UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

TRABAJO FIN DE MASTER

Detección de autoría mediante word embeddings estáticos

Este trabajo presentado cumple con los requisitos del Máster Universitario en Tecnologías del lenguaje

en

ETS de Ingeniería Informática

Autor:

Bosco Gutiérrez Gómez

Dirigido por:

Dr. Víctor Fresno Fernández

Dr. Álvaro Rodrigo Yuste

Febrero de 2024

Madrid

Resumen

Vivimos en la era de la información donde la conectividad, la tecnología y las redes sociales han ido moldeando la forma en que nos relacionamos con la información y el uso que hacemos de ella. Uno de los mayores peligros y dificultades es poder discernir la veracidad del contenido, así como la fuente de la información. En noviembre de 2022, Open AI lanza la aplicación oficial ChatGPT, un modelo de lenguaje generado por inteligencia artificial (IA) generativa, que es el detonante del uso de esta tecnología a nivel general, consiguiendo, según Forbes, más de 100 millones de usuarios en menos de de 2 meses. Esta tecnología abre multitud de posibilidades como mejoras de productividad y calidad, automatización de tareas rutinarias y generación de contenidos de texto, audio o imágenes. Su adopción masiva en ámbitos como el trabajo o la educación y sus capacidades creativas ha generado controversia y mucho debate sobre los peligros de su uso. Algunos de estos miedos entran dentro del ámbito de la educación y del uso que los estudiantes e investigadores pueden hacer de esta herramienta a la hora de generar contenidos para sus trabajos, artículos,. . . El correcto funcionamiento de estos programas hace que sea casi imposible distinguirlos del resultado producido por un ser humano, al menos sin el uso de ninguna herramienta de detección. Este ya no es un problema de detección de plagio para el que existían programas que localizaban el origen de los extractos de texto y sus autores, sino que se incluye dentro de la apropiación de la autoría, entendiendo al autor como la IA generativa que crea el contenido. Nosotros nos planteamos estudiar cómo la representación distribuida en un espacio vectorial nos permite caracterizar un tipo de texto en función de la amplitud y variedad de vocabulario y su especificidad. Estas características deberían ser particulares de cada autor, lo cual nos lleva a profundizar en este planteamiento por el que una IA generativa también podría tener su propio perfil y distinguirse de otros autores. Esta metodología no pretende ser una herramienta que caracterice un texto de manera inequívoca por sí misma, pero si que complemente desde una nueva perspectiva a las distintas técnicas ya existentes para dotarlas de una mayor fiabilidad y precisión. Este trabajo se basa en la detección de autoría en textos mediante el estudio de la orientación semántica de las palabras que lo conforman.

Abstract

We live in the information age where connectivity, technology and social media have been shaping how we relate to information and how we use it. One of the greatest dangers and difficulties is to be able to discern the veracity of the content as well as the source of the information. In November 2022, Open AI launched the official ChatGPT application, a language model generated by generative artificial intelligence, which is the trigger for the use of this technology at a general level, achieving, according to Forbes, more than 100 million users in less than 2 months. This technology opens up many possibilities such as productivity and quality improvements, automation of routine tasks, and generation of text, audio, or image content. Its massive adoption in areas such as work or education and its creative capabilities such as generating text, audio or image content, it has generated controversy and much debate about the dangers of its use. Some of these fears are framed in the educational or academic field and in the use that students and researchers can make of this tool when generating content for their works and articles. The correct functioning of these programs makes it almost impossible to distinguish them from the result produced by a human being, at least without the use of any detection tools. This is no longer just a problem of plagiarism detection, but also of impersonation, such as generative AI that creates the content. The objective of authorship attribution is the identification and assignment of documents to a specific author or authors based on their syntactic, morphological, lexical and semantic style. Its application is very diverse such as plagiarism detection, authorship verification, author profiling and prevention of identity theft. We propose studying how the representation distributed in a vector space allows us to characterize a type of text according to the amplitude and variety of vocabulary and its specificity. These characteristics should be specific to each author, which leads us to deepen this approach by which a generative AI could also have its profile and distinguish itself from other authors. This methodology is not intended to be a tool that characterizes a text unequivocally by itself, but rather complements, from a new perspective, the various existing techniques to provide them with greater reliability and precision. This work is based on the detection of authorship in texts by studying the semantic orientation of the words that comprise it.

Agradecimientos

Me gustaría dedicar este trabajo a mi familia, Marina, Adriana y Belén, y agradecer el cederme nuestro tiempo en familia para poder dedicarlo a este máster.

También le agradezco a Álvaro y a Víctor, mis tutores, por la exigencia y dedicación en la supervisión de este trabajo.

Índice

Índice de figuras

1. Introducción

1.1. Motivación

Vivimos en la era de la información donde la conectividad, internet, la tecnología y las redes sociales han ido modelando la forma en que nos relacionamos con la información y qué uso hacemos de ella. Uno de los mayores peligros y dificultad es poder discernir la veracidad del contenido, así como la fuente de dicha información.

En noviembre de 2022 Open AI lanza la aplicación oficial de ChatGPT[1], un modelo de lenguaje generado por inteligencia artificial generativa, lo que supone el detonante del uso de esta tecnología a nivel general, consiguiendo, según Forbes[2], mas de 100 millones de usuarios en menos de 2 meses. Esta tecnología abre multitud de posibilidades como las mejoras de productividad y calidad, automatización de tareas rutinarias y en la generación de contenido de texto, audio o imágenes.

La adopción masiva en ámbitos como el laboral o educativo y sus capacidades creativas ha generado controversias y mucho debate sobre los peligros de su uso. Algunos de estos miedos se enmarcan dentro del ámbito de la educación y el uso que pueden hacer de esta herramienta los estudiantes e investigadores en la generación de contenido para sus trabajos, artículos,... El buen desempeño de estos programas hace casi imposible distinguirlo a simple vista del resultado producido por un ser humano [Sheinman et al (2023)], al menos sin la utilización de alguna herramienta de detección como Copyleaks[3], Originality.ai[4] o Flint[5].

Esto ya no es solo un problema de detección de plagio, para lo cual existían programas que localizaban la procedencia de extractos de textos y sus autores como Turnitin[6], sino que se engloba dentro de la suplantación de autoría. Es decir, la imitación de la forma e la que escribe un autor o el simple hecho de utilizar una herramienta que

[1] https://chat.openai.com/

[2] https://www.forbes.com/sites/martineparis/2023/02/03/chatgpt-hits-100-million-microsoft-unleashes-ai-bots-and-catgpt-goes-viral/

[3] https://copyleaks.com/

[4] https://originality.ai/

[5] https://flintai.com/

[6] https://turnitin.com/

genere texto a través de la IA, vulnera los códigos deontológicos académicos al apropiarse un sujeto del trabajo de investigación y creación del texto. No entra dentro del ámbito de este trabajo juzgar ni decidir qué porcentaje del uso de estas herramientas sería el apropiado o ético.

Un método lógico para validar la autoría de un texto sería i.ntentar demostrar que todos los escritores dejarían una "huella" estilística y lingüística única en sus textos y es una línea de investigación abierta por Halteren [Halteren et al (2005)] que propone que si la gramática se estabiliza tempranamente, el vocabulario es el único lugar donde el idioma puede adaptarse a los cambios en las condiciones bajo las cuales usamos el idioma. Esto implica que los componentes más útiles de un conjunto de rasgos medibles de los productos lingüísticos se referirían al uso de palabras, mientras que el uso de patrones sintácticos debería ser mucho menos distintivo.

Determinar si la autoría de un texto pertenece a quién se la atribuye es una tarea bastante compleja debido a las posibilidades de autores conocidos o incluso desconocidos. La complejidad de determinar el autor real se incrementa cuando se le añade la dificultad de la validación si el autor es una IA generativa. Por todo esto, parece lógico preguntarnos si es posible caracterizar un texto o tipo de escritura y cómo deberíamos hacerlo tanto si está creada por un ser humano como si lo ha hecho una IA generativa.

Respecto a la caracterización de textos y autores, se han propuesto diferentes iniciativas para su estudio, entre las que debemos mencionar Profiling Authorship in Natural Language Processing and Computational Linguistics[7] (PAN), que son una serie de eventos científicos y tareas compartidas sobre estilometría y análisis de textos digitales organizados anualmente desde 2007 para la detección de plagio y otras tareas de análisis de texto. Suele modelarse como una tarea de aprendizaje supervisado, haciendo uso de corpus de texto etiquetados. Algunos de los métodos mas populares incluyen el uso de clasificadores de máquinas de vectores de soporte (SVM) [Posadas et al (2015)], redes neuronales recurrentes, redes neuronales convolucionales [Zhang et al (2018)] y así como la combinación de distintas técnicas de

[7] https://pan.webis.de/

aprendizaje automático. La efectividad y precisión alcanzada en las distintas tareas a lo largo de los años ha ido incrementándose hasta alcanzar resultados muy fiables, especialmente desde el enfoque sintáctico y estilométrico, aunque encuentra más dificultad cuando el enfoque se centra en la semántica y los significados de las palabras.

La representación del significado de las palabras es una tarea fundamental en el procesamiento del lenguaje natural (PLN). Para la representación de estas palabras a través de vectores utilizamos los word embeddings, que son representaciones numéricas abstractas de las relaciones entre palabras donde conceptos con semántica similar tienen representaciones similares. Con esto nos referimos a la similitud léxico-semántica de los embeddings como la medida en que los vectores de palabras o frases en el espacio vectorial reflejan las similitudes y relaciones semánticas entre ellas, por lo que deberían tener vectores cercanos en ese espacio vectorial.

Un espacio de embeddings es isotrópico si el las direcciones se distribuyen uniformemente. La representación contextual de palabras es generalmente anisotrópico [Xu and Koehn (2021)]. Según el término anisotropía [Liang et al (2021)], el significado de los vectores de palabras se distribuye según una orientación particular en el espacio vectorial. Esta distribución se puede representar geométricamente como puntos en un espacio vectorial continuo, donde a partir de su posicionamiento podemos extraer características específicas de cada conjunto de palabras.

Definimos dos espacios como isométricos si las distancias euclídianas relativas entre vectores son idénticas entre espacios. En el espacio vectorial de representación de word embeddings contextuales, asumimos la isometría semántica referida a la capacidad de preservar relaciones semánticas entre diferentes espacios vectoriales de los embeddings basada en la proximidad según la afinidad de su significado. Los embeddings estáticos (Static Word Embeddings) permiten mantener las propiedades semánticas del significado de las palabras que representan.

Se ha demostrado en los word embeddings que la norma representa la importancia relativa de la palabra, mientras que la dirección representa el significado de la palabra [Oyama et al (2023)]. La norma de un vector de la palabra codifica implícitamente el peso de la importancia de una palabra y que el ángulo entre vectores de la palabra es un buen indicador para la similitud de palabras [Yokoi et al (2020)].

Nosotros nos planteamos estudiar cómo la representación distribuida en un espacio vectorial nos permite caracterizar un tipo de texto en función de la amplitud y variedad de vocabulario y su especificidad. Estas características deberían ser particulares de cada autor, lo cual nos lleva a profundizar en este planteamiento por el que una IA generativa también podría tener su propio perfil y distinguirse de otros autores. Para el estudio nos basaremos en la comparación de textos según la norma y la dirección de los vectores de las palabras. Este trabajo se basa en la detección de autoría en textos mediante el estudio de la orientación semántica de las palabras que lo conforman.

1.2. Propuesta y objectivos

Los principales objetivos de este trabajo son, por un lado, estudiar la aplicación de representaciones semántico-léxicas basadas en embeddings estáticos a la detección de autoría comparando distintos textos y, por otro lado, demostrar que tanto un texto creado por un ser humano como otro generado por una inteligencia artificial tienen un perfil definido que nos permite extraer características especificas para su comparación y ayuda a determinar su autoría.

Para delimitar el alcance del estudio planteamos la siguiente hipótesis: La forma en que se distribuyen los vectores que representan las palabras de un texto en un espacio vectorial, así como la longitud y orientación de estos vectores, son característicos de cada autor en particular, ya sea un ser humano o una inteligencia artificial, ya que están definidos por la riqueza del vocabulario que utilizan.

Un vector se puede representar geométricamente como puntos en un espacio vectorial continuo y el tamaño del vector es proporcional a la especificidad semántica de la palabra. La orientación de estos vectores representados nos permite conocer cómo de similares son las palabras en función de su significado.

Las preguntas de investigación que se plantean en este trabajo son:

- ¿Se pueden extraer características específicas de un texto basado en la orientación semántica y en la especificidad del vocabulario dentro de espacios de representación de embeddings estáticos?

- ¿Es la riqueza del vocabulario utilizado y su significado lo que determina la caracterización de un texto?

- ¿Los textos generados por inteligencia artificial generativa tienen un perfil determinado que les caracteriza?

- ¿Se pueden distinguir los textos creados por inteligencia artificial generativa de los creados por seres humanos según la caracterización de cada texto en base a la especificidad y similitud?

Para la consecución de los objetivos vamos a realizar tareas específicas utilizando las técnicas existentes y necesarias para el desarrollo de un trabajo de fin de master (TFM) como son:

- Crear una colección suficientemente significativa con fragmentos de textos de autores seleccionados, así como textos creados por inteligencia artificial generativa, que tratan diferentes tématicas y con distintos estilos.

- Estudiar y comparar la similitud y la amplitud de la orientación semántica del vocabulario de los textos de un determinado autor con otro.

- Estudiar y comparar la especificidad de las palabras individuales y con mayor información de los textos de un determinado autor con otro.

- Analizar los resultados obtenidos, comparar los distintos enfoques empleados y redactar las conclusiones.

- Preparación para la investigación futura proporcionando una base sólida para un proyecto de tesis de doctorado.

Esta metodología no pretende ser una herramienta que caracterice un texto de manera inequívoca por sí misma, pero si que complemente desde una nueva perspectiva a las distintas técnicas ya existentes para dotarlas de una mayor fiabilidad y precisión.

1.3. Estructura del documento

El trabajo se estructura de la siguiente forma:

Capítulo 2. Estado del arte. Repasamos las investigaciones que son referencia dentro de este ámbito y aquellas que aportan una nueva visión en aspectos concretos del estudio. Explicaremos la evolución de algunos métodos utilizados para la atribución de autoría y la nueva problemática generada por la inteligencia artificial generativa. Nos centraremos especialmente en los embeddings estáticos como enfoque principal de este estudio.

Capítulo 3. Fundamentos teóricos. Explicaremos los fundamentos que dan pie a esta investigación y que permiten delimitar el alcance de este estudio.

Capítulo 4. Análisis de datos. En este capítulo describimos la creación de la base de datos y los fundamentos utilizados para la elección de los textos y los autores. También explicamos los criterios y el proceso de limpieza y preparación de los datos.

Capítulo 5. Evaluación. Para esta sección comenzaremos describiendo las métricas utilizadas para la evaluación. Comentaremos el diseño experimental, desde la metodología utilizada hasta los resultados obtenidos en la fase de experimentación.

Capítulo 6. Resultados y discusión. Presentaremos los resultados obtenidos en la experimentación y haremos un análisis de la comparativa realizada entre ellos.

Capítulo 7. Conclusiones y futuros desarrollos. Comentaremos los resultados y sus conclusiones, así como posibles futuros enfoques como continuación de este trabajo.

2. Estado del Arte

El objetivo de la atribución de autoría es la identificación y asignación de documentos a un autor o autores determinados en función de su estilo sintáctico, morfológico, léxico y semántico. Su aplicación es muy diversa como la detección de plagio, verificación de autoría, perfilado del autor, evitar la suplantación de identidad,...

Para ello es necesaria la extracción de características, y tanto los métodos estadísticos, como los de contenido, obtienen un buen desempeño. Posadas [Posadas et al (2015)] utiliza un modelo de espacio vectorial para representar los tweets de un autor e introducir el uso de n-gramas sintácticos como marcadores de personalidad junto con el uso de clasificadores SVM tradicionales. Zhang [Zhang et al (2018)] construye un embedding para cada palabra de la oración, codificando la ruta en el árbol de sintaxis correspondiente a la palabra. En contenido de los textos, Sari [Sari et al (2018)] en su estudio explora cómo las características de un conjunto de datos afectan la utilidad de diferentes tipos de características (estilo, contenido e híbridos) para la tarea de atribución de autoría utilizando representaciones discretas y continuas.

En la tarea de capturar la información sintáctica y semántica de los textos encuentran una mayor dificultad [Wu et al (2021)]. A pesar de la incorporación de las redes neuronales [Zhang et al (2018)] [Jafariakinabad et al (2019)], que han mejorado la captura de información sintáctica y semántica, no han resuelto la falta de transparencia e interpretación de los resultados obtenidos, lo que deslegitima las posibles conclusiones en los estudios de textos para poder respaldar los datos obtenidos.

Mientras que el avance basado en las características formales de un texto (Número de palabras de las oraciones, n-grams, utilización de signos de puntuación, etc) ha sido mayor gracias a la objetividad de estas medidas, nos encontramos mayores obstáculos cuando nos enfocamos en la semántica, ya que muchas palabras son polisémicas o el propio contexto de la oración condiciona el significado de las palabras. Bajo esta perspectiva encontramos una mayor necesidad de investigación y desarrollo hasta obtener resultados suficientemente fiables y comprensibles y es hacia donde orientamos nuestro estudio.

2.1. Evaluaciones PAN

PAN es una serie de eventos científicos y tareas compartidas sobre estilometría y análisis de textos digitales, creado en 2007, centradas en la evaluación de autoría en el procesamiento del lenguaje natural y la lingüística computacional. Se divide en tres categorías que son:

- Análisis de autoría.

- Ética Computacional.

- Originalidad.

2.1.1. Definición de las subtareas

Dentro del análisis de autoría se han ido desarrollando tareas específicas:

- Identificación del autor (dado un documento y un conjunto de autores candidatos, determinar quién escribió el documento, PAN 2011–2012, 2016–2020).

- Ocultación de autoría (dado un conjunto de documentos del mismo autor, parafrasear uno o todos ellos para que su autor ya no pueda ser identificado, PAN 2016-2018).

- Evaluación de ocultación (diseñar e implementar medidas de desempeño que cuantifiquen la seguridad, solidez y/o sensatez de un software, PAN 2016–2018).

- Perfil del autor y análisis de múltiples autores (dado un par o colección de documentos, determinar si están escritos por el mismo autor, PAN 2013-2015, 2021).

La categoría de originalidad se centra en detección de plagio (PAN 2009-2015). No es casualidad que el reto lanzado para 2024 sea verificación de autoría de IA generativa.

2.1.2. Descripción de los distintos enfoques

Durante estos foros, los investigadores trabajan en el desarrollo de técnicas y sistemas relacionados con la identificación de autores o perfiles de autor a partir de textos escritos. Los retos propuestos de perfiles de autor PAN han tenido como objetivo explorar cómo se aplican las técnicas de perfilado de autores para la identificación de género y grupo de edad[8]. La mayoría de los estudios han utilizado métodos tradicionales de aprendizaje automático como:

- Combinación de palabras funcionales y características de Part-of-Speech (PoS) para predecir el género [Argamon et al (2003)]. En este artículo explora las posibles variaciones entre los estilos de escritura masculinos y femeninos, identificando varias clases de características léxicas y sintácticas simples cuya aparición en los textos difiere sustancialmente según el género del autor.

- Características estilísticas y basadas en contenido [Schler et al (2006)]. Considera las diferencias entre blogs masculinos y femeninos y de diferentes edades según sus características distintivas relacionadas con el estilo y con el contenido midiendo la frecuencia con la que aparece en el corpus.

- Características basadas en jerga y longitud promedio [Goswami et al (2009)] de oraciones. Se centra en las diferencias estilométricas en los blogs según la variación de género y grupo de edad basados en el uso de jerga, y la variación en la duración promedio de las sentencias entre varios grupos de edad y género y se complementan con resultados de estudios previos informados en la literatura para el análisis estilométrico de edad y género.

- Características basadas en estructura, estilometría y semántica utilizando métodos de conjunto ponderados basados en la mayoría (Naive Bayes, Random Tree y SVM) [Meina et al (2013)]. Para todos los documentos se calculan características que describen la estructura de las conversaciones, como el número de conversaciones, párrafos, oraciones, caracteres especiales y palabras por oración.

- Características estilométricas, de n-gramas y de segundo orden y aplicación de un clasificador SVM [Bougiatiotis and Krithara (2016)]. Asocia los diferentes términos de la colección con perfiles objetivo (clases de edad o género) calculando los vectores de clases de palabras según la frecuencia de las palabras.

[8]https://pan.webis.de/clef18/pan18-web/author-profiling.html

- Modelos de n-gramas para la extracción de características entrenando un SVM con la combinación de caracteres y n-gramas tf-idf [Basile et al (2017)].

- CNN para la identificación de grupos de edad [Guimaraes et al (2017)]. La principal aportación de este trabajo es demostrar que parámetros como el uso de la puntuación que incluye iconos de emociones, el número de caracteres del mensaje o la longitud de las frases, la jerga, el número de personas que sigue el usuario, el número de seguidores, el número total de tuits publicados en la red social y los temas abordados son relevantes para incrementar la asertividad y precisión en la clasificación del grupo de edad. Los pasos seguidos por la CNN en el trabajo de entrenamiento son la utilización del modelo de lenguaje neuronal word2vec, el funcionamiento de una red neuronal convolucional para refinar el corpus supervisado y al final, se utilizaron los parámetros para inicializar la red obtenida en una etapa previa al entrenamiento del corpus supervisado.

- RNN con incrustaciones de palabras basadas en skip-gram y CNN [Miura et al (2017)]. Los modelos combinan información de palabras e información de caracteres con redes neuronales complejas que consisten en una capa de red neuronal recurrente, una capa de red neuronal convolucional y una capa de mecanismo de atención para clasificar un rasgo de perfil.

- Modelo de red neuronal densa de tres capas para la tarea de identificación de género [Raiyani et al (2018)]. La idea es tratar el texto del tweet como una colección de palabras del diccionario, se realiza un preprocesamiento de los tweets, seguido de la representación del texto y la construcción del modelo de clasificación para luego utilizar la indexación con una arquitectura densa y simple.

- RNN para la clasificación de texto y CNN para imágenes y agrupación al máximo [Takahashi et al (2018)]. Se describe un modelo de red neuronal para la identificación de género en Twitter llamado Red neuronal de fusión de imágenes de texto (TIFNN). Este método consiste en extraer información de mensajes escritos e imágenes compartidas por los usuarios. Para aprovechar la sinergia de los textos y las imágenes, el modelo propuesto calcula la relación entre ellos utilizando el producto directo.

Se han explorado diversas técnicas en las distintas tareas o retos lanzados a lo largo de los años como son los algoritmos de clasificación supervisada utilizados en tareas de verificación de autoría: máquinas de vectores de soporte, árboles de decisión, análisis discriminante, redes neuronales y algoritmos genéticos [Stamatatos et al (2014)]. Otras estrategias se enfocan en calcular la similitud entre textos y utilizarla para predecir si ambos textos están escritos por el mismo autor o no [Koppel et al (2011)], como con el método de los impostores [Koppel and Winter(2014)] que propone utilizar un conjunto de documentos impostores recopilados de una fuente externa con un tema similar al documento conocido y desconocido. También se define un conjunto de características como son palabras funcionales, n-gramas de palabras y n-gramas de caracteres. El esquema de esta solución es el siguiente, producir un conjunto de documentos "impostores" y medir la similitud de los documentos en comparación con el original seleccionando aleatoriamente subconjuntos de características que sirven como base para comparar documentos. Los documentos no contienen más de 500 palabras. Para medir la similitud, con la similitud coseno, entre documentos, representa cada documento como un vector numérico que contiene las frecuencias respectivas de cada carácter de 4-grams en el documento. En este sentido, tomamos la verificación de la autoría como un problema de clasificación de conjuntos abiertos entre muchos autores posibles.

El método de desenmascaramiento, es otro método propuesto por Koppel como tarea de verificación de autoría, intenta medir qué tan profunda es la diferencia entre dos textos comparándolos varias veces [Koppel et al (2007)]. Se basa en la idea de que el estilo de los textos de un mismo autor sólo difiere en algunos rasgos superficiales. Al eliminar iterativamente estas características de estilo más discriminatorias, se puede medir la velocidad a la que se degrada la precisión de la validación cruzada entre conjuntos de fragmentos de los dos textos. Para textos escritos por el mismo autor, la precisión tiende a disminuir más rápido que en otros casos. Bevendorff et al. [Bevendorff et al (2019)] propusieron un método de desenmascaramiento alternativo que obtiene una gran precisión en textos cortos. Permite la verificación de la autoría de textos cortos con una precisión muy alta reduciendo el material requerido y haciendo que el desenmascaramiento sea aplicable a casos de autoría de proporciones más prácticas. Lo hace explotando la naturaleza de bolsa de palabras de las funciones de desenmascaramiento y creando los fragmentos de sobremuestras de palabras de manera agregada mediante bootstrap. Trata cada texto como un conjunto aleatorio de palabras del que poder extraer sin reemplazo para completar una parte. Una vez que el grupo se agota, lo repone y extrae nuevamente hasta que haya generado una cantidad suficiente. Esto es para garantizar que cada palabra se dibuje al menos una vez.

En los últimos años, se han propuesto arquitecturas de redes neuronales para resolver la tarea y aunque requieren de una mayor capacidad computacional, han obtenido mejores resultados como es el caso de Bagnall [Bagnall (2015)] en el PAN 2015 que utiliza una red neuronal recurrente (RNN) para modelar el lenguaje de varios autores al mismo tiempo, donde el texto de cada autor se representa mediante salidas separadas que dependen de un estado recurrente compartido. Esto permite que la capa recurrente modele el lenguaje como un todo sin sobreajustarlo, incluso con corpus muy pequeños.

Jafariakinabad [Jafariakinabad et al (2019)] en el PAN de 2012, introdujo una red neuronal recurrente sintáctica para codificar los patrones sintácticos de un documento en una estructura jerárquica de oraciones a partir de la secuencia de etiquetas PoS demostrando que supera al modelo léxico.

Weerasinghe y Greenstadt [Weerasinghe and Greenstadt (2020)] propusieron un modelo con un enfoque estilométrico. Extrajeron características estilométricas de cada par de texto y utilizaron las diferencias absolutas entre los vectores de características como entrada para un clasificador. Evaluaron un modelo utilizando un clasificador de regresión logística y otro utilizando un enfoque de red neuronal.

Los avances en el campo semántico dentro de las tareas PAN han sido limitados, ya que tradicionalmente se ha hecho hincapié en las características sintácticas y estilométricas. Sin embargo, el campo del procesamiento del lenguaje natural (PNL) y el análisis semántico han visto avances más amplios en los últimos años, y hay potencial para que estos influyan en futuras tareas de PAN. Algunas tendencias generales y avances en el campo semántico, que eventualmente pueden impactar en la investigación relacionada con PAN, incluyen:

- Embeddings y aprendizaje de representación:

Los avances en embeddings (por ejemplo, Word2Vec, GloVe) y los embeddings contextuales (por ejemplo, BERT, GPT) han mejorado la representación semántica de palabras y frases de una manera significativa, aunque no suficiente.

- Etiquetado semántico de roles:

Las técnicas de etiquetado semántico de funciones (roles), que implican identificar las relaciones entre las palabras en una oración y asignar funciones (por ejemplo, agente, paciente), han mejorado la comprensión del significado y la estructura del texto.

- Identificación de temáticas:

Los algoritmos de identificación de temáticas, como LDA (Latent Dirichlet Allocation) y NMF (Non-Negative Matrix Factorization), permiten la catalogación de temas dentro de una colección de documentos.

- Reconocimiento de entidades (NER):

Los sistemas de reconocimiento de entidades nombradas se han vuelto más sofisticados en la identificación de entidades (por ejemplo, nombres, ubicaciones, organizaciones) en el texto. La utilización de salidas NER podría añadir una capa semántica a los modelos de atribución de autoría.

- Análisis Cross-Modal:

La integración de información de diferentes modalidades, como texto e imágenes, para una comprensión más holística del contenido es un área de investigación en curso y pueden beneficiarse considerando información semántica derivada de diversas fuentes.

- Medidas de similitud semántica:

El desarrollo de medidas de similitud semántica más avanzadas, más allá de la similitud tradicional del coseno, puede contribuir a una comprensión más profunda de las relaciones semánticas entre los textos. Estas medidas pueden aplicarse para mejorar la exactitud de la atribución de autoría.

- Análisis Semántico:

Los avances en el análisis semántico, donde el lenguaje natural se convierte en una representación formal, pueden ayudar a extraer un significado semántico más profundo del texto. Esto podría ser relevante para entender la semántica de los estilos de autoría.

Es importante señalar que los logros de PAN continúan evolucionando a medida que se organizan nuevas tareas compartidas cada año, abordando los desafíos emergentes en el campo del análisis de textos. Investigadores y participantes contribuyen al éxito de PAN desarrollando soluciones innovadoras y avanzando en el estado del arte en la atribución de autoría, perfiles y detección de plagio.

2.2. Embeddings Semánticos

Como hemos comentado previamente, los word embeddings son representaciones numéricas abstractas de las relaciones entre palabras donde conceptos con semántica similar tienen representaciones similares. Hablamos de la similitud léxico-semántica de los embeddings, que se refiere a la medida en que los vectores de palabras o frases en el espacio vectorial reflejan las similitudes y relaciones semánticas entre ellas. Esto significa que palabras o frases similares en significado deberían tener vectores cercanos en ese espacio vectorial. En esencia, se basa en una idea conocida como "hipótesis distributiva" [Harris (1954)]: las palabras que son semánticamente similares también se usan de manera similar y es probable que aparezcan en contextos similares. Esto es especialmente útil para tareas como recuperación de información, agrupamiento de texto y búsqueda de similitud. Word2Vec es uno de los avances clave en el proceso del lenguaje natural (PLN), ya que presenta una forma eficiente de aprender word embeddings a partir de un corpus de texto determinado [Mikolov et al (2013a), Mikolov et al (2013b)].

El word embedding se centra específicamente en la representación vectorial de palabras, mientras que el embedding semántico puede abarcar representaciones más amplias que incluyen conceptos más grandes que palabras individuales. Los embeddings semánticos conforman una serie de técnicas de aprendizaje de representación (o aprendizaje de características) que codifican la semántica de datos como secuencias y gráficos en vectores, de modo que puedan ser utilizados por tareas de análisis estadístico y predicción de aprendizaje automático posteriores. Para la representación mediante embeddings semánticos se utilizan las redes neuronales feedforward [Mikolov et al (2013a)], las redes neuronales recurrentes [Peters et al (2018)] y los transformers [Devlin et al (2018)], mostrando un buen rendimiento. Dos arquitecturas clásicas de codificación automática para aprender representaciones de elementos secuenciales son Skip-gram y Continuous Bag-of-Words (CBoW) [Mikolov et al (2013a), Mikolov et al (2013b)]. El primero tiene como objetivo predecir el entorno de un elemento, mientras que el segundo tiene como objetivo predecir un elemento en función de su entorno.

Los enfoques antes mencionados (CBoW y skip-gram) se limitan al contexto local limitado por el tamaño de la ventana de contexto. Global Vectors (GloVe) aborda ese problema capturando estadísticas globales del corpus con una matriz de probabilidad de co-ocurrencia de palabras [Pennington et al (2014)]. Combina las ventajas de las dos principales familias de modelos de la literatura: factorización matricial global y métodos de ventana de contexto local. Este modelo aprovecha eficientemente la información estadística al entrenar solo en los elementos distintos de cero en una matriz de coocurrencia palabra-palabra, en lugar de en toda la matriz dispersa o en ventanas de contexto individuales en un corpus grande.

Si bien Word2Vec y GloVe ofrecen mejoras sustanciales con respecto a los métodos anteriores, ninguno de ellos logra codificar palabras desconocidas (tokens que no se procesaron en los corpus utilizados). FastText refina los word embeddings al complementar la matriz de embeddings aprendida con subpalabras para superar el desafío de los tokens que no se encuentran dentro del vocabulario [Joulin et al (2017)], lo que permite al modelo aprender representaciones de palabras más generales. A diferencia de los vectores de palabras entrenados sin supervisión de word2vec, las características de las palabras se pueden promediar para formar buenas representaciones de oraciones. Proponen un modelo para aprender representaciones de palabras teniendo en cuenta la morfología [Bojanowski et al (2017)]. Modela la morfología considerando unidades de subpalabras y representando palabras mediante la suma de sus n-grams de caracteres. En varias tareas, fastText obtiene un rendimiento a la par de los métodos propuestos inspirados en el aprendizaje profundo, aunque es mucho más rápido.

Una deficiencia importante de los modelos anteriores es su incapacidad para capturar descripciones contextuales de palabras, ya que todos producen una representación vectorial fija para cada palabra [Alshaabi et al (2022)].

2.3. Norma y dirección en word embeddings

Las representaciones distribuidas de palabras recopilan la información léxico-semántica a través de la norma y la dirección de los vectores de las palabras. La norma representa la importancia relativa de la palabra mientras que la dirección representa el significado de la palabra [Yokoi et al (2020)].

La norma de un vector de palabras codifica implícitamente el peso de la importancia de una palabra e indica la medida en que la palabra contribuye al significado general de una oración. El ángulo entre dos vectores de palabras (es decir, la diferencia entre la dirección de estos vectores) nos referencia la similitud de dos palabras.

La ganancia de información representa cuánta de ella obtenemos sobre la distribución de palabras de contexto [Oyama et al (2023)]. Según la composición de una oración como la suma de los vectores de las palabras [Mitchell and Lapata (2010)], la norma en los embeddings de palabras se considera que representa la importancia de la palabra en una oración porque cuanto más largo son los vectores, mayor es su influencia en la suma del vector. Los vectores los podemos representar en espacio euclidiano como vemos en la figura 1.

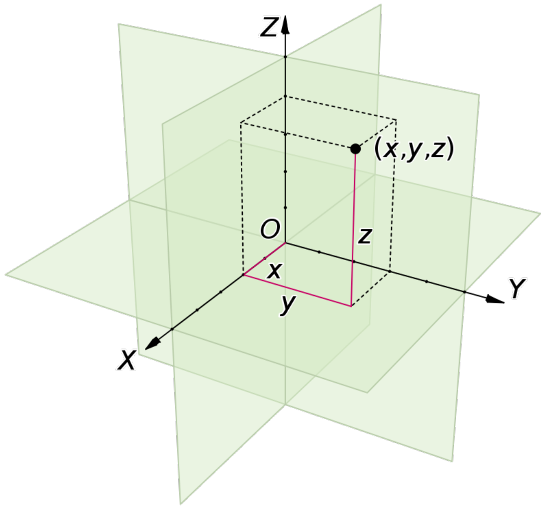

Figura 1: Representación de un punto en un espacio Euclidiano tridimensional

La norma euclideana de un vector es:

$$||x|| = \sqrt{x \cdot x}$$

La distancia euclidiana entre dos puntos de un espacio euclidiano es la norma del vector de traslación que asigna un punto al otro.

$$d(P, Q) = \overrightarrow{||PQ||}$$

El ángulo (no orientado) $\angle$ entre dos vectores distintos de cero es:

$$\angle = arcos(\frac{x \cdot y}{|x||y|})$$

Schakel afirmó que la norma de embeddings de palabras y la frecuencia de las palabras representan el significado de las palabras y demostró experimentalmente que los nombres propios tienen embeddings con normas mayores que las palabras de función [Schakel and Wilson (2015)]. Además, se ha demostrado experimentalmente que la norma de los embeddings de palabras son más pequeños para los tokens menos informativos [Arefyev et al (2018)].

En un embedding contextual, cuando una palabra aparece en distintos contextos, su vector se mueve en diferentes direcciones durante las actualizaciones. El vector final entonces, representa algún tipo de promedio ponderado sobre los diversos contextos.

Promediar sobre vectores que apuntan en diferentes direcciones típicamente resulta en un vector que se acorta con un número creciente de diferentes contextos en los que aparece la palabra. En un embedding estático como word2vec, cada palabra mantiene su norma y dirección.

Nuestro estudio se basa en la medida de similitud textual utilizando por separado la norma y la dirección de los vectores de palabras para comparar diferentes textos mediante embedding estático.

2.4. Stopwords

El concepto de “Stopword” introducido por H. P. Luhn en 1958 juega un papel muy importante en la práctica actual del PLN. Las stopwords se utilizan para reducir los datos de texto ruidosos, eliminar palabras desinformativas, acelerar el procesamiento de texto y minimizar la cantidad de memoria necesaria para almacenar datos. Ejemplos de stop words son los determinantes, preposiciones, conjunciones, etc. Eliminar stopword es un paso de procesamiento previo en la mayoría de las aplicaciones PNL, incluyendo IR (Recuperación de información) y TC (Clasificación de texto). Este paso no solo reduce el tamaño del corpus de texto, sino que también optimiza la potencia de procesamiento, reduce las entradas de índice para la búsqueda y, posteriormente, disminuye la complejidad en la necesidad del espacio y tiempo [Ladani et al (2020)].

Son palabras con bajo poder de discriminación. Se usan para conectar palabras de alto poder de discriminación mientras se construyen oraciones. No tienen sentido por si solas y tienen una alta frecuencia de aparición en el texto.

En nuestro caso, utilizaremos la librería NLTK (Natural Language Toolkit library) donde existe un listado de palabras en diferentes idiomas.

2.5. Sistemas generativos basados en Transformers

Los mecanismos de atención que han evolucionado a la inteligencia artificial generativa basada en el entrenamiento de gran cantidad de datos y una ingente capacidad computacional, han provocado la disrupción de modelos con gran capacidad creativa y de generación de contenido. No solo de textos (GPT[9], Bard[10], Llama2[11],...), sino de imágenes (Dall-e[12], Stable Diffusion[13],...), música (Beatoven[14], Aiva[15],...) video (Peech[16], Synthesia[17],...) y no solo unimodal, sino multimodal como generación de imágenes a través de texto (Midjourney[18],...) música a través de imágenes, de texto,... Esta tecnología abre un campo de oportunidades que ya está impactando tanto a nivel profesional como lúdico, hogar, salud,... pero, como dice un antiguo adagio del siglo I a.C. “Un gran poder conlleva una gran responsabilidad” y así como grandes beneficios, existen grandes peligros alrededor de esta tecnología. Alguno de esos peligros son la posible suplantación de la identidad en los textos y para ello trabajaremos sobre la atribución de autoría.

El análisis de un texto para realizar un perfil de su autor a través de su estilo sintáctico, léxico y semántico es un área de gran interés que está consiguiendo grandes avances. La extracción de características juega un papel fundamental y la mayoría de los métodos estadísticos que utilizan las características de estilo (longitud de las oraciones, tipo de oración,...) y características de contenido (frecuencia de palabras, riqueza de vocabulario, n-grams,...) obtienen unos buenos resultados [Kou et al (2020), Ahmed (2018)]. Los métodos de selección de características se categorizan como filtros, wrappers o métodos embedding.

Mientras que la clasificación sintáctica, léxica y morfológica mediante regresión o algunos métodos de aprendizaje automático obtiene resultados bastante precisos, el análisis de los textos desde el punto de vista semántico tiene mayor dificultad y carecen de transparencia e interpretación, que son esenciales en muchas aplicaciones reales como la detección de plagio o la elaboración de perfiles de autores que requieren evidencia o explicación explícita [Haiyan Wu et al (2021)].

[9]https://chat.openai.com/
[10]https://bard.google.com/
[11]https://ai.meta.com/llama/
[12]https://openai.com/dall-e-3
[13]https://stablediffusionweb.com/
[14]https://www.beatoven.ai/
[15]https://www.aiva.ai/
[16]https://www.peech-ai.com/
[17]https://www.synthesia.io/
[18]https://www.midjourney.com/

Con la llegada a finales de 2017 de modelos basados en la atención profunda denominados transformers [Vaswani et al (2017)], se produjo un avance en el panoráma del lenguaje natural siendo uno de los detonantes de la inteligencia artificial generativa. El marco codificador-decodificador, impulsado por bloques de atención, permite un procesamiento más rápido de la secuencia de entrada y al mismo tiempo preserva el contexto. Las adaptaciones recientes de los componentes básicos de Transformers continúan batiendo récords, mejorando el estado del arte en todos los puntos de referencia de PNL con aplicaciones recientes a la visión por computadora y el reconocimiento de patrones [Dosovitskiy et al (2021)].

Los transformers son una arquitectura de modelos de aprendizaje profundo introducida por Vaswani et al. en el artículo “Attention is All You Need” en 2017. El fundamento principal de los transformers es el mecanismo de atención, que permite a la red focalizarse en diferentes partes de la entrada de manera ponderada. En lugar de depender de la recurrencia o la convolución para procesar secuencias, los transformers utilizan capas de atención para evaluar todas las posiciones de la entrada simultáneamente. Esto mejora la eficiencia computacional y facilita el entrenamiento paralelo. El funcionamiento básico de los transformers se puede dividir en los siguientes componentes clave:

- Capa de atención multi-cabeza: La atención se calcula en paralelo en múltiples cabezas (subconjuntos de la capa). Cada cabeza aprende diferentes aspectos de la relación entre las palabras.

- Codificación posicional: Dado que los transformers no tienen una noción inherente de la posición en la secuencia, se incorpora información posicional adicional para conservar el orden de las palabras.

- Capa de codificación/decodificación: La entrada se procesa a través de múltiples capas de atención y redes completamente conectadas. En el caso de modelos de lenguaje, se utilizan capas de codificación para procesar la entrada, mientras que en la traducción automática, se utilizan capas de codificación y decodificación.

- Conexiones residuales y normalización por lotes: Se utilizan conexiones residuales para ayudar al flujo de información durante el entrenamiento y se aplica normalización por lotes para estabilizar y acelerar el proceso de entrenamiento.

- Función de activación: La función de activación, comúnmente la función de activación de GELU (Gaussian Error Linear Unit) o ReLU (Rectified Linear Unit), se aplica después de cada capa completamente conectada.

Los transformers han revolucionado el campo del aprendizaje profundo al proporcionar una arquitectura altamente paralelizable y eficiente para procesar secuencias de datos. Su capacidad para modelar relaciones complejas y captar patrones de largo alcance los ha convertido en la elección preferida en muchas aplicaciones de procesamiento del lenguaje natural y más allá.

El entrenamiento de un modelo Transformer, como el utilizado en GPT (Generative Pre-trained Transformer) de OpenAI, implica varios pasos. A continuación, se describe el proceso de entrenamiento de manera general:

- Recopilación de datos: Se reúne un conjunto de datos grande y diverso que refleje el dominio y las tareas para las cuales se desea que el modelo sea competente. En el caso de modelos de lenguaje, este conjunto de datos suele consistir en texto en diversos contextos y géneros.

- Tokenización y preprocesamiento: El texto se divide en unidades más pequeñas llamadas "tokens". Cada palabra o subpalabra puede ser un token. Además, se puede aplicar preprocesamiento como la normalización de texto y la tokenización para adaptar el formato del texto a la entrada del modelo.

- Arquitectura del modelo Transformer: Se define la arquitectura del modelo Transformer como hemos comentado anteriormente, que consta de capas de atención, capas completamente conectadas, normalización por lotes y funciones de activación. La arquitectura incluye componentes como la atención multi-cabeza y conexiones residuales.

- Inicialización de pesos: Los pesos del modelo se inicializan aleatoriamente. Esto es crucial para permitir que el modelo aprenda durante el entrenamiento.

- Función de pérdida y optimización: Se selecciona una función de pérdida que mide la discrepancia entre las predicciones del modelo y los objetivos reales. La optimización se realiza mediante algoritmos como el descenso de gradiente estocástico (SGD) o variantes más avanzadas como Adam.

- Entrenamiento no supervisado: El modelo se entrena de manera no supervisada para predecir el siguiente token en secuencias de texto. Se utilizan técnicas como el aprendizaje por atención para capturar patrones y dependencias en el texto.

- Ajuste de hiperparámetros: Se ajustan los hiperparámetros del modelo, como la tasa de aprendizaje y el tamaño del lote, para optimizar el rendimiento del modelo en el conjunto de datos de entrenamiento.

- Entrenamiento a gran escala: Los modelos Transformer, especialmente los utilizados en GPT, se entrenan a gran escala en hardware potente y con grandes cantidades de datos para aprovechar al máximo su capacidad de aprendizaje.

- Afinación fina (opcional): Es una técnica utilizada para mejorar y adaptar modelos de Deep Learning a tareas específicas, aprovechando un modelo preentrenado, donde se han validado la efectividad de los pesos utilizados y ajustándolo con datos adicionales relevantes enfocados a la tarea específica que se desarrolla. En algunos casos, el modelo preentrenado se puede afinar para tareas específicas o en dominios particulares utilizando conjuntos de datos adicionales.

Este proceso se basa en la premisa de aprendizaje no supervisado, donde el modelo se entrena para comprender la estructura y la semántica del lenguaje sin supervisión explícita para tareas específicas. Una vez entrenado, el modelo puede ser utilizado en tareas de generación de texto, traducción automática, respuesta a preguntas, entre otras aplicaciones.

La utilización de modelos neuronales para la captura de la información sintáctica y semántica ha acelerado los avances y se han desarrollado modelos cada vez más transparentes como los mecanismos de atención que ayudan a mejorar la transparencia de las redes neuronales y proporcionar una evaluación explícita y comprensible de la importancia de los pesos de las características introducidas [Wang et al (2014)], lo cual se ha mostrado como un método eficaz para capturar las relaciones no lineales entre las características y los objetivos de clasificación [Gui et al (2019)]. Ning Gui presenta una nueva arquitectura de selección de características basada en redes neuronales, denominada selección de características basada en la atención (AFS) que consta de dos módulos desmontables: un módulo de atención para la generación de pesos de características y un módulo de aprendizaje para el modelado de problemas.

Dentro de las tareas del procesamiento del lenguaje natural (PLN) podemos diferenciar, por un lado, los modelos no contextuales como los word embedding donde cada palabra tiene un único vector independientemente del contexto, lo que les convierte en estáticos. Esto plantea diferentes problemas [Ethayarajh (2019)] como que las palabras polisémicas comparten la misma representación. Por otro lado y en contraposición, están los modelos de representación de palabras contextualizadas que son sensibles al contexto donde aparece lo que les dota de cierto dinamismo. Modelos como ELMO [Peters et al (2018)] y BERT [Devlin et al (2018)].

ELMO crea representaciones contextualizadas de cada token concatenando los estados internos de un LSTM de 2 capas entrenada en una tarea de modelado bidireccional. En contraposición, BERT es un transformer bidireccional de 12 capas. Los modelos de representación contextualizada están basados en la premisa de que en un modelo lineal simple entrenado, la información semántica y sintáctica se codifica implícitamente en la representación contextualizada pero no nos dice cuánto de contextual es la representación y cuánto se podría sustituir por un word embedding estático si no completamente.

Los nuevos modelos de generación de texto como chatGPT y BARD[19] nos plantean una última cuestión sobre si es posible identificar características específicas de estos textos generados por estos modelos. GPT, desarrollado por Open AI, utiliza un decodificador autoregresivo para extraer características y predecir la siguiente palabra basado en las primeras palabras [Zhou et al (2023)]. ChatGPT se ha entrenado

[19] Durante la realización de este estudio se producen constantemente lanzamientos de nuevos modelos y nuevas versiones de modelos existentes mejorados como ChatGPT 4, Bard con su modelo GeminiPro, la versión de Apple y otros. La velocidad de desarrollo y el interés generado nos sitúa en una época de incertidumbre en los impactos que producirá en la sociedad.

utilizando aprendizaje supervisado. En este proceso, a ChatGPT se le presenta un conjunto de datos de texto y código, y se le enseña a predecir la siguiente palabra o frase en cada secuencia. Esto se hace mediante el uso de una red neuronal artificial, que es un tipo de algoritmo de aprendizaje automático que puede aprender de datos. El conjunto de datos utilizado para entrenar ChatGPT incluye texto de una variedad de fuentes, incluyendo libros, artículos, sitios web y código. Esto le permite a ChatGPT aprender sobre una amplia gama de temas y generar texto que sea natural y fluido. La cantidad exacta de datos de texto extraídos de la web utilizados para entrenar ChatGPT no se ha especificado públicamente en la documentación de OpenAI. Sin embargo, se sabe que modelos de la serie GPT, como GPT-3, se entrenaron en una vasta cantidad de datos, que puede ser del orden de cientos de gigabytes o incluso terabytes de texto de la web (Se estiman 175 mil millones de parámetros). OpenAI se enfocó en entrenar estos modelos con grandes conjuntos de datos para que puedan aprender de manera efectiva el lenguaje natural y generar respuestas coherentes y contextuales. Además del tamaño de los datos, la calidad y diversidad de los datos de entrenamiento también son factores importantes para el éxito del modelo.

BARD, desarrollado por Google, está basado en LaMDA (Language model for dialogue application), construido sobre la arquitectura transformer. Aplica la aproximación de fine tuning con datos anotados de gran calidad y fuentes externas de conocimiento para mejorar el desempeño del modelo. Es un modelo de lenguaje factual, tiene 137 mil millones de parámetros, entrenado en un conjunto de datos masivo de texto y código, que incluye libros, artículos, sitios web y código. BARD puede generar texto, traducir idiomas, escribir diferentes tipos de contenido creativo y responder a tus preguntas de forma informativa. BARD tiene un conjunto de datos de entrenamiento diverso, lo que le permite generar texto natural y fluido. Además, BARD se ha entrenado específicamente para aplicaciones de diálogo, lo que le permite entender y responder a las preguntas de forma informativa.

3. Fundamentos teóricos

En este apartado vamos a centrar los fundamentos en que nos basamos para la elección de las medidas para el cálculo de las distancias que utilizaremos en la experimentación.

Ya hemos comentado previamente que nuestro planteamiento está orientado al análisis semántico y no al sintáctico o estilo métrico. La base de nuestro planteamiento es la representación vectorial no contextual, de tal forma que la fijación de cada término en un espacio de representación de embeddings estáticos nos permita el análisis y la comparación entre distintos textos de un determinado autor con los de otro.

Para la experimentación de nuestra colección, nos vamos a centrar en medir la especificidad y la distribución semántica sobre los resultados obtenidos en nuestro análisis en Word2vec para demostrar que cada autor tiene un perfil característico basado en la especificidad y en la variedad y amplitud del vocabulario que utiliza, es decir, que sea característica del individuo de forma continua a lo largo de su obra. También queremos demostrar que los textos generados por la IA generativa se caracterizan por estos mismos atributos.

A pesar de que la similitud coseno es la medida estándar en la representación de la distribución, la cual, nosotros también utilizaremos, es una medida demasiado genérica para los objetivos que buscamos. La isometría angular indica la correspondencia entre la similitud semántica de las unidades del texto y sus ángulos de los embedding, es decir, la similitud entre la representación de dos vectores se incrementa si el coseno del ángulo formado por los vectores crece, esto es, si el ángulo se reduce. Dos pares de palabras con la misma apertura de ángulo pero diferente distancia al origen, aunque proporcional, tienen la misma similitud coseno. Para una mayor comprensión de los datos obtenidos en la similitud coseno, añadimos otras cuatro medidas que nos permitan comprobar cómo las propiedades geométricas se mantienen en el espacio de representación semántico y cómo la relación de estas cinco medidas nos ayudan en la comparación y caracterización de los distintos textos. A la medida de “similitud coseno” y trabajando con distancias euclidianas, añadimos la “distancia del centroide al origen”, la “media de la distancia entre todas las palabras”, la “media de la distancia de las palabras al centroide" y la "media de la distancia de las palabras al origen".

Los modelos contextualizados no siempre mantienen la isometría con respecto a la similitud semántica de las palabras. La representación de las palabras debería estar ampliamente distribuida para representar los diferentes significados semánticos, sin embargo, esta representación de las palabras limita la expresividad del aprendizaje del modelo. Esto es denominado el problema de la degeneración de la representación [Gao et al (2019), Li et al (2020), Cai et al (2021), Wu et al (2020)]. El resultado es que los modelos contextuales son más predictivos como modelos del lenguaje pero no muy efectivos en términos de representación de un texto dentro de un espacio semántico [Amigó et al (2022)].

Para la representación se utilizan Corpus entrenados como modelos distribuidos a base de billones de palabras de textos escritos por diferentes autores para una audiencia determinada, un género específico, con diferentes estilos y en diferentes épocas que varían el uso y significado de las palabras [Johns (2023)].

En definitiva, el significado de las palabras no deja de ser la media de los usos de los diferentes autores en los diferentes tipos de textos escritos y no el significado específico de ese autor individual. Esto representa una ventaja en los modelos contextuales para poder inferir con precisión las palabras buscadas o en la búsqueda del significado dentro del contexto de la oración. En contraposición, se requiere de Corpus entrenados de gran tamaño y no siempre de fácil acceso o de una capacidad computacional de gran tamaño. Además, dota a los modelos contextuales de una característica antes mencionada que hemos denominado dinamismo y que significa que se adapta al contexto de la frase dentro de ese texto y comparado a los otros textos dentro del Corpus. Aunque es una cualidad que ha demostrado ser de una gran utilidad, nos dificulta el poder comparar distintos textos sin que varíen ligeramente los resultados tras cada experimentación.

Brendan T. Johns en su estudio [Johns et al (2018)] demuestra que los libros escritos en el mismo género son más parecidos, desde el punto de vista semántico, que los de distintos géneros, y que los libros del mismo autor son mucho más similares que los de diferentes autores. Esto demuestra que los autores tienen su propio y único lenguaje. Para el estudio se construye un Corpus con 1850 libros y 240 millones de palabras de novelas de ficción de 7 géneros distintos. El modelo trabaja a nivel de oración y recopila dos tipos de información estadística: cada palabra tiene su propia representación en función de su contexto y el orden. La similitud entre 2 libros se evalúa tomando el vector coseno de sus respectivos vectores que representan a cada libro. Este método nos plantea la complejidad de crear un Corpus específico y en cierto modo aleatorio (la elección de las obras no tienen ningún criterio definido), por lo que no serviría para un propósito más general, por ejemplo, si queremos compararlo con el género periodístico o filosófico. Además, el contexto de cada palabra, es comparado con el conjunto de las obras recopiladas y si escogiéramos otras distintas, el contexto cambiaría.

Según estudios previos [Levy and Goldberg (2014), Arora et al (2016)], tanto los modelos de embeddings estáticos (Skip-Gram Negative Sampling (SGNS), GloVe) como contextuales (BERT, GPT) mantienen la correspondencia entre la norma del vector de embedding y la especificidad de la unidad lingüística representada. Por lo tanto, la norma vectorial se puede utilizar como función de Contenido de Información. Tambien demuestran la correspondencia entre la distancia angular y la similitud semántica de palabras en modelos de embedding estáticos como SGNS y GloVe.

Cuanta menos información específica contenga una palabra mas cerca estará respecto al origen de coordenadas y debe estar equidistante de cualesquiera palabras con la misma cantidad de información. En definitiva, tiene sentido que las representaciones estén distribuidas alrededor del origen de coordenadas a una distancia proporcional a la cantidad de información que contienen.

Los word embeddings estáticos tienen una representación única que aproxima a relaciones léxico-semanticas, mientras que los contextuales no lo son. Es por ello que utilizamos los word embedding estáticos generados por un Corpus genérico y universal como es el modelo word2vec [Skrlj et al (2019)], de carácter estático, para comprobar si la singularidad y la unicidad del uso del lenguaje se refleja en cada texto. Resaltaremos la idea que utilizar un método donde se fije el resultado independientemente del contexto nos permite comparar y entender ciertas características que van a complementar este análisis.

4. Metodología

Para resolver las cuestiones que planteamos en la propuesta que se basan en extraer las características específicas de un texto basado en la orientación semántica y en la especificidad del vocabulario dentro de espacios de representación de embeddings estáticos, vamos a utilizar cinco medidas que son la similitud coseno, la “distancia del centroide al origen”, la “media de la distancia entre todas las palabras”, la “media de la distancia de las palabras al centroide” y la “media de la distancia de las palabras al origen”. Con ello, lo que pretendemos demostrar es que el uso del vocabulario es característico de cada autor y se mantiene a lo largo de su producción.

Respecto a la distancia al origen, Li et al [Li et al (2020)] observó empíricamente que las palabras más frecuentes están más cerca del origen mientras que las menos frecuentes se alejan de él, es decir, su representación se distribuye alrededor del origen de coordenadas a la distancia proporcional a la cantidad de información que contiene. Para nuestro análisis hemos eliminado las stopwords que se encuentran más cerca del origen pero que aportan menor información y solo utilizamos palabras individuales para comprobar si la especificidad resultante es suficiente para caracterizar a cada autor.

Respecto a la distancia de las palabras al centroide y la media de la distancia entre todas las palabras, obtenemos la distancia media y por tanto cómo de grande es la “huella” o el volumen de la proyección de las palabras del texto. Esto nos dará pistas de lo variado que pueda ser el vocabulario empleado y añadirá otra característica más que ayude a caracterizar al autor. Con las medidas de la distancia del centroide al origen y la media de la distancia de las palabras al origen nos dice cómo de específico es el vocabulario empleado y nos permite comparar la especificidad entre cada texto.

También utilizamos el concepto de centroide teórico como centro geométrico de la representación del conjunto de las palabras en el espacio vectorial. Este punto nos servirá en alguna experimentación como punto de referencia para algunas medidas. En lenguaje natural, la diferencia entre un centroide teórico y un centroide real está asociada comúnmente con conceptos en estadística, análisis de datos y geometría. Aquí se explica cada término:

- Centroide Teórico:

El centroide teórico es un punto imaginario o calculado que representa el centro geométrico o centro de masa de un conjunto de puntos o elementos. En estadística y análisis de datos, el centroide teórico puede calcularse utilizando fórmulas específicas, como el promedio ponderado de las coordenadas de un conjunto de datos multidimensional. En geometría, el centroide es el punto donde se equilibran las masas de un objeto.

- Centroide Real:

El centroide real es el punto real en el espacio que corresponde al centro geométrico o de masa de un objeto físico. Este punto es determinado mediante mediciones o evaluaciones directas de las características físicas del objeto. En contextos prácticos, el centroide real puede ser el resultado de mediciones precisas de las posiciones o características de los elementos que componen un objeto. En resumen, la diferencia entre un centroide teórico y un centroide real radica en su naturaleza: el centroide teórico es un punto calculado o imaginario basado en fórmulas matemáticas, mientras que el centroide real es el punto físico medido o determinado mediante observaciones directas. Ambos conceptos son relevantes en diferentes contextos, ya sea en análisis de datos, geometría, física, o en otras disciplinas que involucren la determinación del centro de masa o equilibrio de un conjunto de elementos.

4.1. Análisis de datos y descripción

Para la realización de este estudio hemos confeccionado nuestra propia colección motivado por no haber encontrado una base de datos que se adecue a nuestras necesidades, como es que exista un grupo de control homogéneo, otro grupo de textos generados por autores más heterogéneos y un último grupo de textos generados por IA generativa. Además, entendemos que, dentro de los objetivos de un TFM, se encontraría el comprender y desarrollar todas las fases de una elaboración, preparación, limpieza y procesado de una colección.

Para poder configurar nuestra colección hemos tenido que asegurarnos que los textos que utilizamos son textos originales y que no han sufrido distorsiones o arreglos por parte de ediciones posteriores. Así mismo, hemos descartado bases de textos como las ofrecidas en PAN porque la mayoría se han construido a base de datos sintéticos.

También consideramos necesario que haya una colección de textos generados por inteligencia artificial y otras creadas por seres humanos. Dentro de las generadas por IA, hemos escogido ChatGPT y Bard porque queremos validar nuestras hipótesis con dos herramientas distintas de IA. La información que disponemos de cada modelo, debido al secretismo de estas compañías, es la publicada y no nos permite entrar al detalle técnico y la comprensión profunda de su funcionamiento mas que a alto nivel.

Respecto a la generada por seres humanos, hemos decidido recopilar textos de dos grupos distintos como son los periodistas, que tomaremos como grupo de control por la homogeneidad de sus estilos, y autores clásicos con distintos géneros y diferentes temáticas que sean más heterogéneos.

4.1.1. Creación de los Corpus

La elección de los textos viene determinado por:

- Textos periodísticos como grupo de control. La narrativa utilizada por el género periodístico suele ser bastante similar porque trata de abarcar a un público lo más amplio posible, por lo que el vocabulario es más general y menos técnico. Para poder utilizarlo como grupo de control respecto al resto de textos, hemos elegido, dentro del género de las noticias, temáticas distintas como deportes, religión, opinión y salud y poder comprobar que se mantienen en rangos similares.

- Textos de autores clásicos. Autores reconocidos en la literatura universal con dos géneros específicos como son la novela y el ensayo filosófico. Con estos dos géneros queremos validar cuánto de característico es el vocabulario de cada autor sobre género, independientemente de que el ensayo filosófico sea más técnico que la novela. Cuando nos referimos que un texto es más técnico, queremos decir que utiliza vocabulario específico a esa materia y que normalmente no se utilizaría en una conversación informal. Un ejemplo podría ser la utilización de la palabra “paradigma” en un formato más académico y su sinónimo “modelo” cuyo uso es mas común.

- Textos generados por inteligencia artificial generativa. Generamos textos con dos herramientas de IAG como son ChatGPT y Bard con enfoques más técnicos y generales a través de prompts. Para los enfoques más técnicos, lo que hacemos es indicar a los modelos, no solo que la temática pueda estar relacionada con un tema específico como pueden ser los planetas del sistema solar o el concepto del imperativo moral de Kant, sino que utilice términos técnicos para un auditorio técnico, mientras que cuando no se lo indicamos, lo explica con un vocabulario más coloquial. Del mismo modo, probamos, indicándole al modelo mediante un prompt que utilice un lenguaje sencillo y simple para que lo pueda entender un niño y comprobamos en ambos casos que existe variación en la generación de un mismo texto.

Los datos que hemos utilizado los construimos a través de una agrupación de distintos escritores y temáticas con un total de 175 textos. Por un lado hemos utilizado 15 textos de obras distintas de cada uno (excepto Kant y Hume que son 10). Estos textos son de una longitud aproximada de 1.500 palabras cada uno con el objetivo de que tras la limpieza podamos disponer de textos alrededor de 350 palabras para el análisis. Podríamos dividir esta base de datos en tres grupos que serían:

- Artículos de periodistas de la CNN como grupo de control.

- Fragmentos de las obras de autores clásicos.

- Recopilación de textos generados por IA generativa.

La distribución y características de los textos los podemos ver en la tabla 1.

Fuente	Género	Subtemática	Autor	Textos
CNN	Noticias	Deportes	Ben Church	15
CNN	Noticias	Opinión	John D. Sutter	15
CNN	Noticias	Religión	Daniel Burke	15
CNN	Noticias	Salud	Jessica Ravitz	15
Proyecto Gutemberg	Novela	Varios	Charles Dickens	15
Proyecto Gutemberg	Filosofía	Varios	Immanuel Kant	10
Proyecto Gutemberg	Novela	Varios	Rudyard Kipling	15
Proyecto Gutemberg	Filosofía	Varios	David Hume	10
Proyecto Gutemberg	Novela	Varios	G. K. Chesterton	15
Bing	Ensayo	Varios	GPT	15
Google	Ensayo	Varios	BARD	15
Bing	Ensayo	Prompt Técnico	GPT	15
Bing	Ensayo	Prompt Infantil	GPT	15

Tabla 1: Composición Corpus de textos

Para los textos de los artículos de periodistas de la CNN, hemos utilizado la base datos que se pueden encontrar en kaggle[20]. Este conjunto de datos contiene alrededor de 38000 líneas de artículos de noticias de CNN del año 2011 al 2022. Los datos se recopilaron mediante un rastreador web. Con esta base de datos vamos a configurar nuestro grupo de control para lo cual, elegimos a 4 periodistas de la CNN como son Ben Church, John D. Sutter, Daniel Burke y Jessica Ravitzs. Estos 4 periodistas escriben sobre temáticas distintas y específicas cada uno como es la religión, la salud, la política y los deportes. Los utilizamos como grupo de control porque entendemos que la forma de escribir de un artículo periodístico tiene la finalidad de llegar a un grupo de lectores amplio, por lo que la forma de expresarse y su mensaje debe estar adaptado a un público masivo. Esto condicionará el vocabulario empleado y su variedad.

Además de los periodistas antes mencionados, hemos utilizados textos de las obras de 5 escritores clásicos como Kant, Hume, Charles Dickens, Rudyard Kipling y Ches-

[20]https://www.kaggle.com/datasets/hadasu92/cnn-articles

terton. La fuente utilizada es el proyecto Gutemberg[21] y la recopilación de los textos para conformar el dataset lo hemos hecho manualmente. Respecto a los géneros, hemos elegido el ensayo y la novela. Respecto a la temática hemos seleccionado textos filosóficos y novelas de aventuras, cuentos,... en definitiva muy variado.

Y el tercer grupo son 4 recopilaciones de 15 textos cada uno, generados con GPT de Bing sobre temáticas diversas como filosofía, tecnología, ... así como otros textos generados por Bard, la tecnología de Google. Una vez que obtuvimos los resultados de la experimentación, nos pareció interesante investigar si eramos capaces de orientar los resultados hacia una mayor y menor especificidad y una mayor y menor dispersión de la representación de las palabras respecto a su centroide. Para ello, elaboramos dos nuevas agrupaciones de textos creados por ChatGPT 3.5 donde los textos creados alrededor de las mismas temáticas que con los textos anteriores de GPT y Bard, pero con el prompt más específico. En un caso la redacción de los ensayos se debía hacer con lenguaje más específico como si lo expusiera un técnico para una audiencia tambien técnica. En el caso contrario, le pedimos que se expusiera con un lenguaje mas sencillo como si lo explicara un niño. Intentamos replicar la técnica con Bard, pero la contestación fue que no puede escribir un ensayo de 1.500 palabras para que un niño lo entendiera porque es demasiado largo y complejo.

4.1.2. Limpieza de los datos

Una vez realizada la recopilación de diversas fuentes y documentos para la construcción de la colección y el preprocesado para convertir toda la información a un único formato CSV, comenzaremos con la tokenización y limpieza de los datos mediante la selección del vocabulario que necesitamos, su lematización y la vectorización con el modelo pre-entrenado Word2vec[22]. Para ello utilizamos SpaCy y Gensim.

El preprocesamiento de datos lo hacemos trabajando en Python con SpaCy[23]. SpaCy es una poderosa biblioteca de procesamiento de lenguaje natural que nos ayuda a realizar diversas tareas de preprocesamiento, como eliminar caracteres no deseados

[21] https://www.gutenberg.org/

[22] https://code.google.com/archive/p/word2vec/

[23] https://spacy.io/models

como los signos de exclamación, puntos, comas,... y las stopwords, convertir el texto a minúsculas, dividir el texto en oraciones y tokenizar las oraciones en palabras. En este caso, lo primero que hacemos es etiquetar las palabras en función de si pertenecen al grupo de sustantivos, adjetivos, verbos o adverbios para despues lematizar cada grupo. Finalmente conformamos un único grupo. El propósito es experimentar con aquellas palabras que nos ofrecen información relevante para poder resaltar las diferencias a la hora de comparar textos.

Para el entrenamiento del modelo utilizamos Gensim. Gensim es una biblioteca popular en Python para el procesamiento de lenguaje natural (NLP) y la modelización. Uno de los usos más conocidos de Gensim es la creación y entrenamiento de modelos Word2Vec, que se utilizan para representar palabras en vectores distribuidos. Hay modelos disponibles en línea que podemos usar con Gensim. En este caso utilizamos el modelo de conjunto de datos de Google News[24], que proporciona vectores previamente entrenados en parte del conjunto de datos de Google News (alrededor de 100 mil millones de palabras). El modelo contiene vectores de 300 dimensiones para 3 millones de palabras y frases. En este caso los vectores que representan las palabras se conforman por 300 dimensiones.

El resultado son conjuntos de palabras de alrededor de 350 unidades a las cuales, hemos transformado en embeddings a través de Word2vec. La proyección de estos embeddings son los que nos van a permitir medir distancias para poder comparar los distintos textos.

4.2. Medidas de evaluación

A continuación describimos las métricas de evaluación que utilizaremos.

4.2.1. Similitud Coseno

La similitud coseno de palabras es una medida que se utiliza para determinar cuán similares son dos palabras en función de su significado y su relación en un espacio vectorial de palabras. Esta medida se basa en la idea de que las palabras pueden ser

[24]https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing

representadas como vectores en un espacio multidimensional, y la similitud coseno evalúa la similitud entre estos vectores. La similitud coseno mide la dirección y la proximidad entre los vectores que representan las palabras en un espacio vectorial. Cuanto más cercanos estén los vectores, mayor será la similitud coseno y, por lo tanto, mayor será la similitud entre los documentos. Al ser el coseno del ángulo, el rango de su valor está limitado entre 0 y 1.

$$\hat{A}\ \hat{B}/||A||\ ||B|| = Similitud\ Coseno$$

- A y B son los vectores que representan las dos palabras que se comparan.

- Producto escalar de los vectores A y B.

- ||A|| y ||B|| representan la norma (longitud) de los vectores A y B.

Al calcular esta fórmula, se obtiene un valor que varía entre -1 y 1. Cuanto más cercano sea el resultado a 1, mayor será la similitud entre las dos palabras en términos de su significado. Un valor de 1 indica que las palabras son idénticas, mientras que un valor cercano a 0 indica poca similitud y un valor negativo puede indicar oposición en términos de significado.

Figura 2: Similitud Coseno

Esta medida es fundamental en NLP porque permite capturar la semántica y la similitud de palabras de manera más precisa que enfoques basados en frecuencia de palabras. Los modelos de Word Embeddings, que subyacen a la similitud coseno de palabras, se entrenan en grandes conjuntos de datos y capturan las relaciones semánticas y contextuales entre palabras como podemos observar en la figura 2.

Skip-Grams with Negative Sampling (SGNS) [Mikolov et al (2013a)] busca representar cada palabra y cada contexto como vectores dimensionales, de modo que las palabras que son "similares" entre sí tendrán representaciones vectoriales similares. Lo hace tratando de maximizar una función del producto del vector de la palabra y el vector del contexto. Levy [Levy et al (2015)] interpreta que esta función agrega términos de similitud de primer orden a la función de similitud de segundo orden. Los términos de primer orden miden la tendencia de una palabra a aparecer en el contexto de la otra. Los términos de segundo orden miden la medida en que las dos palabras son reemplazables en base a sus tendencias a aparecer en contextos similares, y son la manifestación de la hipótesis de distribución de Harris [Harris (1954)]. La similitud calculada es una combinación simétrica de similitudes de primer y segundo orden normalizadas. Esta medida de similitud establece que las palabras son similares si tienden a aparecer en contextos similares, o si tienden a aparecer en los contextos del otro (y preferiblemente ambos). En su experimentación encuentra que la normalización estándar usando la medición de similitud del coseno es consistentemente superior.

En la similitud coseno entre vectores lo que determina su valor es la proporción del tamaño de los vectores y el ángulo que lo forma, por lo que, a pesar de ser una medida estándar, nos interesa desagregar en otras dos medidas como es la especificidad y la dispersión de sus vectores.

4.2.2. Distancia al centroide

La distancia al centroide es una medida que se utiliza para evaluar cuán similares o cercanas están un conjunto de palabras o vectores de palabras a un punto central (centroide) en un espacio multidimensional. Esta técnica es comúnmente utilizada en procesamiento de lenguaje natural (NLP) y análisis de datos para identificar grupos de palabras relacionadas en función de su representación vectorial. Esta medida nos va a permitir medir el grado de dispersión de los vectores respecto al centroide y refleja si el vocabulario empleado es variado o heterogéneo, tal y como se observa en la figura 3.

El centroide es un punto en el espacio vectorial que se calcula como el promedio de los vectores de las palabras de interés. En este caso, lo primero que hacemos es calcular el centroide de las palabras de cada texto y calcular la media de la distancia de cada palabra al centroide, con lo que obtenemos otra medida para poder comparar la huella dejada por el conjunto de las palabras de cada texto.

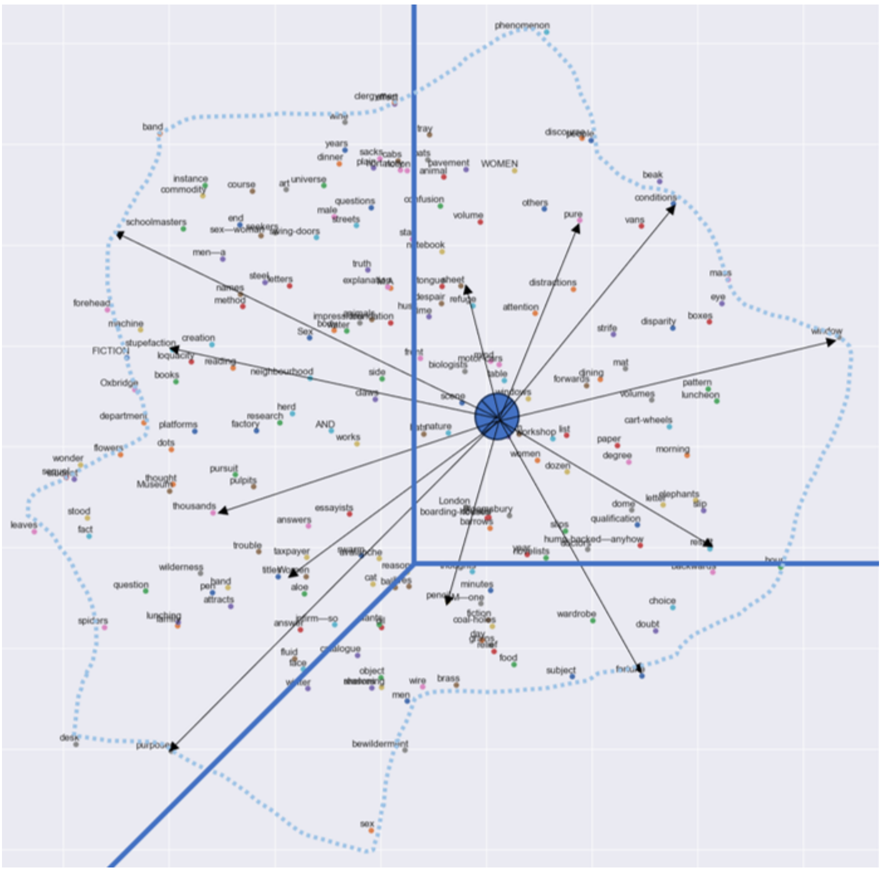

Figura 3: Distancia media al centroide

La distancia al centroide se utiliza en diversas aplicaciones, como la agrupación de palabras relacionadas (clustering de palabras), la búsqueda de palabras similares o relacionadas, la identificación de palabras clave en un conjunto de documentos y la extracción de temas de un conjunto de textos.

4.2.3. Distancia media entre todas las palabras

La distancia media entre todas las palabras es una medida similar a la distancia al centroide y nos permite evaluar cuán similares o cercanas están un conjunto de palabras o vectores de palabras entre sí en un espacio multidimensional, representado en la figura 4. Mientras que la distancia al centroide podría entenderse como la media del radio del volumen generado por la representación espacial de las palabras, la distancia media entre las palabras se podría entender como la media del diámetro de este volumen.

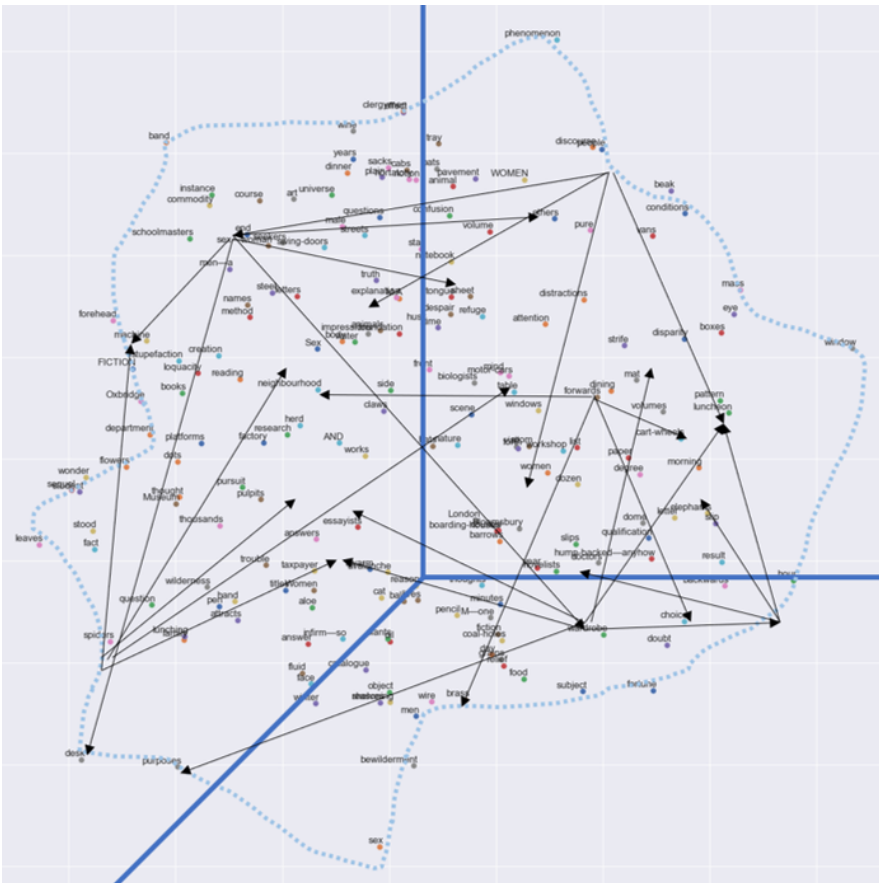

Figura 4: Distancia media entre todas

La ventaja de comparar dos medidas similares en el mismo espacio vectorial es poder verificar que los resultados son coherentes entre sí y que la pérdida de información en el proceso de cálculo no distorsionan los resultados. Esa pérdida de la información se puede dar cuando reducimos los vectores de 300 dimensiones a un vector unidimensional donde podamos medir las distancia euclidianas. Para el caso de la distancia al centroide, una vez hemos obtenido la representación geométrica de las palabras en vectores de 300 dimensiones, es cuando los reducimos a una única dimensión. La perdida de información en este paso es difícil de cuantificar y por eso nos debemos apoyar en los resultados de las distintas medidas para comprobar la coherencia de los distintos resultados.

4.2.4. Distancia del centroide al origen

La distancia del centroide de las palabras representadas al origen de coordenadas en un espacio multidimensional se refiere a la medida de la distancia entre el punto central que representa la media de un conjunto de vectores de palabras (centroide) y el punto de referencia que es el origen de coordenadas (0,0,0,...,0) en ese espacio vectorial. Esta medida se utiliza para evaluar cuán lejos o cerca están las palabras o conceptos representados por el centroide del punto de referencia en ese espacio multidimensional. Esta distancia se calcula utilizando la fórmula de la distancia euclidiana. La distancia euclidiana mide la longitud del segmento de línea que conecta el centroide al origen de coordenadas en el espacio multidimensional. Cuanto mayor sea la distancia euclidiana, más lejos estará el centroide del origen de coordenadas.

Al calcular la distancia del centroide al origen, figura 5, estamos midiendo la especificidad del vocabulario empleado. Al haber delimitado el estudio a palabras individuales, la diferencia de especificidad es menor que con conjuntos de palabras, pero queremos demostrar que las diferencias son suficientes para la caracterización de los textos.

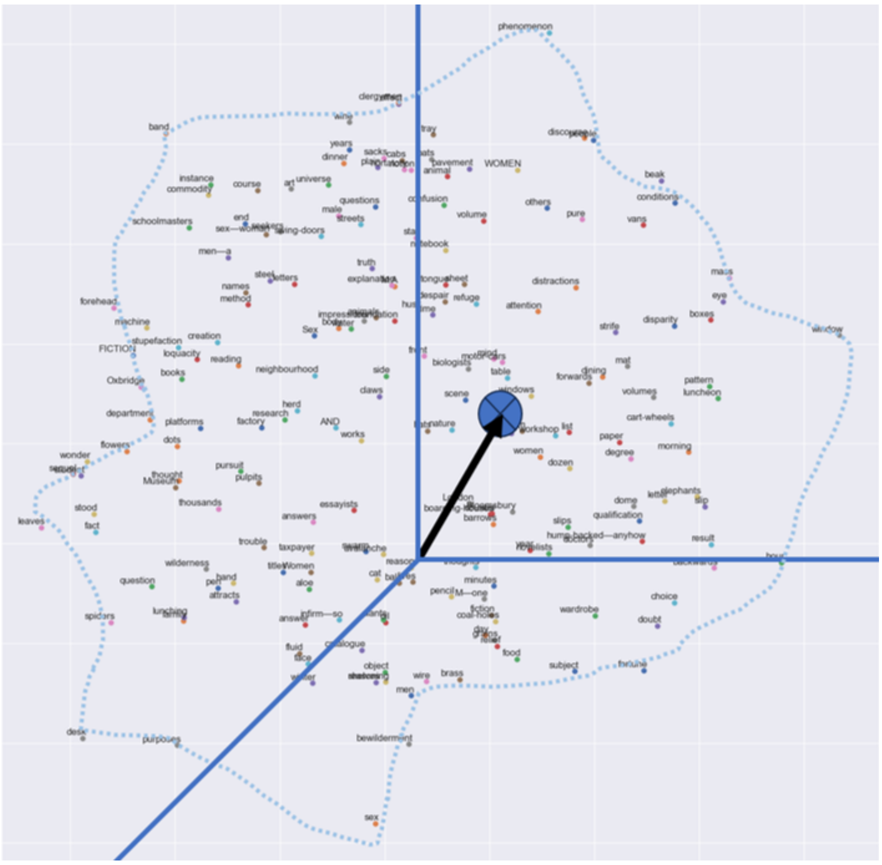

Figura 5: Distancia Centroide al origen

La distancia del centroide al origen de coordenadas se utiliza en diversas aplicaciones para evaluar la centralidad o prominencia de un conjunto de palabras en un espacio vectorial. Cuanto mayor sea esta distancia, más alejadas estarán las palabras representadas por el centroide del punto de referencia en el origen. Esta medida puede ayudar a identificar qué conjuntos de palabras son más distintivos o destacados en relación con el espacio vectorial en el que se encuentran.

4.2.5. Distancia media de las palabras al origen

En este caso utilizamos la medida de la distancia media de las palabras al origen como complemento a la medida del centroide al origen. El propósito de añadir esta medida es la de calcular la media de los vectores de 300 dimensiones para finalmente reducirlo a una dimensión y entonces calculamos la distancia euclidiana. De esta forma podremos comprobar si la pérdida de información durante la reducción escalar es significativa.

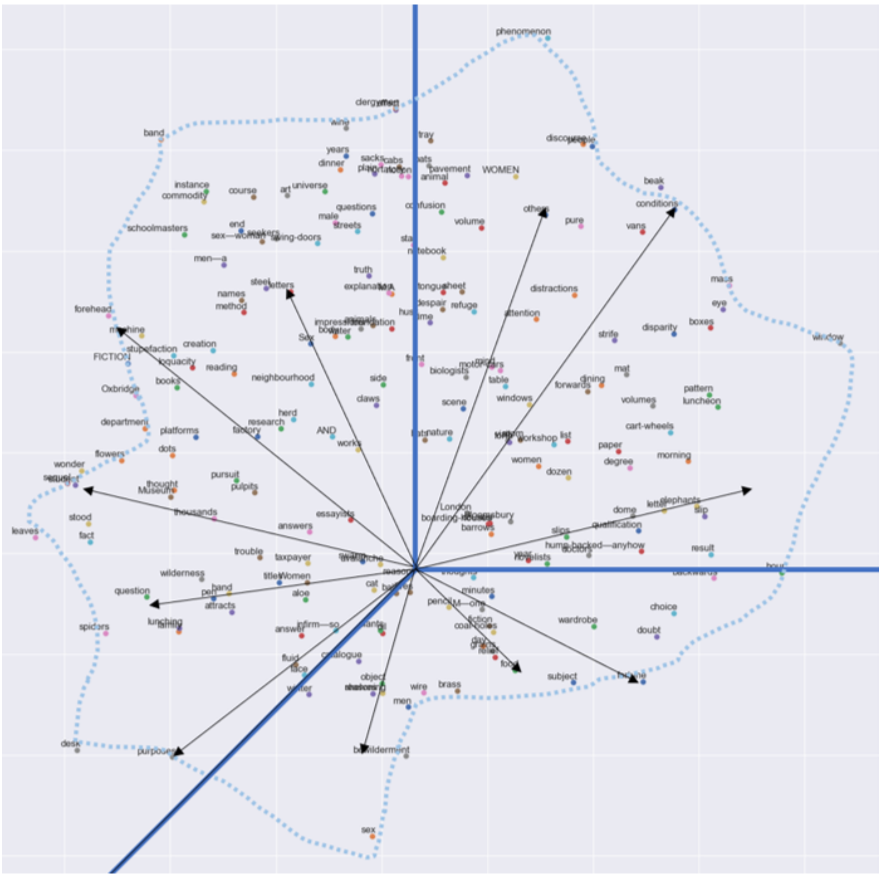

Figura 6: Imagen Dist media al origen

Otra información que podremos obtener es cuanto de desplazado se encuentra el centroide respecto al origen y así hacernos una idea de cómo es la distribución de las palabras en el espacio alrededor del eje de coordenadas.

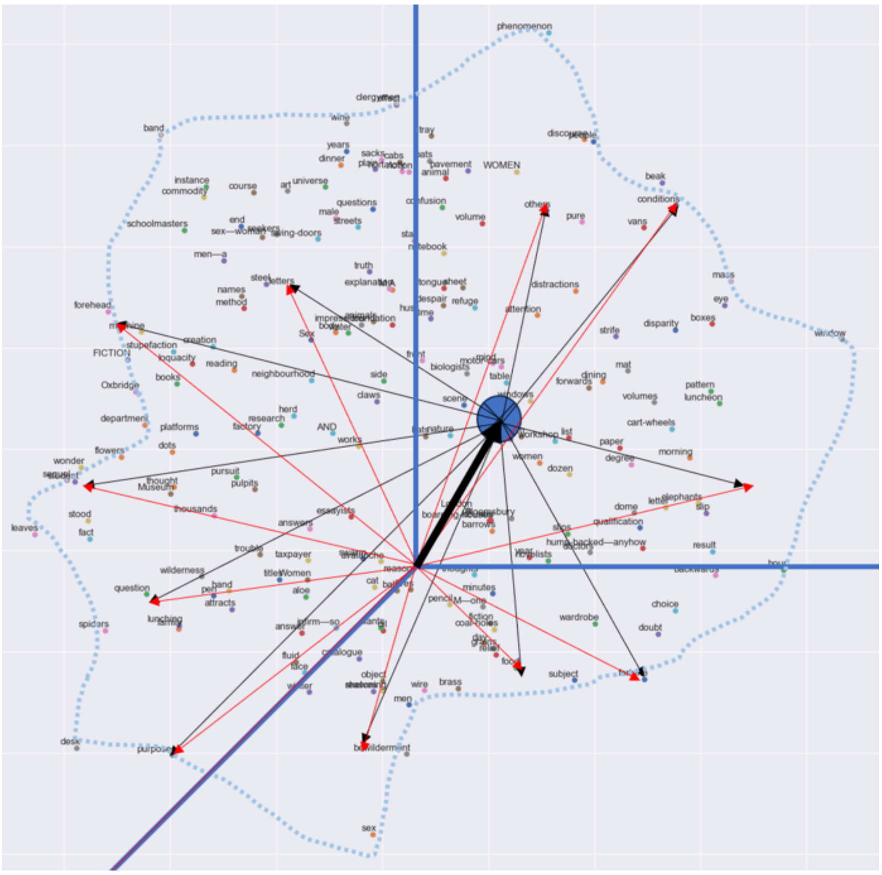

Figura 7: Distancia media al origen y al centroide

Hemos comentado anteriormente que cuanta menos información específica contenga una palabra mas cerca estará respecto al origen de coordenadas. Por lo tanto, para una mayor optimización de la representación de las palabras, eliminaremos las palabras comunes y stopwords y nos quedaremos con los sustantivos, verbos, adverbios y adjetivos, ya que pensamos que son las palabras que mejor representan y caracterizan a un autor.

5. Experimentación y evaluación

Ya hemos comentado previamente que nuestro planteamiento está orientado al análisis semántico y no al sintáctico o estilo métrico mediante la representación vectorial no contextual, y así poder comparar los distintos textos de un determinado autor con los de otro.

Para la experimentación nos centramos en medir la especificidad y la distribución semántica mediante la similitud coseno y añadimos otras cuatro medidas que son:

- La distancia del centroide al origen.

- La media de la distancia entre todas las palabras.

- La media de la distancia de las palabras al centroide.

- La media de la distancia de las palabras al origen.

La razón por la que duplicamos el cálculo en dos medidas como es la distancia del centroide al origen con la media de la distancia de las palabras al origen y la media de la distancia entre todas las palabras con la media de la distancia de las palabras al centroide, es porque en el caso de los centroides, reducimos la multidimensionalidad de 300 dimensiones a una para calcular la distancia euclidiana al principio, mientras que cuando utilizamos las medias de las distancias, lo hacemos al final. El objetivo es comprobar si hay perdida de información en el paso de la multidimensionalidad a la unidimensionalidad.

Una vez hemos establecido las medidas que utilizamos para analizar y comparar los textos, y antes de proceder a detallar la experimentación, podemos ver en las figuras 8 y 9 donde, para facilitar la visualización, representamos como ejemplos, las distancias de una agrupación de palabras de un texto procesado, ordenadas de menor a mayor de la distancia al centroide y del centroide al origen. Para ello hemos utilizado dos textos distintos generados por Bard.

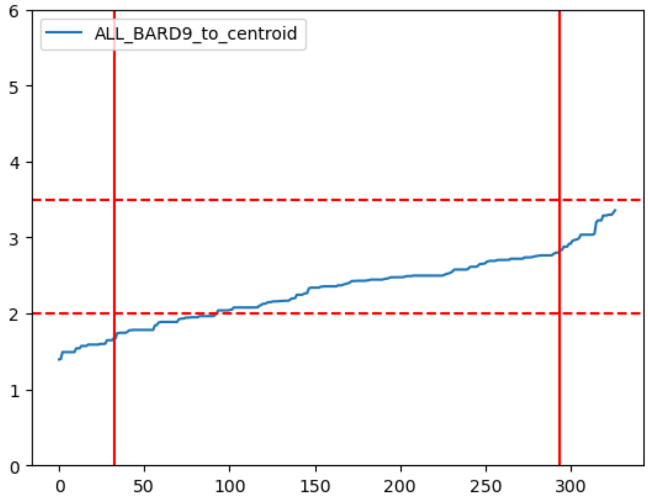

Figura 8: Esquema distancia al centroide

En el eje de ordenadas vemos la distancia euclidiana, mientras que en el eje de abscisas apreciamos el número de palabras utilizadas para el análisis una vez procesado y limpiado el texto.

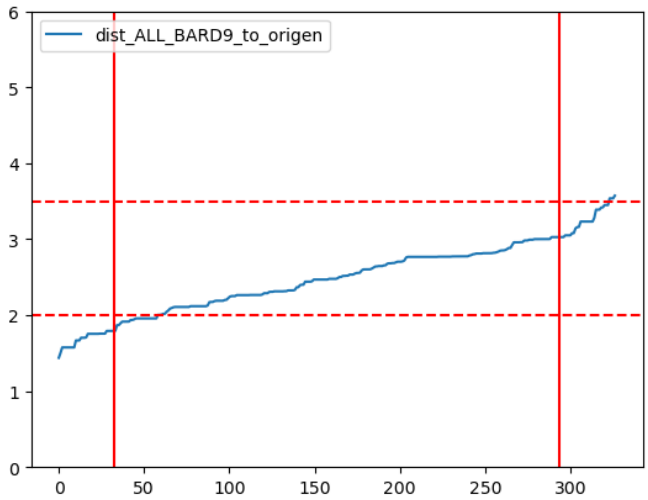

Figura 9: Esquema distancia al origen

Lo que podemos apreciar en esta representación es que al haber eliminado las stopwords, no tenemos resultados cercanos a cero y que al analizar las palabras individualmente, la distancia que nos muestra la especificidad, no se alejan de una manera desproporcionada.

5.1. Similitud Coseno

Analizando los resultados obtenidos, en el cuadro de la figura 10, similitud coseno, podemos observar resaltados los tres grupos de autores. Tomando como referencia el grupo de periodistas de la CNN, vemos como los dos cuartiles centrales, es decir, entre el 25 % y 75 %, se mantienen un rango de variación del 10 % y en posiciones similares. A pesar de que, como hemos explicado anteriormente, sus temáticas son dispares, apreciamos homogeneidad tanto en la dispersión de sus resultados como en la situación de la mediana de cada grupo.

Cuando observamos el segundo grupo de autores, vemos que los dos cuartiles centrales tienen un rango mayor de alrededor del 20 % y además, entre ellos existe mas variedad, llegando a no solaparse en algunos casos. La medida de la similitud coseno es un 6 % mayor, hasta el 0,38, que el grupo de control. En este caso, la dispersión de sus resultados es mayor, habiendo grandes diferencias entre varios de los autores como son Dickens, Hume y Kant. Comparando este segundo grupo con el de los periodistas, nos reafirmamos en la validez de este último como grupo de control.

Finalmente, podemos apreciar como el grupo de textos generados por la IA se diferencian respecto a los otros dos, no solo en que el rango de los dos cuartiles es de un 25 %, sino que la medida en el caso más extremo es de un 15 % mayor, hasta el 0,41, que los escritores y un 21 % mayor que el grupo de control. Si apreciamos que, como grupo, las distancias son mayores que los otros dos, pero todavía no somos capaces de sacar conclusiones significativas.

Como hemos comentado en los objetivos, la similitud coseno, por sí sola, no nos desglosa el ángulo y la norma, por lo que debemos seguir con la experimentación para obtener esos resultados por separado y comentar la experimentación en conjunto.

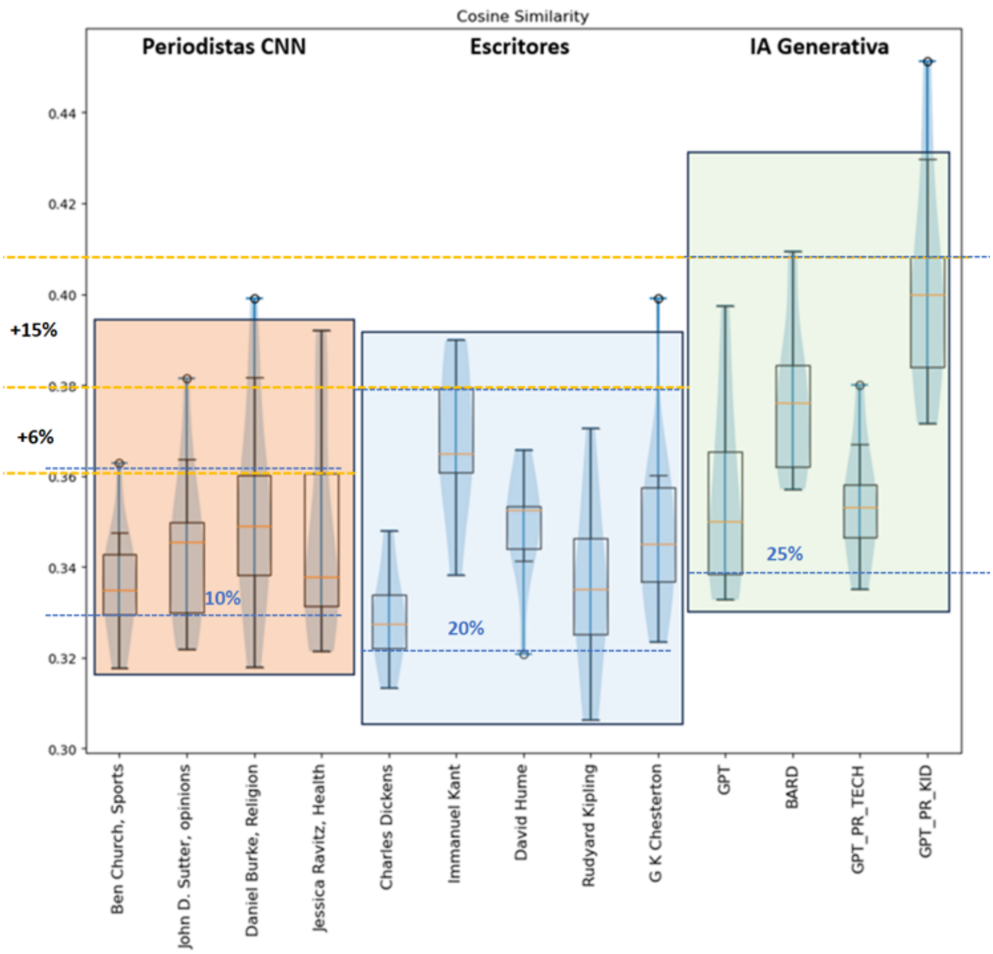

Figura 10: Similitud coseno

5.2. Distancia al centroide

En la figura 11, medimos la distancia al centroide, Esta medida nos va a permitir medir el grado de dispersión de la representación geométrica de los vectores respecto al centroide y refleja cómo de variado o heterogéneo es el vocabulario empleado.

Se aprecian mayores diferencias entre la IA generativa y los dos grupos de textos generados por seres humanos. En este caso se repite el mismo patrón que en la similitud coseno, donde los textos de los periodistas son más similares entre ellos, y que los textos de los escritores clásicos varían más entre ellos. Mientras que el rango de apertura de los cuartiles centrales de los dos primeros grupos es de alrededor de un 10 %, ligeramente por encima de 2,5, el de los textos de la IA generativa es de un

25 %, por encima de 2,8, llegando a tener la mínima distancia y la máxima, siendo la máxima un 10 % mayor respecto a los otros dos grupos. Los textos que marcan la máxima y la mínima son los textos generados por la IA direccionando el prompt técnico y el prompt a modo de un niño.

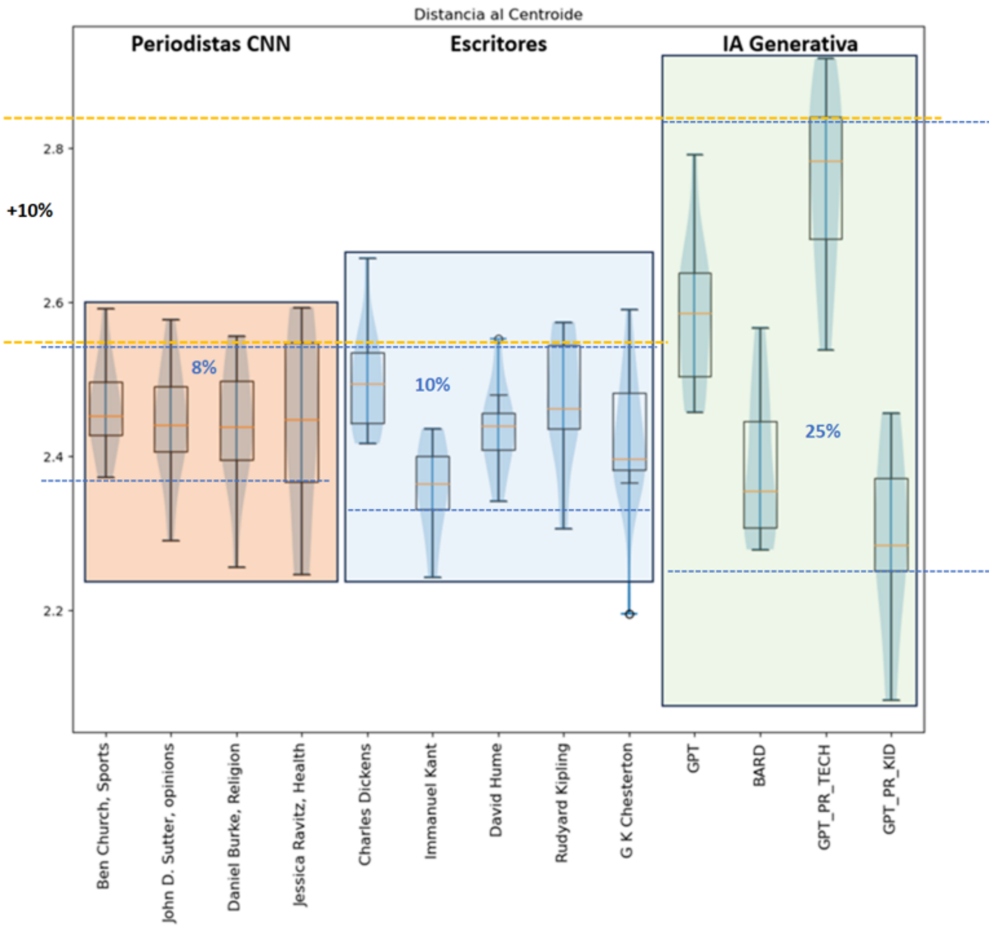

Figura 11: Distancia al Centroide

Si que podemos ver de una forma clara, cómo cada grupo se comporta de una manera característica respecto a lo otros y que deberíamos poder caracterizar según vayamos añadiendo nuevas medidas. Además, apreciamos cómo los textos de los escritores y de la IAG, se diferencian unos de otros, dentro de cada grupo, y empezamos a ver indicios de singularidad individual que permita su caracterización.

5.3. Distancia media entre todas las palabras

En la figura 12, la distancia media entre todas las palabras nos permite evaluar cuán similares o cercanas están un conjunto de palabras o vectores de palabras entre sí en un espacio multidimensional.

Vemos como las diferencias entre los grupos son parecidas de tal forma que el grupo de periodistas se mantiene en un rango de un 8 % y que el grupo de autores clásicos, aunque mantiene la dispersión entre ellos, su rango es de un 10 %, similar a los periodistas.

Cuando nos fijamos en el grupo de textos generados por la inteligencia artificial generativa, vemos que la dispersión es mucho mayor, hasta el 28 %, siendo los valores extremos el prompt técnico y el prompt de lenguaje infantil.

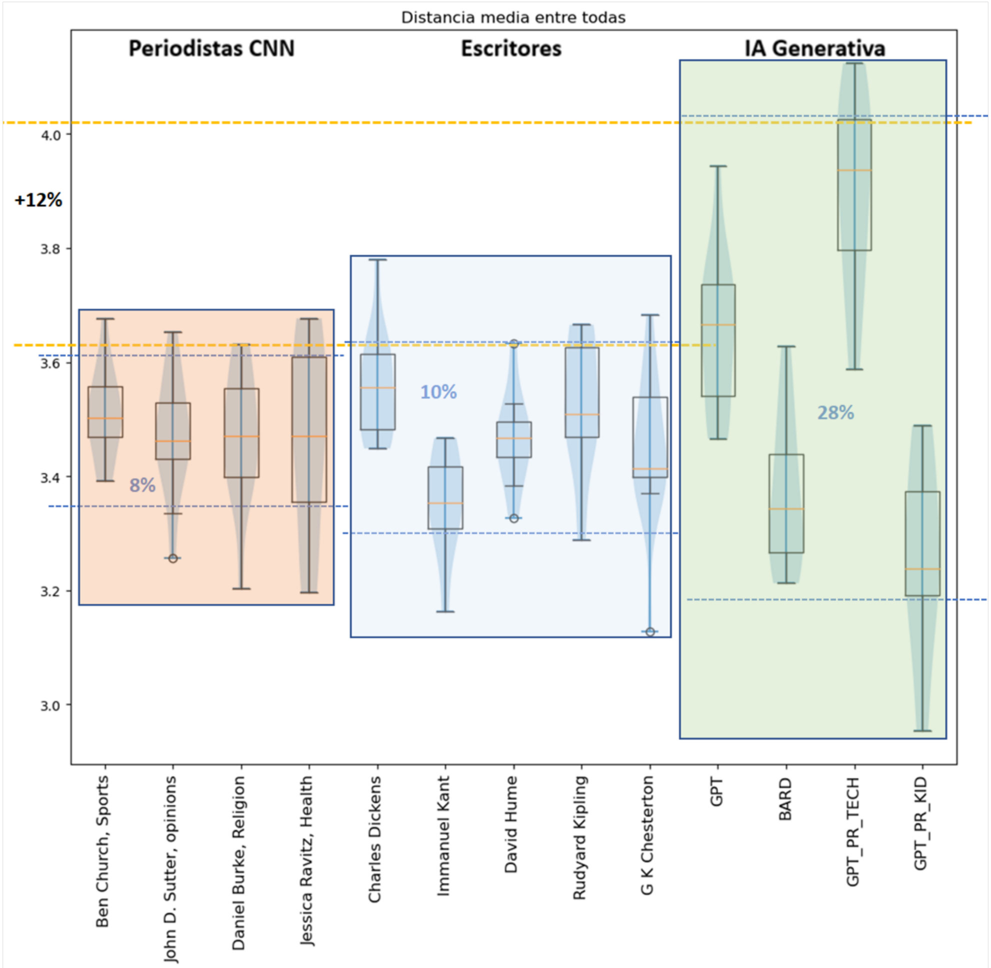

Figura 12: Distancia media entre todas

Utilizamos esta medida, similar a la distancia al centroide, para verificar que los resultados son coherentes entre sí y que la pérdida de información en el proceso de cálculo, cuando reducimos los vectores de 300 dimensiones a un vector unidimensional, no distorsionan los resultados. Comprobamos que los resultados son similares entre la figura 11 y 12, por lo que las dos medidas son válidas y la posible pérdida de información no distorsiona los resultados finales.

5.4. Distancia del centroide al origen

En la figura 13, distancia del centroide al origen, medimos la distancia entre el punto central que representa la media de un conjunto de vectores de palabras (centroide) y el punto de referencia que es el origen de coordenadas en ese espacio vectorial. De esta forma, reflejamos la especificidad de los textos según la distancia euclidiana del centroide al origen. En este caso observamos como en el grupo de control existe un poco mas de variabilidad entre ellos y puede deberse a la temática, aunque no parece demasiado significativo.

En el grupo de escritores clásicos se mantiene, en comparación al grupo de control, el rango de especificidad, pero entre ellos apreciamos características singulares de cada autor. Por ejemplo, Dickens es menos específico en general, a lo largo de toda su obra, mientras Kant lo es algo más, pero también es muy homogéneo a lo largo de toda su obra.

En cambio, cuando observamos el grupo de textos generados por IA, vemos que siempre son más específicos que los generados por los humanos, tanto periodistas como escritores, siendo el prompt técnico el más específico, tal y como era de esperar. Sin embargo, en el prompt con lenguaje de niño, sigue siendo más específico que los textos por humanos, lo que llama la atención ya que debería emplear vocabulario más común.

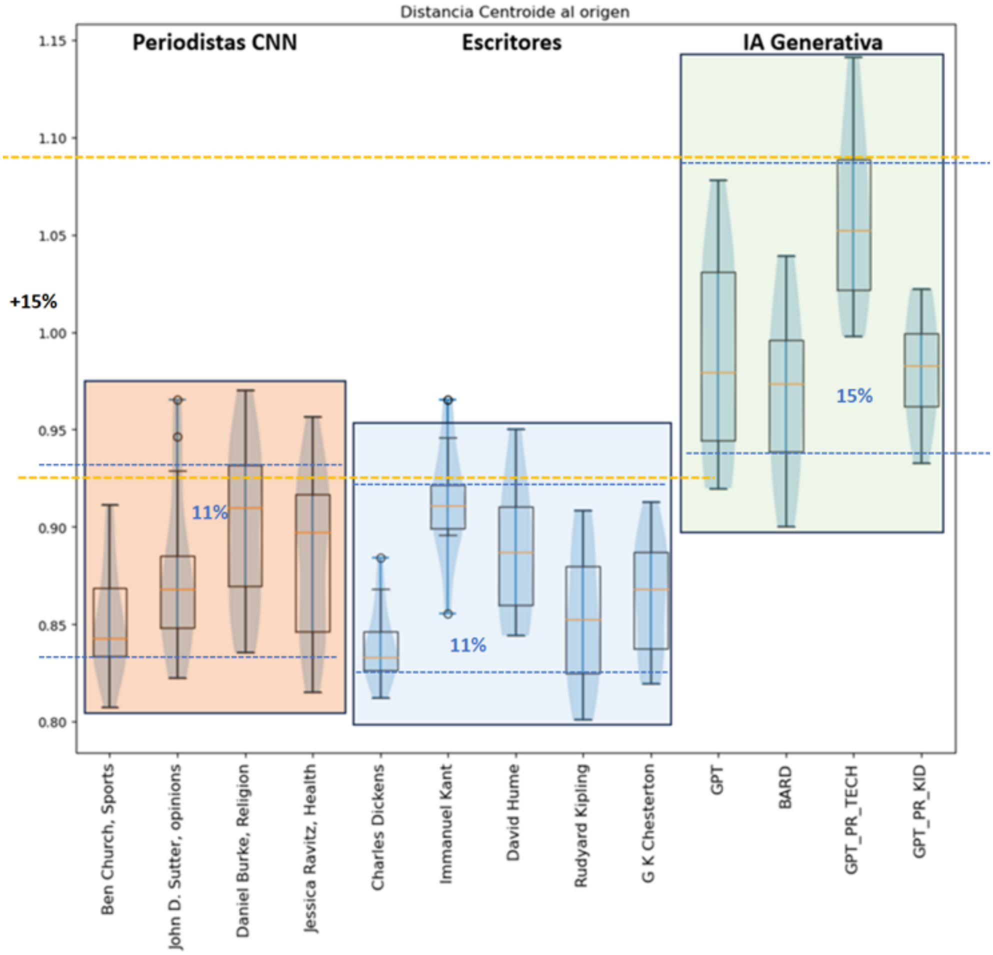

Figura 13: Distancia Centroide al origen

5.5. Distancia media de las palabras al origen

En la figura 14, utilizamos la medida de la distancia media de las palabras al origen como complemento a la medida del centroide al origen y observamos como el grupo de periodistas se mueve en un margen similar al de los autores clásicos, aunque estos últimos son menos homogéneos entre ellos, llegando a no solaparse en algunos casos como Dickens y Kipling con Kant.

Respecto a los texto generados por IA generativa, volvemos a ver como su dispersión en mucho mayor respecto a los otros dos grupos y entre ellos. En este caso, el prompt técnico es el que mayor especificidad refleja en los resultados, mientras que el prompt con lenguaje infantil es el que menor especificidad muestra, no solo respecto al grupo de IA, sino respecto a los otros dos grupos. Esto tiene sentido respecto al análisis del centroide respecto al origen, ya que, como podemos observar en la figura 7, el centro de coordenadas y el centroide son dos puntos geométricos en el espacio desplazados entre sí. En este caso, los resultados del análisis de la distancia media al origen reflejan una situación acorde con lo que podríamos esperar respecto a la especificidad.

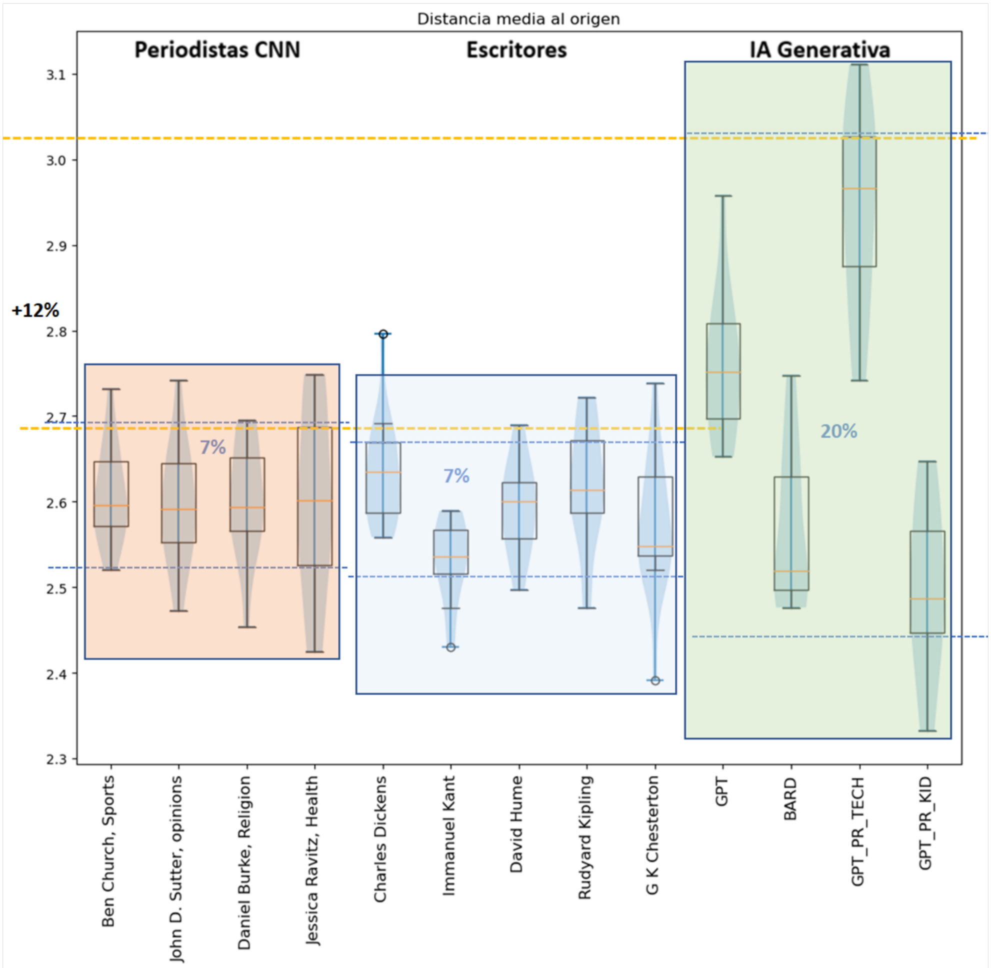

Figura 14: Distancia media al origen

6. Discusión

La primera cuestión que planteábamos al comienzo de esta investigación es si se pueden extraer características específicas de un texto basado en la orientación semántica y en la especificidad del vocabulario dentro de espacios de representación de embeddings estáticos. Si nos fijamos en la figura 10 similitud coseno podríamos concluir que los textos de David Hume y los de GPT PR Tech son similares o incluso pertenecen al mismo autor, pero cuando los desglosamos en especificidad y distribución semántica, vemos cómo estas características son mucho mayores en los textos generados por GPT que los generados por Hume.

Para contestar la cuestión de si es la riqueza del vocabulario utilizado y su significado lo que determina la caracterización de un texto, podemos observar como los 3 grupos se comportan de diferente forma entre ellos, pero según un patrón similar dentro de cada uno. Podemos observar como los periodistas, independientemente de la temática de sus textos, ofrecen resultados similares no solo con el coseno, sino también con la especificidad y la distribución semántica y podemos concluir cómo la forma de dirigirte a una audiencia determinada y sobre todo, a una audiencia amplia y multitudinaria requerirá que tu vocabulario sea menos específico. Esto lo hemos comprobado cuando hemos pedido a la IAG que redactará textos técnicos (GPT PR TECH) y textos para niños (GPT PR KID) donde no solo ha sido más específico con el modo técnico, sino la variedad de vocabulario ha sido significativamente menor en el modo niño.

También hemos demostrado con la experimentación que los textos generados por inteligencia artificial generativa tienen un perfil determinado que les caracteriza y que les diferencia respecto al grupo de control y el grupo de escritores con mayores distancias entre ellas y al centroide lo que denota una mayor variedad de vocabulario y una mayor especificidad, especialmente cuando le pedimos que utilice un lenguaje más técnico. Podemos concluir que se pueden distinguir los textos creados por inteligencia artificial generativa de los creados por seres humanos según la caracterización de cada texto en base a la especificidad y similitud, pero también observamos que se pueden orientar los textos mediante prompts para que sus estilos se adapten a lo que les pedimos.

7. Conclusiones y futuros desarrollos

A lo largo de este trabajo hemos empleado diversas medidas de similitud sobre word embeddings para estudiar si se puede caracterizar la forma en que escriben diversos autores. Hemos podido comprobar que, aunque el uso de las similitud coseno es un estándar utilizado para medir y comparar la similitud de diferentes textos, no nos permite analizar toda la información que recoge. Nos referimos a la especificidad y la distribución semántica. La similitud coseno es una combinación de las dos y sus rasgos de valor están entre 0 y 1. Cuando desagregamos los dos valores que componen la similitud coseno que son la norma y el ángulo del vector que representa a la palabra, se demuestra que, dentro de espacios de representación de embeddings estáticos, se puede extraer características específicas de un texto basado en la orientación semántica y en la especificidad del vocabulario.

El poder medir la especificidad nos aporta una información muy relevante de cómo de específico es el vocabulario empleado, siendo de gran importancia el haber demostrado que esta variable caracteriza a un autor incluso midiendo solo palabras individuales y no grupos de palabras. Si a la especificidad le añadimos la distribución semántica al análisis, podemos entender cómo de amplio y variado es el vocabulario empleado.

Por último, demostramos cómo los textos generados en IA generativa destacan frente a los redactados por humanos en mayor grado de especificidad y como destacan por su capacidad de adaptación a la forma de comunicarse y modelar el mensaje cuando se lo hemos pedido y por tanto, tienen un perfil determinado que ayuda a caracterizarlos individualmente y que permite distinguirlos frente a los textos generados por seres humanos.

Aunque estas características de los textos similitud coseno, especificidad y distribución semántica, no son concluyentes por sí solas para caracterizar e individualizar a un autor o a una IA generativa, sí que pueden ser un complemento que mejore la precisión y soporte en su análisis. Su aplicación no solo se limitaría al análisis de autoría, detección de identidad o plagio, sino que podría emplearse, por ejemplo, en el ámbito de la educación para evaluar el nivel de desarrollo lingüístico de un alumno, medición de progreso y mejora de un estudiante, conectar y adaptar los textos a un nivel de comprensión adecuado que permita la personalización,...

Para futuros desarrollos y teniendo en cuenta lo rápido que evolucionan y salen al mercado nuevos modelos fundacionales como ChatGPT 4, Bard[25] o Amazon Q, habría que crear un corpus más extenso de textos generados por las distintas IA generativas para entender si se podrían caracterizar individualmente, lo que tendría lógica ya que tanto su tamaño, arquitectura y entrenamiento son distintos y deberían parecer, por tanto, autores diferentes. Además y por lógica, cuanto más evolucionados estén, serán más capaces de adoptar caracterizaciones específicas, técnicas, generales, simples,... en base a las peticiones que les hagamos con los prompts, lo cual nos lleva a abrir otra línea de investigación que sería comprobar si serían capaces de imitar una caracterización de un autor humano específico de tal manera que los haga indistinguibles, con el peligro que una suplantación de identidad tan precisa acarrearía.

8. Bibliografía

Referencias

[Posadas et al (2015)] J.-P. Posadas-Durán, I. Markov, H. Gómez-Adorno, G. Sidorov, I. Batyrshin, A. Gelbukh, O. Pichardo-Lagunas: *"Syntactic n-grams as features for the author profiling task"*, Working Notes Papers of the CLEF, (2015)

[Zhang et al (2018)] R. Zhang, Z. Hu, H. Guo, Y. Mao: *"Syntax encoding with application in authorship attribution"*, Conference on Empirical Methods in Natural Language Processing, (2018)

[Sari et al (2018)] Y. Sari, M. Stevenson, A. Vlachos: *"Topic or style? exploring the most useful features for authorship attribution"*, Proceedings of the 27th International Conference on Computational Linguistics, (2018)

[25] En el cierre de este estudio, Google anuncia el lanzamiento de Gemini como evolución de Bard

[Hinh et al (2016)] R. Hinh, S. Shin and J. Taylor: *“Using frame semantics in authorship attribution”*, 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, (2016)

[Argamon et al (2003)] S. Argamon, M. Koppel, J. Fine and A.R. Shimoni: *"Gender, genre, and writing style in formal written texts"* Text-The Hague Then Amsterdam Then Berlin (2003)

[Schler et al (2006)] J. Schler, M. Koppel, S. Argamon and J.W. Pennebaker: *.Effects of age and gender on blogging"* In AAAI spring symposium: Computational approaches to analyzing weblogs, volume 6, (2006)

[Goswami et al (2009)] S. Goswami, S. Sarkar and M. Rustagi: *"Stylometric analysis of bloggers' age and gender"* In Third International AAAI Conference on Weblogs and Social media, (2009).

[Meina et al (2013)] M. Meina, B. Celmer, M. Czokow, J. Pezacki and M.Wilk: *.Ensemble-based classification for author profiling using various features"* the notebook for, In 2013 PAN at the Conference and Labs of the Evaluation Forum (CLEF, (2013)

[Bougiatiotis and Krithara (2016)] K. Bougiatiotis and A. Krithara: *.Author profiling using complementary second order attributes and stylometric features"* In CLEF (Working Notes), (2016)

[Basile et al (2017)] A. Basile, G. Dwyer,M.Medvedeva, J. Rawee, H. Haagsma and M. Nissim: *Ïs there life beyond n-grams? a simple svmbased author profiling system"* In Working Notes of CLEF 2017-Conference and Labs of the EvaluationForum, (2017)

[Guimaraes et al (2017)] R. Guimaraes, R. Rosa, D. De Gaetano, D.Z. Rodriguez and G. Bressan: *.Age groups classification in social network using deep learning"* IEEE Access, 5 (2017)

[Miura et al (2017)] Y. Miura, T. Taniguchi, M. Taniguchi and T. Ohkuma: *.Author profiling with word+ character neural attention network"* In CLEF (Working Notes),

(2017)

[Raiyani et al (2018)] K. Raiyani, T. Gonc¸alves, P. Quaresma and V. Nogueira: *"Multi-language neural network model with advance preprocessor for gender classification over social media"* In CLEF (Working Notes). CLEF'2018, (2018)

[Takahashi et al (2018)] T. Takahashi, T. Tahara, K. Nagatani,Y. Miura, T. Taniguchi and T. Ohkuma: *"Text and image synergy with feature cross technique for gender identification"* Working Notes Papers of the CLEF, (2018)

[Stamatatos et al (2014)] Stamatatos, E.; Daelemans, W.; Verhoeven, B.; Potthast, M.; Stein, B.; Juola, P.; Sanchez-Perez, M.A.; Barrón-Cedeño, A: *.Overview of the Author Identification Task at PAN 2014"* CLEF 2014

[Koppel and Winter(2014)] Koppel, M.; Winter, Y: *"Determining If Two Documents Are Written by the Same Author"* Assoc. Inf. Sci. Technol, (2014)

[Koppel et al (2007)] Koppel, M.; Schler, J.; Bonchek-Dokow, E.; Dokow, B: *"Measuring Differentiability: Unmasking Pseudonymous Authors"* J. Mach. Learn. Res. 2007

[Bevendorff et al (2019)] Bevendorff, J.; Stein, B.; Hagen, M.; Potthast, M: *"Generalizing Unmasking for Short Texts"* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, (2019)

[Koppel et al (2011)] Koppel, M.; Schler, J.; Argamon, S: *.Authorship Attribution in the Wild"* Working Notes Papers of the CLEF, (2011)

[Bagnall (2015)] Bagnall, D: *.Author Identification Using Multi-Headed Recurrent Neural Networks"* arXiv 2015, arXiv:1506.04891.

[Jafariakinabad et al (2019)] Jafariakinabad, F.; Tarnpradab, S.; Hua, K.A: *"Syntactic Recurrent Neural Network for Authorship Attribution"* arXiv 2019, arXiv:1902.09723

[Weerasinghe and Greenstadt (2020)] Weerasinghe, J.; Greenstadt, R: *"Feature Vector Difference Based Neural Network and Logistic Regression Models for Authorship Verification"* In Notebook for PAN at CLEF 2020

[Kou et al (2020)] G. Kou, P. Yang, Y.Peng, F.Xiao, Y.Chen: *"Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods"*, Appl. Soft Comput 86. (2020)

[Ahmed (2018)] Hossam Ahmed: *"The role of linguistic feature categories in authorship verification"*, The 4th international conference on arabic computational linguistics. (2018)

[Mikolov (2012)] Tomas Mikolov: *"Statistical Language Models Based on Neural Networks. PhD thesis"*, PhD thesis, PhD Thesis, Brno University of Technology. (2012)

[Mikolov et al (2013a)] Mikolov, T., Chen, K., Corrado, G., and Dean, J: *"Efficient estimation of word representations in vector space"*. International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4,Workshop Track Proceedings (2013a)

[Mikolov et al (2013b)] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. *"Distributed representations of words and phrases and their compositionality"* in Proceedings of the 26th International Conference on Neural Information Processing Systems (2013b)

[Ethayarajh (2019)] Kawin Ethayarajh: *"How contextualized word representations? Comparing the geometry of Bert, ELMo and GPT-2 embeddings"*, Stanford University. (2019)

[Peters et al (2018)] Peters, Matthew. Neumann, Mark. Iyyer, Mohit. Gardner, Matt, Clark, Christopher. Lee, Kenton. Zettlemoyer, Luke: *"Deep contextualized word representations"*, Association for Computational Linguistics. (2018)

[Devlin et al (2018)] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understan-*

ding", Google AI Language. (2018)

[Gao et al (2019)] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang and Tie-Yan Liu: -*epresentation degeneration problem in training natural language generation models"*, Conference paper at ICLR. (2019)

[Li et al (2020)] Li, Bohan, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang and Lei Li: *.On the sentence embeddings from pre-trained language models"*, Conference on EMNLP. (2020)

[Cai et al (2021)] Cai, Xingyu, Jiaji Huang, Yuchen Bian and Kenneth Church: *Ïsotropy in the contextual embedding space: Clusters and manifolds"*, In international conference on learning representations. (2021)

[Wu et al (2020)] Wu, Jhon, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi and James Glass: *"Similarity analysis of contextual representations models"*, In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. (2020)

[Amigó et al (2022)] Enrique Amigó, Alejandro Ariza-Casabona, Víctor Fresno y M. Antònia Martí: *Ïnformation Theory-based compositional distributional semantics"*, Association for computational Linguistics. (2022)

[Johns (2023)] Brendan T. Johns: *Çomputing word meanings by aggregating individualized distributional models: Wisdom of the crowds in lexical semantic memory"*, Cognitive Systems Research. (2023)

[Johns et al (2018)] Brendan T. Johns, Randall K. Jamieson: *.A Large-Scale Analysis of Variance in Written Language"*, Cognitive Science. (2018)

[Skrlj et al (2019)] Blaz Skrlj, Jan Kralj, Nada Lavrac and Senja Pollak: *"Towards robust text classification with semantics-aware recurrent neural architecture"*, MDPI journals (2019)

[Zhou et al (2023)] Ce Zhou, Qian Li, Chen Li, Jun Yu et al: *.A comprehensive survey on pretrainedfoundation models: A history from Bert to ChatGPT"*, (2023) [Muntina et al (2022)] Eddy Muntina Dharma, Ford Lumban Gaol, Harco Leslie Hendric Spits Warnars, Benfano Soewito: *"The accuracy comparison among Word2vec, glove, and fastText towards convolution neural network (CNN) text classification"*,Journal of Theoretical and Applied Information Technology (2022)

[Schulte and Frassinelli (2022)] Sabine Schulte im Walde and Diego Frassinelli: *"Distributional measures of semantic abstraction"*, Frontiers in Artificial Intelligence (2022)

[Johns and Jones (2022)] Johns, B. T., Jones, M. N: *Çontent matters: Measures of contextual diversity must consider semantic content"* Journal of Memory and Language, 123, Article (2022)

[Levy and Goldberg (2014)] Levy, O., Goldberg, Y.: *"Neural word embedding as implicit matrix factorization."* ,In Advances in Neural Information Processing Systems (pp. 2177-2185).(2014)

[Levy et al (2015)] Levy, O., Goldberg, Y., Dagan, I.: *Ïmproving distributional similarity with lessons learned from word embedding."* Transactions of the Association for Computational Linguistics, 3, 211–225.(2015)

[Wang et al (2019)] Suhang Wang, Charu Aggarwal, Huan Liu: *"Beyond word2vec: Distance-graph Tensor Factorization for Word and Document Embeddings"* In The 28th ACM International Conference on Information and Knowledge Management.(2019)

[Sheinman et al (2023)] Michael Sheinman Orenstrakh, Oscar Karnalim, Carlos Aníbal Suárez, Michael Liut: *"Detecting LLM-Generated Text in Computing Education: A Comparative Study for ChatGPT Cases"* ACM.(2023)

[Liang et al (2021)] Liang, Y., Cao, R., Zheng, J., Ren, J., Gao, L.: *"Learning to remove: Towards isotropic pre-trained BERT embedding"* In International Conference on Artificial Neural Networks.(2021)

[Halteren et al (2005)] H. van Halteren, H. Baayen, F. Tweedie, M. Haverkort, A. Neijt: *New machine learning methods demonstrate the existence of a human stylome*, Journal of Quantitative Linguistics 12 (2005)

[Wu et al (2021)] Haiyan Wu, Zhiqiang Zhang, Qingfeng Wu: *Exploring syntactic and semantic features for authorship attribution*, Applied Soft Computing, Elsevier (2021)

[Wang et al (2014)] Q. Wang, J. Zhang, S. Song, Z. Zhang: *Attentional neural network: Feature selection using cognitive feedback*, in: Advances in Neural Information Processing Systems, (2014)

[Gui et al (2019)] N. Gui, D. Ge, Z. Hu, AFS: *An attention-based mechanism for supervised feature selection* in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, (2019)

[Haiyan Wu et al (2021)] Haiyan Wu, Zhiqiang Zhang, Qingfeng Wu: *Exploring syntactic and semantic features for authorship attribution* in: Applied Soft Computing, Elsevier, (2021)

[Vaswani et al (2017)] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. *Attention is all you need* in Advances in Neural Information Processing Systems, (2017)

[Dosovitskiy et al (2021)] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. *An image is worth 16x16 words: transformers for image recognition at scale* in Proceedings of the International Conference on Learning Representations ICRL (2021)

[Harris (1954)] Harris, Z. S. *Distributional structure*. Word 10, 146–162. (1954)

[Alshaabi et al (2022)] Thayer Alshaabi, Colin M. Van Oort, Mikaela Irene Fudolig, Michael V. Arnold, Christopher M. Danforth and Peter Sheridan Dodds. *Augmenting Semantic Lexicons Using Word Embeddings and Transfer Learning*. Front. Artif. Intell., Sec. Language and Computation (2022)

[Pennington et al (2014)] Jeffrey Pennington, Richard Socher, Christopher D. Manning. *GloVe: Global Vectors for Word Representation*. in Proceedings of the

2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)(2014)

[Joulin et al (2017)] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov. *Bag of Tricks for Efficient Text Classification*. in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics(2017)

[Bojanowski et al (2017)] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov. *Enriching Word Vectors with Subword Information*. Transactions of the Association for Computational Linguistics, vol. 5 (2017)

[Arora et al (2016)] Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. *A latent variable model approach to PMI-based word embeddings*. Transactions of the Association for Computational Linguistics, (2016)

[Xu and Koehn (2021)] Haoran Xu and Philipp Koehn. *Cross-Lingual BERT Contextual Embedding Space Mapping with Isotropic and Isometric Conditions*. arXiv:2107.09186, (2021)

[Oyama et al (2023)] Momose Oyama, Sho Yokoi, Hidetoshi Shimodaira. *Norm of Word Embedding Encodes Information Gain*. arXiv:2212.09663v3, (2023)

[Yokoi et al (2020)] Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki and Kentaro Inui. *Word Rotator's Distance*. Conference on Empirical Methods in Natural Language Processing (EMNLP), (2020)

[Agirre et al (2012)] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. *A Pilot on Semantic Textual Similarity*. SemEval-2012 Task 6, (2012)

[Schakel and Wilson (2015)] Adriaan M. J. Schakel and Benjamin J. Wilson. *Measuring word significance using distributed representations of words*. ArXiv 1508.02297., (2015)

[Mitchell and Lapata (2010)] Jeff Mitchell and Mirella Lapata. *Composition in distributional models of semantics*. Cognitive Science, 34(8):1388–1429, (2010) [Arefyev et al (2018)] Nikolay Arefyev, Pavel Ermolaev, and Alexander Panchenko. *How much does a word weigh? weighting word embeddings for word sense induction*. ArXiv 1805.09209, (2018)

[Ladani et al (2020)] Ladani, Dhara and Nikita Desai. *Stop word Identification and Removal Techniques on TC and IR applications*. A Survey. 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 466–472. https:// doi.org/10.1109/ICACCS48705.2020.9074166, (2020)