038
innovationand	0.530286
enterpreneurial	0.514795
uw	0.507220
cranfield	0.491797
cariocca	0.490303
dartington	0.488870
universitynow	0.485878
1990_CA_entrepreneurship	
volunteerism	0.715057
lifelong_learning	0.659610
advancement	0.638344
financial_literacy	0.625011
experiential_learning	0.620480

1995_CA_entrepreneurship	
lifelong_learning	0.632986
excellence	0.627398
volunteerism	0.622569
self_reliance	0.615168
pedagogy	0.600071
innovation	0.594343
experiential_learning	0.589409
tertiary_education	0.588407
vocational_education	0.580555
intrapreneurship	0.577721
2000_CA_entrepreneurship	
innovation	0.745477
fostering	0.685704
internationalization	0.652298
wealth_creation	0.640214
2005_CA_entrepreneurship	
philanthropy	0.615785
experiential_learning	0.598362
catalysing	0.586829
internationalization	0.584551
advancement	0.575554
civic_engagement	0.570706
wealth_creation	0.568979
financial_literacy	0.568583
fostering	0.561295
volunteerism	0.557980
2010_CA_entrepreneurship	
wealth_creation	0.652798
innovation	0.645381
fostering	0.641295
volunteerism	0.635147
intrapreneurship	0.627225
self_reliance	0.626497
experiential_learning	0.618966
critical_thinking	0.610426
philanthropy	0.602367
lifelong_learning	0.602070
2015_CA_entrepreneurship	
innovation	0.689756

unisa	0.635198
lifelong_learning	0.633100
qut	0.631572
dcu	0.618710
behavioral_science	0.607187
critical_thinking	0.607072
informatics	0.606101
pedagogy	0.604316
uvic	0.601522
2019_CA_entrepreneurship	
civic_engagement	0.677542
lifelong_learning	0.668340
gender_equality	0.668307
environmental_sustainability	0.661993
fostering	0.659996
social_enterprise	0.649362
innovation	0.647382
ashoka	0.636977
sustainability	0.629059
outreach	0.623801

Tabla 5.7: Top 10 de las palabras mas similares de Australia a lo largo de los registros.

words	Similitud de Coseno
1990_AU_entrepreneurship	
leadership	0.523048
statesmanship	0.500268
voluntarism	0.499070
universalism	0.498684
mindedness	0.498101
innovation	0.492472
dynamism	0.490069
experimentation	0.489667
excellence	0.489463
ethic	0.486301
1995_AU_entrepreneurship	
creativity	0.655237
self_reliance	0.642218
inventiveness	0.621060
critical_thinking	0.615298
innovation	0.612885

volunteerism	0.612595
risktaking	0.609510
risk_taking	0.607756
innovativeness	0.606194
excellence	0.598469
2000_AU_entrepreneurship	
critical_thinking	0.757379
pedagogy	0.727119
problem_solving	0.712058
innovation	0.702209
experiential_learning	0.689999
lifelong_learning	0.684321
behavioral_science	0.671513
specialization	0.662478
financial_literacy	0.660251
internationalization	0.656710
2005_AU_entrepreneurship	
lifelong_learning	0.689780
innovation	0.681127
unisa	0.677618
critical_thinking	0.677192
experiential_learning	0.674021
swinburne_university	0.669523
qut	0.663429
swinburne	0.663373
pedagogy	0.662910
centennial_college	0.660358
2010_AU_entrepreneurship	
innovation	0.732006
creativity	0.725937
risk_taking	0.678340
self_reliance	0.676773
fostering	0.659034
innovativeness	0.647364
volunteerism	0.641830
wealth_creation	0.637990
technological_innovation	0.637426
environmental_sustainability	0.633562
2015_AU_entrepreneurship	
innovation	0.816401
lifelong_learning	0.765363

critical_thinking	0.736966
experiential_learning	0.733783
creativity	0.711144
employability	0.709671
intrapreneurship	0.701649
2019_AU_entrepreneurship	
innovation	0.761641
fostering	0.720716
lifelong_learning	0.716083
experiential_learning	0.697807
commercialization	0.663640
employability	0.662133
advancement	0.655566
volunteerism	0.648213