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

Análisis del lenguaje en estados alterados de conciencia

Tesis presentada para optar por el título de Doctora de la Universidad de Buenos Aires en el área Ciencias Físicas

Lic. Camila Sanz

Director de tesis: Dr. Enzo Tagliazucchi
Consejero de estudios: Dr. Carlos Acha

Lugar de trabajo: Laboratorio de Conciencia Cultura y Complejidad

25/03/2024, Buenos Aires.

"I'm not young enough to know everything"

Oscar Wilde

Resumen

Análisis del lenguaje en estados alterados de conciencia

En esta tesis aplicamos y desarrollamos herramientas de procesamiento de lenguaje natural (NLP) para identificar estados cognitivos alterados producidos tanto por sustancias psicoactivas como por enfermedades neurodegenerativas. En particular, estudiamos las alteraciones que el LSD y la microdosificación con hongos Psilocybe cubensis generaban en el lenguaje, las comparaciones se realizaron con una condición de placebo. En el estudio de las enfermedades neurodegenerativas, nos enfocamos en la enfermedad de Alzheimer (EA) que actualmente resulta preocupante tanto por el costo económico que implica como por el deterioro en la calidad de vida de los pacientes y sus allegados. Las comparaciones se realizaron con un grupo de control y pacientes diagnosticados con la enfermedad de Parkinson (EP) para determinar la especifidad de nuestros hallazgos.

Todas las publicaciones que resultaron de esta tesis presentan una estructura similar. En primer lugar identificamos características lingüísticas que reflejan aspectos cognitivos del estado estudiado. En la investigación del discurso bajo los efectos del LSD, nos centramos en caracterizar la desorganización del lenguaje. En el estudio de la microdosificación, nos enfocamos en lograr una cuantificación objetiva de los efectos que suelen reportarse por los usuarios de esta práctica pero carecen de sustento experimental. En la investigación de enfermedades neurodegenerativas, desarrollamos métodos automatizados y objetivos para medir efectos de la EA con sustento clínico. Una vez identificadas las características de interés, aplicamos análisis estadísticos y algoritmos de *Machine Learning* (ML) para determinar si las métricas consideradas resultaban distintivas de los grupos considerados.

PALABRAS CLAVE: Estados cognitivos alterados, Aprendizaje Automático, Procesamiento de Lenguaje Natural, Sustancias psicodélicas, Enfermedades neurodegenerativas.

Abstract

Language analysis in altered states of consciousness

We applied and developed Natural Language Processing (NLP) tools to identify altered cognitive states produced by both psychoactive substances and neurodegenerative diseases. Specifically, we studied changes in natural speech under the effects of LSD and microdosing with Psilocybe cubensis; comparisons were made with a placebo condition. In the research of neurodegenerative disease, we focused on Alzheimer disease (AD), which is becoming a major problem due to the growing economic burden and the decreasing life quality of patients and their relatives. Comparisons were made with a control group and patients diagnosed with Parkinson's disease (PD) to determine the specificity of our results.

Publications arise from this thesis share a similar structure. First, we identified linguistic features that mirror cognitive aspects of the studied state. Our research on speech under LSD was based on characterizing the disorganization of language. While the microdosing study focused on achieving an objective quantification of effects that are often reported by users but lack experimental support. In the research of neurodegenerative diseases, we develop automated and objective methods to measure clinical supported effects of AD. Once the characteristics of interest were identified, we applied statistical analyses and Machine Learning (ML) algorithms to determine whether the metrics were distinctive for the groups under consideration.

KEY WORDS: Cognitive states, Machine Learning, Natural Language Processing, Psychoactive substances, Neurodegenerative diseases.

Agradecimientos

Agradezco al CONICET por otorgarme la Beca de Doctorado que hizo posible esta investigación. También quiero darle las gracias a la Facultad de Ciencias Exactas y Naturales y a sus profesores, que me brindaron una excelente formación académica totalmente gratuita. A Enzo Taglizucchi que me acompañó y asesoró desde mi Tesis de Licenciatura hasta el día de hoy. Si tuviese que agradecerle a Enzo por todo lo que hizo por mi durante estos años, los agradecimientos resultarían más largos que la tesis. Por ende, me limito a decirle gracias por convencerme de hacer el Doctorado, crear una ambiente muy cálido para trabajar, tener infinita paciencia y siempre haber estado a mi lado para guiarme. Quería aprovechar para agradecerle a Facu Carrillo que fue mi segundo mentor durante el Doctorado que me brindó mucha ayuda y supervisó todas mis publicaciones. Soy muy afortunada por haber recibido siempre el apoyo de mi familia en todo lo que quise hacer, en particular hago una mención especial a mi mamá que me aconsejó, acompañó e incetivó para terminar esta Tesis. También quiero resaltar el apoyo de mi pareja que siempre está para ayudarme cuando no puedo sola. Por último, quería dedicar parte de estos agradecimientos a mis compañeros de laboratorio, que me acompañaron durante todo el trayecto del Doctorado y lograron que COCUCO sea un ambiente familiar y amistoso.

Índice general

Capítulo 1

Introducción

El lenguaje es un fenómeno social que asigna un significado a determinados sonidos y signos, permitiendo que los individuos puedan comunicarse entre sí (Lewis, 1975). A diferencia de una teoría matemática que funda sus bases en axiomas y sus resultados en demostraciones, se estipula que el proceso de formación del lenguaje fue inverso. En principio constituía un sistema desestructurado y tomó forma por medio de regularidades, es decir, por sonidos, gestos y signos dependientes del estado cognitivo del emisor (Lewis, 1975). Finalmente, estas regularidades se formalizaron como convenciones que esbozaron la estructura del lenguaje. Hoy en día, las antiguas convenciones lingüísticas se precisan con reglas sintácticas y gramaticales. Este fenómeno es un sistema dinámico, donde continuamente se excluyen/incluyen elementos y el significado de los mismos muta según los sucesos culturales o históricos independientemente de los individuos particulares (De Saussure et al., 1987). Dado este conjunto de reglas y relaciones conceptuales, el lenguaje se constituye como una expresión del pensamiento, por lo tanto, el estudio del mismo permite (hasta cierto punto) analizar el estado cognitivo de los individuos.

Múltiples sucesos contribuyeron al desarrollo de técnicas computacionales para estudiar el lenguaje, los más relevantes corresponden al incremento de la potencia de cómputo, el almacenamiento/producción masiva de datos (Hirschberg and Manning, 2015) y el acceso público a esta información provisto por Internet. De esta forma, alrededor de los 70's surgió el "Procesamiento del Lenguaje Natural" (NLP); una rama de la computación dedicada al estudio del lenguaje escrito con métodos objetivos, automatizados y escalables (Khurana et al., 2022). Dentro de este campo, se pueden destacar diferentes niveles (Liddy, 2001) de comprensión lingüística, en esta tesis nos enfocaremos en tres: (1) El semántico: permite la interpretación de frases analizando el significado de las palabras y sus relaciones tanto conceptuales como contextuales; (2) El sintáctico: corresponde al análisis estructural de las frases; (3) El discursivo: analiza documentos enteros basado

en las propiedades y conexiones de sus partes. Para ejemplificar las distinciones entre niveles consideremos las frases "Fui al banco de la plaza" y "Fui al banco a sacar dinero", sintácticamente las oraciones son equivalentes, sin embargo, el análisis semántico de las frases permitiría distinguirlas por la relación contextual banco-plaza y banco-dinero. De forma análoga, frases radicalmente diferentes solo encuentran su distinción en la sintáctica, por ejemplo "Argentina le ganó a Francia en el mundial" y "Francia le ganó a Argentina en el mundial". Por su parte, el nivel discursivo excede las frases o términos individuales y contribuyen a la comprensión del texto como un todo (Liddy, 2001); ejemplos del mismo se pueden encontrar en la detección de tópicos o métodos de reducción de dimensionalidad. El NLP combinado con técnicas de aprendizaje automático (ML) (ver Sección 2.6) encuentra su aplicación en diversos campos como traducciones automatizadas (Luong et al., 2015), tareas de pregunta-respuesta (Wiese et al., 2017), resúmenes de texto (Yu et al., 2018), entre otros. Investigaciones basadas en NLP mostraron que el pensamiento espontáneo puede ser estudiado directa y objetivamente con el análisis del discurso (Elvevåg et al., 2007; Mota et al., 2012; Carrillo et al., 2013; Mota et al., 2014; Bedi et al., 2014).

Uno de los mayores desafíos que enfrentamos actualmente radica en la falta de interpretabilidad. Por un lado, la diversificación en las técnicas de NLP permitieron la extracción automatizada de inmensas cantidades de *features* lingüísticas. Por otro lado, el aumento de complejidad de los algoritmos de ML condujo a modelos predictores de alto rendimiento que funcionan como "cajas negras", es decir, modelos que no podemos explicar. En consecuencia, muchas investigaciones extraen múltiples dominios lingüísticos que utilizan como entrada en algoritmos de ML para obtener resultados. Si bien este constituye un enfoque válido, muchas veces los dominios considerados se encuentran descorrelacionados de la temática investigada (de la Fuente Garcia et al., 2020; Fraser et al., 2016; Orimaye et al., 2014).

En esta tesis, exploramos la relación entre el lenguaje y las alteraciones de las funciones cognitivas generadas por dos estados cualitativamente diferentes, pero que pueden mostrar similitudes en cuanto se asocian a una fuerte afectación de capacidades específicas que incluyen la atención, la memoria, las funciones ejecutivas, entre otras. Por un lado, investigamos los efectos de drogas psicodélicas (Sección 1.1), las cuales resultan en un estado transitorio y reversible al cabo de pocas horas. Por otro lado, estudiamos los efectos de enfermedades neurodegenerativas (Sección 1.2) tales como la enfermedad de Alzheimer, las cuales resultan en un deterioro progresivo patológico e irreversible de la cognición. Cuando nos referimos a "efectos de las enfermedades neurodegenerativas" apuntamos a los síntomas de la enfermedad y cuando hablamos de estados de conciencia alterados entendemos que vienen dados por atrofias cerebrales. A pesar de las diferencias fundamentales que existen entre estos dos tipos de estados, proponemos la hipótesis de

que marcadores basados en lenguaje natural podrían ser capaces de caracterizarlos y distinguirlos de un estado usual u ordinario de conciencia.

1.1. Sustancias psicodélicas

Los psicodélicos son compuestos con el potencial de alterar profundamente el estado de conciencia del usuario. Estas sustancias pueden clasificarse en grupos o familias según su acción farmacológica (Tagliazucchi et al., 2017; Iversen et al., 2017); en particular, los psicodélicos serotonérgios "clásicos" actúan en el cerebro como agonistas parciales del receptor de serotonina subtipo 2A ($5-HT_{2A}$). Los grupos funcionales de las moléculas serotonérgicas corresponden a triptaminas (e.g. N,N-demietiltriptamina [DMT], 4-hidroxi-N,N-demietiltriptamina [psilocina]) o a feniletilaminas (e.g. 2-(3,4,5-trimetoxifenil) etanamina [mescalina]), con la excepción de la dietilamida de ácido lisérgico (LSD) y otras moléculas similares tales como el 1-Propionil-LSD (1P-LSD) y el 6-Alil-6-Nor-LSD (AL-LAD), que se tratan de ergolinas y contienen ambos grupos funcionales (Nichols, 2016).

En términos de efectos subjetivos, los agonistas de receptores $5-HT_{2A}$ pueden generar (Preller and Vollenweider, 2016):

- Distorsiones en la información sensorial (sinestesia): Los canales sensoriales se mezclan permitiendo asociar sonidos a colores, gustos a sonidos, etc.
- Distorsiones de la percepción espacio-temporal: Dificultades en la estimación del tiempo entre eventos o distancia entre puntos.
- Cambios en el estado anímico: Inducción de un estado de euforia, éxtasis o pánico. También pueden generar un incremento en los sentimientos de amor, amistad y empatía.
- Unidad con el universo y pérdida de la individualidad: Alteraciones en el sentido del "yo" (disolución del ego). Muchas de estas sustancias generan la sensación de atenuar los límites que separan al individuo del resto del mundo.
- Experiencias místicas y religiosas.
- Otros: Incremento de la creatividad, cambios en el flujo de pensamiento, hiperasociatividad de ideas, etc.

Los compuestos psicodélicos clásicos se encuentran categorizados dentro del *Schedule 1* (i.e. sustancias con potencial de abuso y sin uso médico o científico reconocido en humanos), este hecho ha dificultado la investigación de sus efectos subjetivos y mecanismos

de acción farmacológica. Si bien en muchos casos existen riesgos considerables que no justifican éticamente la experimentación con humanos, se ha argumentado que los psicodélicos clásicos no conllevan grandes riesgos y encuentran uso tanto en la psiquiatría como en la neurociencia (Nutt et al., 2013a,b).

1.2. Demencia

La demencia es un conjunto de enfermedades neurodegenerativas caracterizadas por el deterioro progresivo de las funciones cognitivas. Los síntomas varían ampliamente y dependen tanto de las características específicas del trastorno como de los individuos en sí, es decir, la misma enfermedad puede tener diferentes manifestaciones dependiendo del caso particular. Sin embargo, se pueden establecer ciertos procesos cognitivos que suelen verse afectados en todas las demencias como pérdidas de memoria, déficits de comunicación y lenguaje, incapacidad de reconocer objetos (agnosia) y deterioros de las funciones ejecutivas (razonamiento, juicio y planificación) (Duong et al., 2017). En general, estos trastornos comienzan a manifestarse a partir de los 60 años y pueden clasificarse por métodos clínicos (historial médico, exámenes físicos, evaluaciones cognitivas, etc), por las principales regiones cerebrales afectadas (corteza cerebral, tálamo, lóbulos, entre otras) o por anomalías moleculares basadas en acumulaciones de proteínas específicas (Dugger and Dickson, 2017). Sin embargo, todos los métodos de clasificación presentan superposición entre las patologías y los primeros síntomas de demencia suelen ser extremadamente similares a degeneraciones de procesos cognitivos normales (i.e. causadas por el envejecimientos), estos factores dificultan la tarea de identificación de trastornos.

En el 2019, la Organización Mundial de la Salud (WHO) (`https://www.who.int/`) determinó que alrededor de 55 millones de personas en el mundo padecían demencia y estimó un incremento de 78 millones para el 2030. Además, los costos económicos globales rondaron los 1.3 trillones de dólares en 2019 y se proyecta que alcanzará de 1.7-2.8 trillones para el 2030. Dentro de las variantes de demencia, las más frecuentes corresponden a la enfermedad de Alzheimer (EA) y la enfermedad de Parkinson (EP).

Alrededor de 60 %-70 % de los individuos con demencia son diagnosticados con la EA. El síntoma clínico más distintivo de esta enfermedad corresponde a pérdidas de memoria donde los dominios más afectados son el episódico (recuerdos autobiográficos) (Erkkinen et al., 2018) y el semántico (relacionada con los hechos, vocabularios y conceptos adquiridos por la experiencia) (Rogers and Friedman, 2008; Ozel-Kizil et al., 2020). Las principales atrofias se generan en los lóbulos temporales, frontales y parietales, aunque también suelen encontrarse degeneraciones en la zona del hipocampo (Scheltens et al.,

2021; Dugger and Dickson, 2017). Además, la EA se caracteriza por la acumulación irregular de proteínas β-amiloide ($A\beta$) (Chen et al., 2017).

La EP se constituye como el segundo trastorno neurodegenerativo más frecuente y cuenta con el $\sim 11\,\%$ de los casos (`https://www.who.int/`). Esta enfermedad se manifiesta (principalmente) en los dominios motores causando temblor, rigidez, inestabilidad, lentitud en los movimientos, entre otros (Hoehn and Yahr, 1998). Los efectos no motoros incluyen deterioros de las funciones ejecutivas, trastornos en el estado de ánimo y desórdenes de sueño (Poewe et al., 2017). Además, la EP se caracteriza por acumulaciones de la proteína α-sinucleína y la pérdida de neuronas dopaminérgicas en la zona ventrolateral de la sustancia negra (Dickson et al., 2009).

1.3. Resumen de tesis

Incluyendo la presente introducción (Cap. 1), esta tesis se compone de siete capítulos. En el Capítulo 2 se describe el funcionamiento de las técnicas de NLP, los métodos desarrollados para extraer métricas lingüísticas y los algoritmos de ML aplicados. Los Capítulos 3 y 4 corresponden a contenidos adaptados de dos publicaciones: "The entropic tongue: Disorganization of Natural Language under LSD" (Sanz et al., 2021) y "Natural Language signatures of psilocybin microdosing" (Sanz et al., 2022a), respectivamente. Ambas investigaciones se basan en el análisis de alteraciones lingüísticas generadas por sustancias psicodélicas. En el primero se estudiaron entrevistas conducidas bajo los efectos del LSD y una condición de placebo (PCB). Utilizando herramientas de NLP se extrajeron métricas semánticas, no semánticas (estructurales) y de desorganización directa del lenguaje que permitieron diferenciar las condiciones. En el segundo se analizaron entrevistas conducidas bajo los efectos de microdosificación serotonérgica (con Psilocybe cubensis) frente a una condición de placebo. Esta práctica produce alteraciones más sutiles y menos perceptibles que una dosis completa. Para diferenciar las condiciones, se extrajeron métricas lingüísticas que reflejaban los efectos informalmente reportados de la microdosificación. Además, se compararon diversas medidas que habían resultado distintivas del LSD en nuestra publicación anterior. En el Capítulo 5 corresponde a contenidos adaptados de "Automated text-level semantic markers of Alzheimer's disease" (Sanz et al., 2022b) donde estudiamos las alteraciones lingüísticas generadas por la EA. Utilizando diversas tareas clínicas de lenguaje (i.e. descripción de imágenes, relatos de experiencias, entre otras), se extrajeron dos métricas que reflejaban los efectos neurológicos de la EA. Para determinar si estas medidas eran selectivas de la enfermedad, se realizaron comparaciones con un grupo de control (GC) y pacientes con EP. El Capítulo 6 también corresponde a un análisis de las alteraciones del lenguaje en

la EA. Sin embargo, el enfoque utilizado difiere de las publicaciones anteriores. En lugar de extraer métricas asociadas a efectos reportados en la literatura, se utilizó el algoritmo BERT (ver Sección 2.6.2) para obtener una clasificación automatizada de pacientes con esta enfermedad frente a un GC. Los resultados se interpretaron con el modelo LIME. Además, se utilizó una regresión logística para realizar la misma clasificación y comparar tanto el rendimiento de BERT como la interpretación provista por LIME. En el Capítulo 7 se exponen las conclusiones generales de la tesis así como las futuras líneas de investigación.

Capítulo 2

Métodos

En el transcurso de esta tesis se aplicaron distintos métodos y algoritmos que se definen a continuación.

2.1. Pre-procesamiento de texto

El pre-procesamiento de texto constituye el primer eslabón de NLP ya que permite estandarizar los datos, eliminar el ruido y mejorar el rendimiento de los algoritmos. Dado un documento, el procedimiento puede resumirse en las siguientes instancias (Anandarajan et al., 2019):

- Separar las palabras individuales y puntuación (*Tokenización*) e identificar la categoría gramatical de cada "token" (*Part of Speech* [POS] *Tagging*).
- Transformar las palabras a su término de raíz (*Lemmatización*); e.g. verbos conjugados en distintos tiempos a su infinitivo, términos plurales a singulares, entre otros.
- Descartar caracteres o palabras con escaso contenido semántico. Este paso incluye convertir mayúsculas en minúsculas y remover tanto símbolos de puntuación como un conjunto de palabras de uso frecuente en el lenguaje (conjunciones, preposiciones, pronombres personales, etc) denominadas "stopwords".

Las primeras dos instancias pueden resolverse con diversos modelos computacionales basados en cadenas de Markov (Kupiec, 1992), redes neuronales (Schmid, 1994), entre otros. La mayoría de estos métodos se entrenan o testean con ejemplos etiquetados

manualmente generando una fuerte dependencia respecto al lenguaje del documento (inglés, español, alemán, etc).

La implementación de la última instancia depende del problema que se esté analizando. Tanto los símbolos de puntuación como la distinción en mayúsculas pueden ser fundamentales para algunos algoritmos (e.g. Análisis de Sentimiento [SA]) (Symeonidis et al., 2018). De igual forma, el conjunto de "stopwords"puede incluir otro tipo de términos (Anandarajan et al., 2019). A modo de ejemplo se puede considerar la clasificación de documentos provenientes de sujetos que presentan una enfermedad frente a un grupo de control. Las "stopwords"de este problema incluirán nombres, sinónimos y abreviaciones de la enfermedad para lograr clasificaciones no triviales. Descartando estas salvedades, un documento pre-procesado correspondería a una lista de palabras en minúscula y "lemmatizadas" (manteniendo el orden original), sin símbolos de puntuación ni palabras de uso frecuente.

Para ejemplificar, consideremos la frase "Las manadas viajan en busca de presas.". Aplicamos la librería TreeTagger de Python con el Corpus de Ancora en Español (`http://clic.ub.edu/corpus/es/ancora`) para efectuar los dos primeros pasos del pre-procemiento. De esta forma, obtenemos una lista de vectores con tres componentes que contienen el "token" original, la categoría gramatical y el "lemma" (Tabla 2.1).

Token	**POS Tagg**	**Lemma**
Las	DET.Def.Fem.Plur.Art	el
manadas	NOUN.Fem.Plur	manada
viajan	VERB.Ind.Plur.3.Pres.Fin	viajar
en	ADP.Prep	en
busca	NOUN.Fem.Sing	busca
de	ADP.Prep	de
presas	NOUN.Fem.Plur	presa
.	PUNCT.Final	.

Tabla 2.1: Ejemplo de aplicación de las dos primeras instancias de pre-procesamiento a la frase: "Las manadas viajan en busca de presas.".

Implementando la última instancia se obtendría: ["manada", "viajar", "busca", "presa"]; donde se utilizó el conjunto de palabras de uso frecuente incluido en NLTK (`https://www.nltk.org/`).

2.2. Modelos de Embedding

En NLP coexisten diversos métodos para generar representaciones estructuradas del lenguaje. Entre ellos se destacan los *Word Embeddings* que identifican las palabras con

vectores en un espacio de $\mathbb{R}^n$ donde n es un hiper-parámetro del modelo. Se entrenan sobre extensos corpus de documentos con el objetivo de posicionar "cerca" en el espacio aquellos términos semánticamente similares (Rong, 2014). La cercanía entre vectores de palabras se determina bajo una métrica, la más frecuente corresponde a la similitud semántica o coseno (Wang et al., 2019): $sim(u, v) = \frac{u.v}{\|u\|\|v\|}$ con $u, v \in \mathbb{R}^n$. A modo de ejemplo, consideramos las palabras "perro", "gato" y "logaritmo", un modelo de *embedding* razonable debería posicionar "cerca" (i.e. alta similitud semántica) los dos primeros términos. Mientras que el término matemático debería ubicarse "lejos" (i.e. baja similitud semántica) de las palabras referidas a animales. En efecto, al computar las similitudes con el modelo *FastText* (`https://fasttext.cc/`) se obtiene el resultado esperado: $sim(perro, gato) = 0{,}82 > sim(logaritmo, gato) = 0{,}11 \simeq sim(logaritmo, perro) = 0{,}05$. Además, estos métodos funcionan como reductores de dimensionalidad (Yin and Shen, 2018), ya que identificar relaciones semánticas entre palabras permite representar el vocabulario del corpus en un espacio de dimensionalidad menor al original.

Los algoritmos para generar relaciones entre palabras y vectores varían ampliamente (Ghannay et al., 2016), desde matrices de co-ocurrencia que cuentan frecuencias de palabras en el corpus de entrenamiento (e.g. Latent Semantic Analysis [LSA], TfIdfVectorizer [TF-IDF]) hasta redes neuronales que modelan la probabilidad de aparición de un término según el contexto (e.g. Word2Vec). Además, algunos métodos de *embedding* permiten operaciones aritméticas (Mikolov et al., 2013) con resultados lingüísticamente razonables, por ejemplo: $v(\text{“}rey\text{”}) - v(\text{“}hombre\text{”}) + v(\text{“}mujer\text{”}) = v(\text{“}reina\text{”})$; donde $v(\text{“}w\text{”})$ corresponde a la representación vectorial de la palabra w en el *embedding*.

Basados en estos modelos, desarrollamos dos técnicas para cuantificar características que discriminen alteraciones en estados de conciencia inducidos tanto por sustancias psicoactivas como por enfermedades neurodegenerativas: la variabilidad semántica y el rango del *embedding*.

2.2.1. Variabilidad semántica

La coherencia es un concepto que engloba nociones tanto locales como globales y puede definirse como "La cualidad de ser lógico o consistente"[1]. En particular, la coherencia semántica está vinculada a relaciones conceptuales entre palabras o partes que contribuyen a la comprensión del lenguaje (Silverman, 2021). Esta cualidad suele verse comprometida en enfermedades neuropsiquiátricas tales como la esquizofrenia (Elvevåg et al., 2007) y podría verse afectada en los efectos agudos de psicodélicos.

[1] Oxford Languages: `languages.oup.com`

Motivados por estas nociones, desarrollamos un método objetivo para cuantificar la coherencia de un documento por medio de su variabilidad semántica que esquematizamos en la Figura 2.1. Dado un texto pre-procesado, aplicamos un modelo de *embedding* para obtener la representación vectorial de cada palabra $(v_1, v_2, ...v_n)$. Calculamos las distancias semánticas (i.e. $d = 1 - sim(u, v)$) entre palabras consecutivas $d_i = d(v_i, v_{i-1})$ y obtenemos la variabilidad de las distancias computando la varianza (σ^2):

$$\sigma(d)^2 = \frac{1}{1-n}\sum_{i}^{n-1}(d_i - \mu)^2 \tag{2.1}$$

$$\mu = \frac{1}{n}\sum_{i} d_i$$

De esta forma, un documento presenta una varianza elevada cuando las palabras consecutivas tienden a estar "alejadas" en el espacio vectorial del *embedding*. Esto implica que el documento contiene cambios repentinos en el flujo semántico de los conceptos y, por ende, menor coherencia.

Cabe destacar que este método requiere el pre-procesamiento del documento analizado para eliminar la interferencia que generan los derivados de un término (e.g. verbos conjugados), las palabras con escaso contenido semántico ("stopwords") y los símbolos de puntuación.

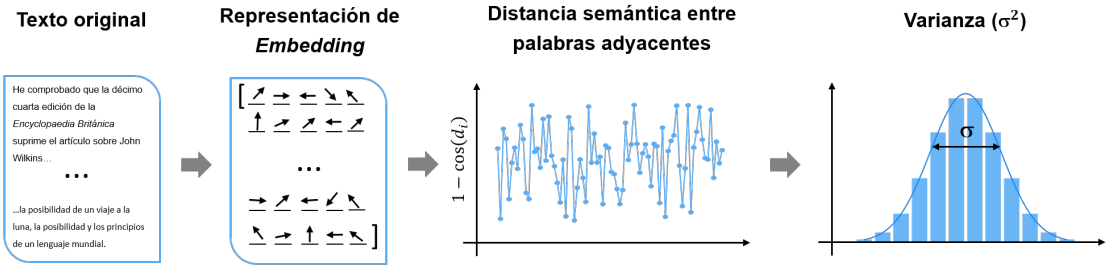

FIGURA 2.1: Esquema metodológico desarrollado para medir la variabilidad semántica de un documento y cuantificar su coherencia.

2.2.2. Rango de un documento

En el ámbito de Álgebra Lineal, el rango de una matriz corresponde a la dimensión del espacio vectorial generado por sus filas (o columnas). Es decir, el número máximo de vectores linealmente independientes que conforman la matriz (Hefferon, 2020). Aplicado al lenguaje, el rango se asemeja al concepto de sinopsis o resumen, ya que los mismos

constituyen explicaciones breves y no redundantes de las ideas principales contenidas en una materia.

Inspirados en estas nociones, desarrollamos un método para cuantificar la compresibilidad de un documento, es decir, la mínima dimensión a la que puede reducirse sin perder información. Dado un texto y un modelo de *embedding*, conformamos una matriz cuyas filas correspondían a las representaciones vectoriales de las palabras. Aplicando la Descomposición en Valores Singulares (SVD) determinamos el rango de la matriz contando el número de autovalores que excedían un valor umbral (Press et al., 2007).

En esta métrica, la redundancia se manifiesta como la probabilidad de formular los vectores de palabras de un documento con combinaciones lineales de otros vectores presentes. Por lo tanto, un conjunto arbitrario de términos abarcará palabras con diferente contenido semántico resultando en la independencia lineal de la mayoría de los vectores ("alto" rango). Por el contrario, en el espacio vectorial generado por un documento altamente redundante, los vectores de palabras tenderán a reflejar la información contenida en otros ("bajo" rango).

2.3. Análisis de sentimiento

El análisis de sentimiento (SA) es una técnica de NLP que permite identificar las emociones expresadas en un texto y clasificar su polaridad (positivo, neutral o negativo) (Medhat et al., 2014).

Los algoritmos de SA pueden dividirse en tres categorías con diferentes enfoques (Madhoushi et al., 2015; Sankar and Subramaniyaswamy, 2017). Los primeros son aquellos basados en diccionarios de palabras, donde la polaridad de los términos se etiquetan manualmente. Los segundos basan su funcionamiento en técnicas de aprendizaje automático (ML) tanto supervisado como no supervisado que extraen patrones de los datos. Los últimos consisten en modelos híbridos que combinan los enfoques anteriores para mejorar el rendimiento. En el caso supervisado, el entrenamiento se realiza con documentos de polaridad rotulada, por ejemplo, reseñas de productos donde el usuario etiqueta (dentro de un rango) la calidad del mismo.

Para ejemplificar, consideramos el modelo híbrido *senti-py* (`github.com/aylliote/senti-py`) aplicado en la publicación (Sanz et al., 2022a) que asigna la polaridad de una frase puntuando de 0 (negativa) a 1 (positiva). Extrajimos tres oraciones de los datos utilizados en el mismo trabajo para ilustrar las puntuaciones:

- Sentimiento negativo (puntuación = 0.06): “Me siento cansado, estuve cansado todo el día, no creo que... sigo pensando que tomé dos placebos”.
- Sentimiento neutro (puntuación = 0.49): “No, creo que fue bastante normal eh... como a grandes rasgos siento que estuve más tranquilo como que al presentarse las cosas que pasan normalmente pude verlo como un paso más atrás viste, como más relajado pero aparte de eso bastante normal, no vi grandes cambios”.
- Sentimiento positivo (puntuación = 0.83): “Ahora, completa tranquilidad”.

2.4. Grafos

Los grafos (o redes) son estructuras matemáticas que modelan sistemas complejos. Están compuestos por grupos de nodos (N) conectados por enlaces (E) y presentan propiedades topológicas que reflejan las relaciones entre objetos del sistema (Newman, 2010). Nos propusimos investigar la estructura del lenguaje mediante redes, donde los nodos son palabras y las conexiones representan secuencialidad.

Las propiedades topológicas de los grafos se cuantifican por medidas globales y locales (ver Figura 2.2), también llamadas métricas o características. A continuación, se definen una serie de métricas frecuentemente utilizadas para caracterizar un grafo de N nodos y E enlaces (Mota et al., 2014).

Métricas globales:

- Densidad (ρ): Proporción entre E y el máximo número posible de conexiones en el grafo, es decir: $\rho = \frac{E}{N(N-1)/2}$.
- Camino más corto promedio (ASP): El camino más corto (SP) entre un par de nodos corresponde, para un grafo no pesado, al número mínimo de enlaces que conectan el par. Sea d_{ij} el SP entre los nodos i y j, entonces: $ASP = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}$.
- Diámetro (D): Longitud del mayor SP entre pares de nodos.
- Grado medio total (ATD): El grado del nodo i (k_i) corresponde al número de enlaces que lo conectan con otros nodos de la red. El ATD refiere a la suma normalizada de todos los grados del grafo: $ATD = \frac{1}{N} \sum_i k_i$
- Mayor componente fuertemente conectada (LSC): Número de nodos incluidos en el máximo sub-grafo completo (existe un camino del nodo i al j y uno del j al i).

- Coeficiente medio de clustering (CC): El coeficiente de clustering del nodo i corresponde a la proporción entre el número de enlaces con sus primeros vecinos y el número de enlaces que tendría si los primeros vecinos estuvieran completamente conectados. El ACC se computa como el promedio del coeficiente de clustering de todos los nodos.

Métricas locales:

- "Loops"de primer orden (L1): Número de ciclos que contienen un nodo.
- "Loops"de segundo orden (L2): Número de ciclos que contienen dos nodos.
- "Loops"de tercer orden (L3): Número de ciclos que contienen tres nodos.
- Enlaces repetidos (RE): Número de enlaces que vinculan el mismo par de nodos.
- Enlaces paralelos (PE): Número de enlaces que vinculan el mismo par de nodos considerando que el nodo de origen de un enlace puede ser el nodo objetivo del enlace paralelo.

2.5. Grafos de discurso

Co-existen diversas maneras de implementar elementos de la Teoría de Grafos al NLP (e.g. modelos de *Word Embedding* y SA). Sin embargo, una de las aplicaciones más directas corresponde a los "grafos de discurso" (Mota et al., 2012) donde la red representa un documento (o conjuntos de ellos). Una forma de configurar estos grafos consiste en asignar un nodo a cada palabra diferente de un texto y los enlaces a términos consecutivos del documento. Este tipo de red presenta las propiedades de ser dirigido (las conexiones son unidireccionales), no pesado (las conexiones tienen igual peso entre sí) y multi-enlace (más de un enlace dirigido puede conectar al mismo par de nodos). Para ejemplificar, la Figura 2.3A ilustra el grafo de la siguiente frase, extraída de los datos utilizados en (Sanz et al., 2022a): "Estafado, me siento estafado, por ahí me tocó el placebo no... no me siento drogado.".

Las estructuras presentes en los "grafos de discurso" permiten visualizar y cuantificar el contenido no semántico del lenguaje por medio de las características topológicas definidas anteriormente. Las métricas locales reflejan la recurrencia presente en un documento y las globales se asocian a la conectividad (Mota et al., 2014). A pesar de que la mayoría de estas medidas presentan una fuerte dependencia con la longitud del documento, los efectos pueden descartarse aplicando normalizaciones. En nuestra publicación (Sanz

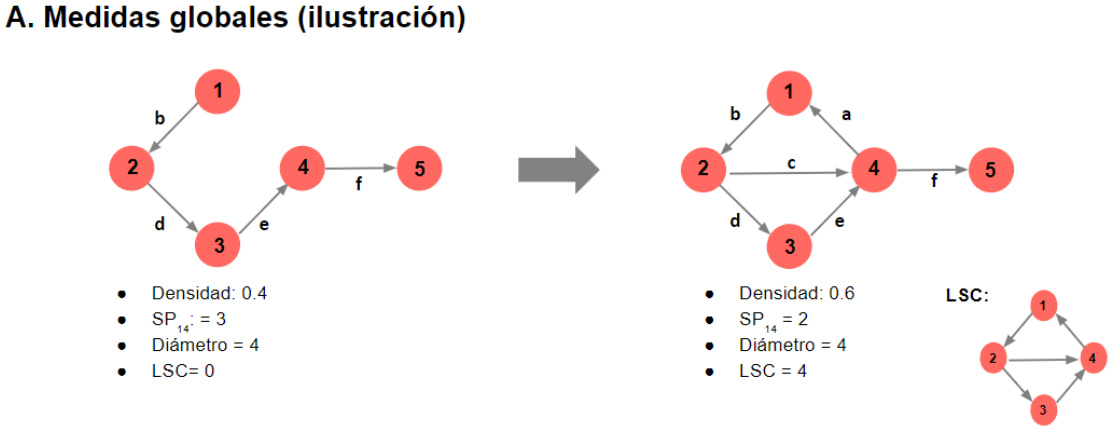

FIGURA 2.2: Ilustración gráfica de las medidas globales (**A**) y locales (**B**) de un grafo. **A.** Los grafos presentan la misma cantidad de nodos y diferente número de enlaces, en consecuencia el grafo de la derecha (con mayor cantidad de enlaces) es más denso. En el grafo de la izquierda, se puede establecer un único camino entre los nodos 1 y 4 (enlaces $\{(b, d, e)\}$) por ende el $SP_{1,4} = 3$. Mientras que en el otro grafo se pueden trazar dos caminos $\{(b, c), (b, d, e\}$, tomando el más corto se obtiene $SP_{1,4} = 2$. El diámetro de ambos coincide y corresponde a los enlaces $\{(b, d, e, f)\}$ (distancia entre los nodos 1 y 5), por lo tanto el diámetro es 4. El LSC corresponde a 0 para el grafo de la izquierda y 4 en el de la derecha, se puede ver que en el sub-grafo se puede trazar un camino entre cualquier par de nodos. **B.** Ilustración gráfica de cinco medidas locales de un grafo. De izquierda a derecha: Ciclos de orden 1 (L1), 2 (L2) y 3 (L3); enlaces repetidos (RE) y enlaces paralelos (PE).

et al., 2021), utilizamos un método basado en modelos nulos. Para cada documento, se computaron 1000 grafos aleatorios generados por la distribución azarosa de sus palabras. El promedio de las métricas obtenidos con los grafos aleatorios se utilizó para normalizar las características reales.

2.5.1. Granularidad semántica

La taxonomía puede definirse como la clasificación ordenada de un conjunto de objetos que comparten características comunes [2]. Dentro del lenguaje, diversos grupos presentan una estructura jerarquizada (Resnik, 1999). En particular, la categoría gramatical de los sustantivos conforma una taxonomía donde los hiperónimos (e.g. "familia") engloban hipónimos (e.g. "nieto") (Yang and Powers, 2005).

[2] Oxford Languages: `languages.oup.com`

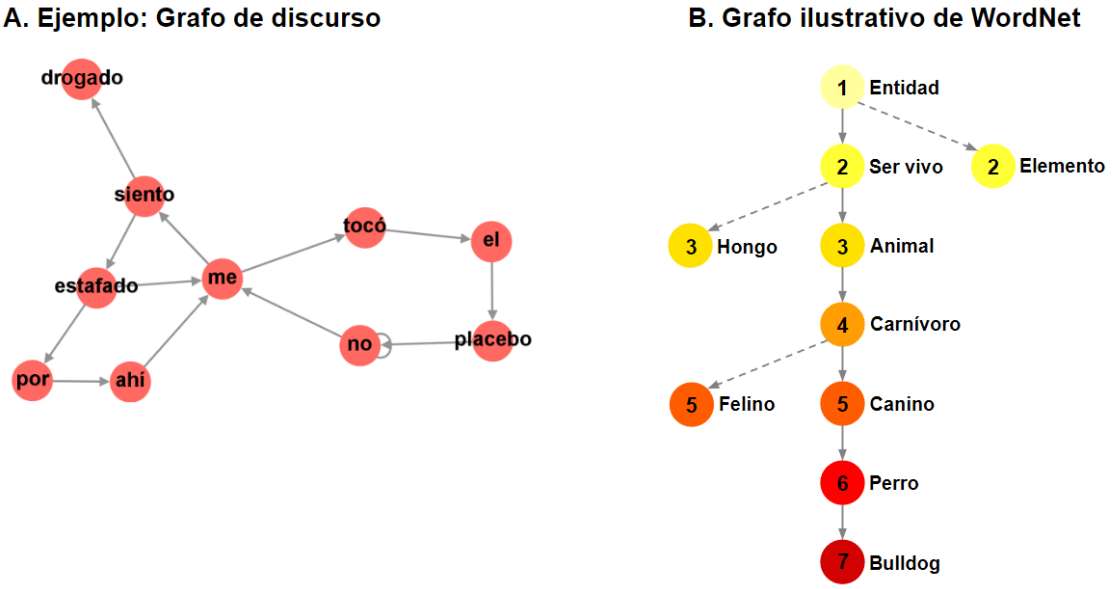

FIGURA 2.3: **A.** Ejemplo del grafo de discurso de una frase. **B.** Segmento simplificado del grafo *WordNet* que muestra relaciones jerárquicas (algunos nodos intermedios se omitieron por claridad del ejemplo). Los valores de granularidad se identifican con el color y número del nodo. Las líneas puntadas representan bifurcaciones de la red que no conducen al nodo "Bulldog".

Una posible jerarquización de los sustantivos puede encontrarse en la base de datos léxica WordNet (https://wordnet.princeton.edu/). Esta contiene alrededor de 80.000 sustantivos ordenados que pueden visualizarse como un grafo (dirigido, acíclico y no-pesado) que inicia con el hiperónimo "Entidad" para continuar con términos progresivamente menos generales (ver Figura 2.3B). Se puede entonces cuantificar la granularidad de un sustantivo como el número de nodos que lo separan del término "Entidad". Bajo esta métrica conceptos generales como "animal" o "comida" tendrán menor granularidad que términos específicos como "bulldog" o "zanahoria".

2.6. Clasificadores de aprendizaje automático

El aprendizaje automático (ML) involucra un conjunto de técnicas que producen modelos predictores a partir de la identificación de patrones en los datos (Nasteski, 2017). La principal propiedad de estos modelos radica en la capacidad de inferencia que presentan frente a nuevos conjuntos de datos. Actualmente, las técnicas de ML se aplican en una inmensa variedad de problemas como reconocimiento de imágenes, predicciones de clima, detección de fraude, entre otros.

El modelo predictor generado depende del proceso de aprendizaje que utilice el algoritmo. En esta sección nos enfocaremos en las técnicas de clasificación supervisadas donde los datos se encuentran asignados a categorías. La tarea de aprendizaje radica en generar una

función que mapee las *features* (variables) de entrada a sus respectivas clases (Mahesh, 2020). Estos algoritmos se dividen en dos etapas: entrenamiento y evaluación (testeo). En la primera instancia, se utiliza un subconjunto de los datos etiquetados (*training set*) para extraer patrones e inferir la relación funcional de mapeo. La segunda etapa consiste en predecir las categorías del subconjunto de datos restantes (*testing set*). El rendimiento del clasificador se determina con la comparación de las clases predichas con las reales (Mahesh, 2020), las métricas más frecuentemente utilizadas corresponden a la exactitud (*accuracy*) (i.e. fracción de predicciones correctas) y el área bajo la curva (AUC) ROC (*Receiver Operating Characteristics*) (Huang and Ling, 2005).

En este trabajo, utilizamos algoritmos de ML que presentan robustez y alto rendimiento sin necesidad de ajustar los parámetros del modelo (Friedman, 2002; Breiman, 2001). En la Sección 2.6.1 se describen los clasificadores *Random Forest* y *Gradient Boosting* aplicados en (Sanz et al., 2021, 2022a) y (Sanz et al., 2022b), respectivamente. El proceso de entrenamiento fue análogo en todas las publicaciones. En principio, los datos se dividieron aleatoriamente en k subgrupos mediante la validación cruzada estratificada para preservar la proporción de clases. Luego, se realizaron k iteraciones donde $k-1$ subconjuntos se utilizaron para entrenar y el restante para evaluar. Dado que cada subconjunto atraviesa la instancia de testeo, obtenemos un *score* para cada una de las muestras y calculamos el AUC utilizando estos valores. Una de las ventajas de utilizar este método es que se evitan sesgos en los resultados provenientes de *testing sets* particularmente informativos. Este proceso se repitió N veces, tanto para las categorías reales como para las categorías aleatoriamente distribuidas que rompen la relación funcional entre *features* de entrada y etiquetas, dando como resultado un modelo nulo. Finalmente, se construyó un p-valor contando el número de veces que el AUC del modelo nulo excedía al obtenido por el real, normalizado por el número total de repeticiones.

2.6.1. Ensembles de árboles

Entre las técnicas de ML, los árboles de decisión (DT) se destacan por su interpretabilidad (James et al., 2013). La idea básica de estos algoritmos radica en dividir una decisión compleja (i.e. una clasificación que depende de múltiples *features*) en la unión de decisiones más simples (i.e. respuestas binarias que dependen de una única *feature*) (Safavian and Landgrebe, 1991).

La estructura de los DT corresponde a un grafo (no pesado, acíclico y dirigido) donde los nodos representan preguntas binarias relacionadas a las *features* de los datos y las respuestas se simbolizan con enlaces (ver Figura 2.4.A). El algoritmo comienza del "nodo raíz" (asociado a la *features* más informativa) de donde se desprenden "ramas"

(posibles respuestas) que conectan con "nodo internos" (otras *features*). El árbol se conforma repitiendo este proceso, los nodos finales de cada rama se denominan "hojas" y representan el resultado del algoritmo.

En particular, los clasificadores buscan minimizar medidas de "impureza" que reflejan la proporción de clases en cada división del árbol. De esta forma se considera que una variable es más informativa si la subdivisión de datos que produce contiene la menor mezcla de clases posible (Gupta et al., 2017). En general, la impureza de la hoja m, se determina con el Criterio de Gini (ecuación 2.2) o con la Entropía (ecuación 2.3).

$$G_m = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) \tag{2.2}$$

$$E_m = -\sum_{k=1}^{K} p_{mk} log(p_{mk}) \tag{2.3}$$

Donde K es el número total de clases y p_{mk} corresponde a la proporción de los datos que están en la hoja m y pertenecen a la categoría k. La impureza total del árbol se computa como la suma de las impurezas de cada hoja.

El problema principal que presentan los árboles de decisión se encuentra en la inestabilidad del algoritmo, pequeñas perturbaciones en los datos de entrenamiento pueden alterar significativamente las predicciones (Breiman, 1996). Para solventar esta desventaja, surgieron modelos compuestos de conjuntos de predictores simples denominados *ensembles*. El algoritmo "Random Forest" (RF) es un *ensemble* conformado por árboles independientes de idéntica estructura (ver Figura 2.4.B). La independencia se alcanza al considerar diferentes subconjuntos de las *features* de entrada para cada predictor, el tamaño del subconjunto suele ser similar a la raíz cuadrada del número de *features* totales y las variables que contienen se seleccionan aleatoriamente. De esta forma, se evita la presencia de *features* altamente informativas en todos los árboles del *ensemble* (James et al., 2013). En particular, los RF de clasificación, asignan categorías a los datos mediante la votación de las predicciones particulares de cada árbol. La robustez viene de promediar modelos independientes con bajo sesgo y alta varianza (DT individuales) que resultan en un nuevo modelo de bajo sesgo y varianza reducida, como consecuencia del Teorema Central del Límite (Breiman, 2016).

Otro *ensemble* de árboles frecuentemente utilizado es el "Gradient Boosting" (GB). A diferencia del anterior, los predictores individuales se agregan de forma secuencial (ver Figura 2.4.C). Cada nuevo árbol se entrena para reducir el error total que presenta el

ensemble hasta esa instancia (Natekin and Knoll, 2013). Para obtener robustez (Breiman, 1996), estos modelos utilizan subconjuntos de los datos para entrenar cada árbol (*boostraping*) junto con técnicas de regularización (Friedman, 2002).

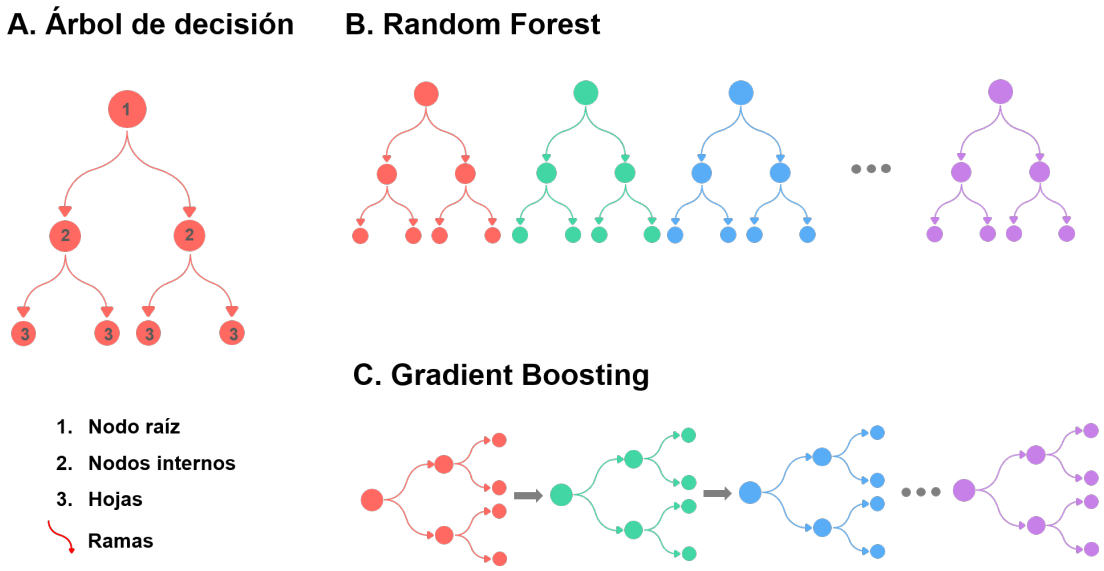

FIGURA 2.4: Representaciones estructurales de **A.** Un árbol de decisión y sus componentes principales (nodos, hojas y ramas); **B.** Un algoritmo *Random Forest* conformado por N árboles de decisión independientes; **C.** Un algoritmo *Gradient Boosting* compuesto por N árboles de decisión concatenados de forma secuencial.

2.6.2. BERT

Dentro del ML, se encuentra la rama de algoritmos secuenciales (*seq2seq*) en donde tanto los datos de entrada como de salida corresponden a una serie de características. Se utilizan para resolver tareas en donde los elementos individuales de la secuencia se modifican o interactúan entre sí, es decir, tareas contextuales. De las aplicaciones más comunes al NLP se pueden mencionar la traducción de un documento, problemas de pregunta-respuesta, resumen de texto, entre otras.

Los *Transformers* se destacan frente a otros modelos secuenciales tanto por su eficiencia computacional frente a series de entrada extensas como por la calidad que presentan sus resultados (Vaswani et al., 2017). Se componen de dos redes neuronales profundas denominadas *Encoder* y *Decoder*. El primero recibe los datos de entrada y los codifica en vectores localizados en un espacio de dimensión reducida (respecto a la original) mientras que el segundo decodifica la información. Los *Transformers* utilizan mecanismos de atención permitiendo que el modelo detecte información contextual entre elementos relevantes de la secuencia (Lin et al., 2022).

En el 2018, investigadores de Google AI introdujeron el modelo "Bidirectional Encoder Representations from Transformers" (BERT) (Kenton and Toutanova, 0189) para resolver diversos problemas de NLP aplicando un ajuste de parámetros pequeños (*Fine-tuning*) a *Encoders* pre-entrenados en inmensas bases de datos. La arquitectura de BERT "base" (ver Figura 2.5) se constituye por una capa de *embedding* y doce bloques de *Encoder* que contienen dos sub-capas. La primera es un mecanismo de atención múltiple y la segunda corresponde a una capa *feed-foward* completamente conectada. Además, cada bloque contiene conexiones residuales, agregados y normalización (Vaswani et al., 2017).

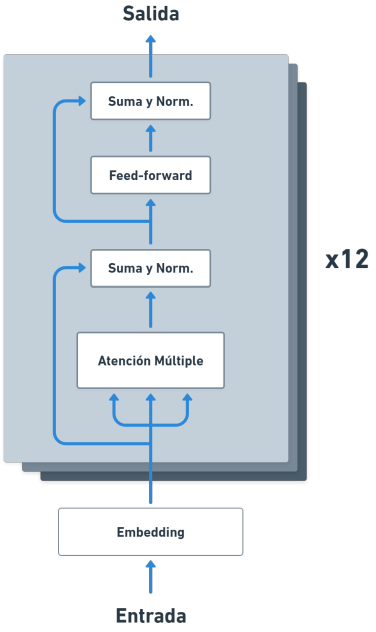

FIGURA 2.5: Arquitectura de BERT: una capa de embedding y doce bloques de *Encoder* conformados por mecanismos de atención múltiple, una capa *feed-foward*, conexiones residuales, agregado (suma) y normalización.

A diferencia de los modelos de lenguaje previamente mencionados, BERT utiliza el vocabulario de WordPiece que contiene palabras y sub-palabras. Estas últimas corresponden a fragmentos de términos que permiten reducir el tamaño del vocabulario y representar palabras de uso poco frecuente como conjunción de estos fragmentos (Wu et al., 2016). En principio, el algoritmo diversifica la secuencia (i.e. frase o conjuntos de frases) de entrada en tres vectores: palabras o sub-palabras (*tokens*) de la serie, posición de los *tokens* de forma estrictamente monótona creciente, identificación de la frase (segmento) a la que corresponde cada *token*. Estos vectores se utilizan como entrada en la capa de *embedding*.

BERT se pre-entrena con dos métodos no supervisados. Por un lado, se utilizan modelos de lenguaje enmascarados (MLM) que predicen *tokens* ocultos (*masked*) de la secuencia de entrada en base al contexto dado por los elementos a izquierda y derecha del mismo.

Por otro lado, se aplican modelos binarizados de predicción de la frase siguiente (NSP). En estos se utilizan dos oraciones de entrada (A y B), la tarea consiste en determinar si B sigue o no a A. De esta forma, un modelo pre-enternado de BERT contiene tanto nociones de contexto como de relación entre frases.

Como se mencionó anteriormente, el pre-entrenamiento de BERT se realiza sobre extensas bases de datos que requieren considerables costos computacionales. En la publicación original se emplearon los datos provenientes de BooksCorpus (800M de palabras) (Zhu et al., 2015) en conjunto con los artículos de Wikipedia en Inglés (2,500M de palabras). Sin embargo, una vez pre-entrenado el modelo, la flexibilidad de BERT permite aplicar *Fine-tuning* para resolver problemas de NLP de diferente índole. Este método consiste en modificar el modelo para resolver una tarea específica al agregar una capa adicional al final de la red y re-entrenarlo con datos relativos a la tarea. A modo de ejemplo, consideramos la clasificación binaria de documentos, la capa agregada correspondería a una lineal con una única neurona y función de activación sigmoidea. Debido a que los pesos iniciales se fijan en el pre-entrenamiento, el *Fine-tuning* alcanza un alto rendimiento en menor tiempo computacional y con menor cantidad de datos (Kenton and Toutanova, 0189).

Capítulo 3

El lenguaje natural bajo los efectos de LSD

Este capítulo abarca contenidos adaptados de "The entropic tongue: disorganization of natural language under LSD" (Sanz et al., 2021) donde estudiamos los efectos de la dietilamida de ácido lisérgico (LSD) en el lenguaje natural.

El LSD se constituye como uno de los psicodélicos serotonérgicos más pontentes en cuanto a su dosis activa mínima (Nichols, 2016). Como se explicó previamente (ver Capítulo 1) esta sustancia actúa mediante agonismo en el receptor de serotonina 2A ($5 - HT_{2A}$); en consecuencia los principales efectos subjetivos incluyen producción de imágenes visuales simples y complejas, distorsiones en la percepción del tiempo y espacio, sinestesia, cambios en el estado de ánimo, alteraciones en la cognición, distorsiones de la conciencia propia (*self-awareness*), entre otros (Kaelen et al., 2015; Dolder et al., 2016; Nichols, 2016; Kometer and Vollenweider, 2016; Preller and Vollenweider, 2016).

En este trabajo hipotetizamos que el LSD genera un aumento en marcadores de desorganización del discurso que permiten distinguirlo del estado basal. Analizamos aspectos semánticos (i.e. contenido del lenguaje) y no semánticos (i.e. estructura del lenguaje) extraídos de entrevistas conducidas bajo los efectos del LSD (75μg) y una condición de placebo (PCB). Además, identificamos métricas asociadas a la desorganización directa del discurso. Para verificar si estas características eran distintivas del estado psicodélico realizamos comparaciones estadísticas y aplicamos dos algoritmos de ML, uno con las métricas semánticas y otro con la no-semánticas. Ambos modelos fueron significativamente distinguibles del azar y resultaron en valores de AUC ~ 0.75. En resumen, nuestros resultados pueden verse como una caracterización cuantitativa y objetiva de la desorganización del lenguaje natural como rasgo del estado psicodélico.

3.1. Participantes y protocolo

Todos los datos utilizados en los siguientes análisis fueron recolectados por la División de Neurociencia de la Facultad de Medicina del *Imperial Collage of London*.

Los participantes de este experimento fueron convocados informalmente, por medio de redes sociales y comunicación oral (de boca en boca). Los principales criterios de exclusión (Carhart-Harris et al., 2015) involucraban: ser menor de 21 años, presentar historial de diagnóstico de enfermedades psiquiátricas, poseer familiares cercanos con desórdenes psicóticos, embarazo, consumo elevado de alcohol (más de 40 unidades por semana), ausencia de experiencia con psicodélicos clásicos, entre otros. En total, se reclutaron veinte sujetos de lengua nativa inglesa que asistieron a las instalaciones experimentales dos veces y con un intervalo de (al menos) dos semanas. Los participantes recibieron LSD ($75\mu g$ en $10mL$ de solución salina) y placebo ($10mL$ de solución salina) por vía intravenosa en orden aleatorio (i.e. desconocido por los sujetos e investigadores). Los días de pruebas experimentales incluyeron estudios de resonancia magnética funcional (fMRI) y magnetoencefalografía (MEG) (con una duración de ∼60 minutos cada uno). Después de cada estudio, se entrevistó a los participantes sobre su experiencia general y se les pidió que relataran (sin límite de tiempo) los sentimientos y pensamientos espontáneos que habían experimentado. El entrevistador mantenía intervenciones mínimas con el objetivo de que los participantes profundizaran los relatos. El escáner de fMRI se llevó a cabo 120-150 min después de la administración de la dosis (momento donde los efectos del LSD alcanzan su auge), por tanto la primera entrevista se realizó a 180-210 min de la infusión y lo identificamos como "Tiempo 1". Mientras que el escáner de MEG se condujo ∼225 min después de la administración, por tanto la entrevista se realizó a ∼285 min de la infusión y lo identificamos como "Tiempo 2". Uno de los participantes experimentó ansiedad en la instancia de escaneo y optó por abandonar el experimento, en consecuencia, sus datos no se utilizaron en los análisis siguientes. La descripción detallada de los participantes y del protocolo experimental puede encontrarse en (Carhart-Harris et al., 2016).

Este estudio fue aprobado por el comité "National Research Ethics Service" de Londres y fue conducido en acuerdo con la declaración de Helsinki (2000), las indicaciones de "The International Committee on Harmonization Good Clinical Practice" y de "National Health Service Research Governance Framework".

3.2. Métodos

En los análisis siguientes solo se consideró el discurso de los participantes, es decir, se descartaron las preguntas e intervenciones de los entrevistadores. El contenido semántico se representó con diez *features* calculadas a partir de la similitud semántica (coseno) entre cada entrevista y una serie de términos predefinidos (Sección 3.2.2). Las *features* no-semánticas se codificaron con nueve métricas basadas en elementos de la Teoría de grafos (Sección 3.2.3). Por último, se calcularon 3 medidas referidas a la desorganización directa del discurso (Sección 3.2.4): la entropía de Shannon, la variabilidad semántica y el rango de las entrevistas.

3.2.1. Pre-procesamiento

Las trascripciones de las entrevistas se pre-procesaron (Sección 2.1) con la librería Natural Language Toolkit (NLTK) (`https://www.nltk.org/`) (Bird et al., 2009) y la lista de palabras de uso frecuente ("stopwords") se extrajo de la librería Gensim (`https://radimrehurek.com/gensim/`). Al considerar el conjunto completo de entrevistas se encontró que 3.6 % de las palabras no estaban incluidas en el vocabulario de NLTK. Dado que la mayoría correspondía a sonidos emitidos por los participantes (i.e. "mmm", "oooooh", "uhm"), se optó por descartarlas. Un breve análisis sintáctico de las partes más relevantes del discurso (i.e. sustantivos, verbos, adjetivos y adverbios) se realizó contando la ocurrencia de estos para las dos condiciones (LSD y PCB) y tiempos. Las comparaciones estadísticas no mostraron diferencias significativas entre condiciones, dando p-valores mayores a 0.04 (Tiempo 1) y 0.036 (Tiempo 2) que no superaron las correcciones de Bonferroni ($\alpha = 0.013$) (ver Apéndice A.1).

3.2.2. Métodos: contenido semántico

El análisis semántico de las transcripciones se realizó con el *embedding* de Word2Vec (Rong, 2014) pre-entrenado con la técnica "Skip-Gram" (i.e. predicción de las palabras de contexto según la palabra objetivo) en un espacio vectorial de 300 dimensiones y con la base de datos de Google News (`https://news.google.com/`). Dado que las palabras polisémicas pueden introducir ruido en el análisis semántico, se estimó si la cantidad de este tipo de palabras era similar en ambas condiciones. Bajo la suposición de que los primeros vecinos de una palabra polisémica presentarán una similitud semántica menor entre ellos, se calcularon las similitudes entre el primer y segundo término más cercanos

de cada palabra contenida en las transcripciones. La comparación estadística entre condiciones no resultó significativa ($p = 0.35$) (Figura 3.1) asegurando la cancelación del ruido introducido por la polisemia en los análisis siguientes.

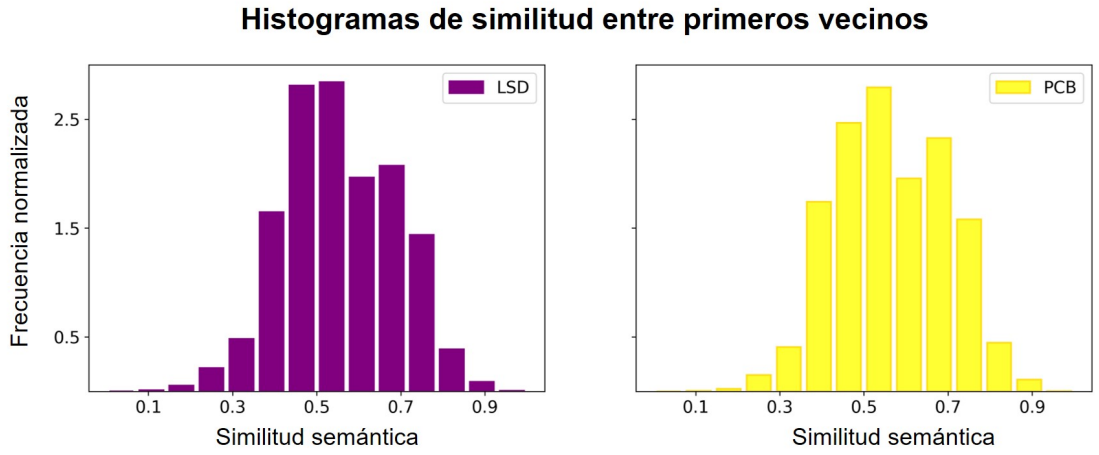

FIGURA 3.1: Distribuciones de similitud semántica entre primeros vecinos de cada palabra en las entrevistas conducidas bajo las condiciones de LSD (izquierda) y PCB (derecha). No se encontraron diferencias estadísticas entre las condiciones ($p = 0.35$).

El *embedding* de Word2Vec se aplicó para mapear los términos de cada transcripción con sus correspondientes vectores. Basados en investigaciones previas de Bedi y sus colegas ((Bedi et al., 2014)), se representó el contenido semántico del discurso como la similitud (métrica coseno) entre las transcripciones y una serie de diez términos predefinidos que correspondían a: "visual" (visuales), "pattern" (patrón), "relax" (relajación), "listen" (escuchar), "mood" (estado de ánimo), "stimulate" (estimular), "normal" (normal), "ego" (ego), "fear" (miedo) y "reality" (realidad). Estos términos se obtuvieron utilizando modelado de tópicos con Análisis Semántico Latente (LSA) (Landauer et al., 1998) en una publicación previa (Zamberlan et al., 2018) y de las nubes de palabras presentadas en la Figura 3.2. Las *features* semánticas se codificaron con el promedio de las similitudes entre las transcripciones y la lista de términos predefinidos.

3.2.3. Métodos: contenido no semántico

El análisis del contenido no semántico se realizó con la librería *SpeechGraphs v1.0* (`https://neuro.ufrn.br/softwares/speechgraphs`) que permitió visualizar los grafos de discurso de cada transcripción y computar el conjunto de métricas descriptas en la Sección 2.4. Para este análisis, el pre-procesamiento consistió en remover los símbolos de puntuación y transformar mayúsculas en minúsculas. En consecuencia, cada transcripción se consideró como una frase única, es decir, todos los nodos del grafo compartían (como mínimo) una conexión dirigida con otro nodo. Dado que las transcripciones conducidas

bajo el estado psicodélico presentaban una mayor verbosidad, se computaron modelos nulos para normalizar las métricas (ver Sección 2.5) y descartar influencias en esta variable. Finalmente, las *features* no semánticas se codificaron con nueve métricas: ASP, CC, Diámetro, LSC, L1, L2, L3, RE and PE.

3.2.4. Medidas de desorganización directa

La representación vectorial de cada transcripción se utilizó para calcular su variabilidad semántica (ver Sección 2.2.1) y estimar su rango (ver Sección 2.2.2). En esta última, se fijó un valor umbral de 0.5 para determinar la independencia entre vectores.

Además, se consideró cada participante como una fuente de secuencias de palabras que trasladan cierta cantidad de información. Para cuantificarla, se calculó la tasa media de producción de información con la entropía de Shannon: $\sum_i p_i log(p_i)$; donde p_i representa el número veces que el término $i-$ésimo figura en la transcripción, divido por el número total de palabras.

3.2.5. Clasificadores y comparaciones estadísticas

Con el objetivo de verificar si las *features* semánticas y no semánticas podían utilizarse como marcadores distintivos entre condiciones, se utilizó la librería sckit-learn (`https://scikit-learn.org/`) para entrenar clasificadores *Random Forest* (Sección 2.6.1) (Breiman, 2001) con validación cruzada estratificada de 5 subconjuntos. El modelo constaba de 1000 árboles de decisión entrenados en subgrupos de 3 *features* aleatoriamente seleccionadas. Aplicando el procedimiento descripto en la Sección 2.6, se determinó el rendimiento del modelo computando el AUC y el p-valor para 1000 iteraciones. Además, se implementaron dos clasificadores diferentes, uno para las *features* semánticas y otro para las no semánticas. En ambos casos, solo se consideraron las transcripciones pertenecientes al Tiempo 1 debido a que el mismo coincide con el momento donde los efectos del LSD son más pronunciados. El Apéndice A.2, contiene una descripción detallada de todas las *features* utilizadas como entrada para los clasificadores, así como las matrices de correlaciones por pares entre estas métricas.

La significancia estadística de las comparaciones por pares se evaluó mediante pruebas no paramétricas de rango con signo de Wilcoxon de dos colas con correcciones de Bonferroni para las comparaciones múltiples. Estadísticas resumidas de estudios anteriores que investigaron métricas similares, principalmente la coherencia (Elvevåg et al., 2007) y las medidas de grafos de discurso (Mota et al., 2012, 2014), fueron utilizadas para estimar el tamaño de la muestra necesario según la fórmula: $n = \frac{(\sigma_1^2+\sigma_2^2)(Z_{1-\alpha/2}+Z_{1-\beta})^2}{\Delta^2}$;

con $\alpha = 0.05$ (probabilidad de error tipo 1) y $\beta = 0.2$ (probabilidad de error tipo 2). Δ corresponde a la diferencia absoluta entre la media de las condiciones, σ a la varianza de cada condición y Z al valor crítico para los valores α, β (Rosner, 2015). Asumiendo grupos de igual tamaño se encontró que $n = 20$ resultaba en un poder estadístico adecuado.

3.3. Resultados

El corpus se dividió en cuatro grupos: transcripciones bajo las condiciones LSD/PCB en el Tiempo 1/Tiempo 2. Las comparaciones entre pares de condiciones se realizaron para cada tiempo por separado, exceptuando el análisis de las palabras polisémicas (previamente presentado) y las representaciones en nubes de palabras de los términos más frecuentes, que se calcularon combinando las transcripciones bajo LSD y PCB en ambos momentos.

3.3.1. Resultados semánticos

La Figura 3.2 muestra nubes de palabras que representan los términos más frecuentes utilizados en las transcripciones bajo la condición de LSD (izquierda) y PCB (derecha). Se descartaron las palabras que aparecían en ambas nubes para destacar las diferencias entre ellas. Encontramos que los participantes bajo la condición de LSD tendían a utilizar términos relacionados con la percepción: “music” (“música”), “hear” (“escuchar”), “space” (“espacio”), “talk” (“hablar”), “visual” (“visuales”), “see” (“ver”), “pattern” (“patrón”), entre otros. Mientras que las palabras más frecuentes de los participantes bajo la condición de PCB estaban relacionadas con un estado de vigilia y tranquilidad: “relax” (“relajar”), “daydream” (“soñar despierto”), “asleep” (“adormecido”), “calm” (“calma”), etc.

Basados en este resultado y en estudios anteriores que aplicaron modelado de tópicos a informes retrospectivos de experiencias con drogas psicoactivas (Zamberlan et al., 2018), seleccionamos diez términos relacionados con el estado psicodélico y calculamos su similitud semántica promedio con las transcripciones (Ver Sección 3.2.2). Participantes en la condición de PCB presentaron aumentos significativos con el término “mood” (“estado de ánimo”) para ambos momentos (Tiempo 1: $Z = 3.36$, $p = 8 \times 10^{-4}$; Tiempo 2: $Z = 3.04$; $p = 0.0024$). Mientras que para el grupo bajo la condición de LSD, se encontraron aumentos significativos con el término “fear” (“miedo”) ($Z = 2.84$, $p =$

FIGURA 3.2: Nubes de palabras de los términos más frecuentes utilizados en las transcripciones bajo la condición de LSD (izquierda) y PCB (derecha). El tamaño de las palabras es proporcional a su frecuencia de aparición en las transcripciones.

0.0046) para el Tiempo 2. El resto de las comparaciones exhibieron p-valores mayores a 0.03 (Tiempo 1) y 0.007 (Tiempo 2) que resultaron no significativos al considerar correcciones de Bonferroni ($\alpha = 0.005$).

En la Figura 3.3, se muestran diagramas de caja de la similitud semántica promedio entre las transcripciones y los cinco términos que presentaron p-valores menores a 0.05 en (como mínimo) uno de los momentos. Al comparar las dos condiciones, se observa que para ambos Tiempos, los participantes bajo los efectos psicodélicos presentaron aumentos en la similitud semántica con las palabras "fear" ("miedo"), "ego" ("ego") y "reality" ("realidad"); mientras que disminuyeron su semejanza con los términos "mood" ("estado de ánimo") y "relax" ("relajación"), respecto a la condición de PCB.

3.3.2. Resultados no semánticos

La inspección visual de los grafos de discurso mostró marcadas diferencias asociadas a la cantidad y organización tanto de nodos como de enlaces. A modo de ejemplo, en la la Figura 3.4.A se presentan los grafos de dos participantes para ambas condiciones. La comparación de las métricas reveló que el LSD produjo aumentos en la verbosidad (WC) en ambos momentos (Tiempo 1: $Z = 3.47$, $p = 5 \times 10^{-4}$; Tiempo 2: $Z = 4.38$, $p = 1 \times 10^{-5}$) e incrementos en las conexiones entre nodos (ATD) dando $p < 1 \times 10^{-4}$ para

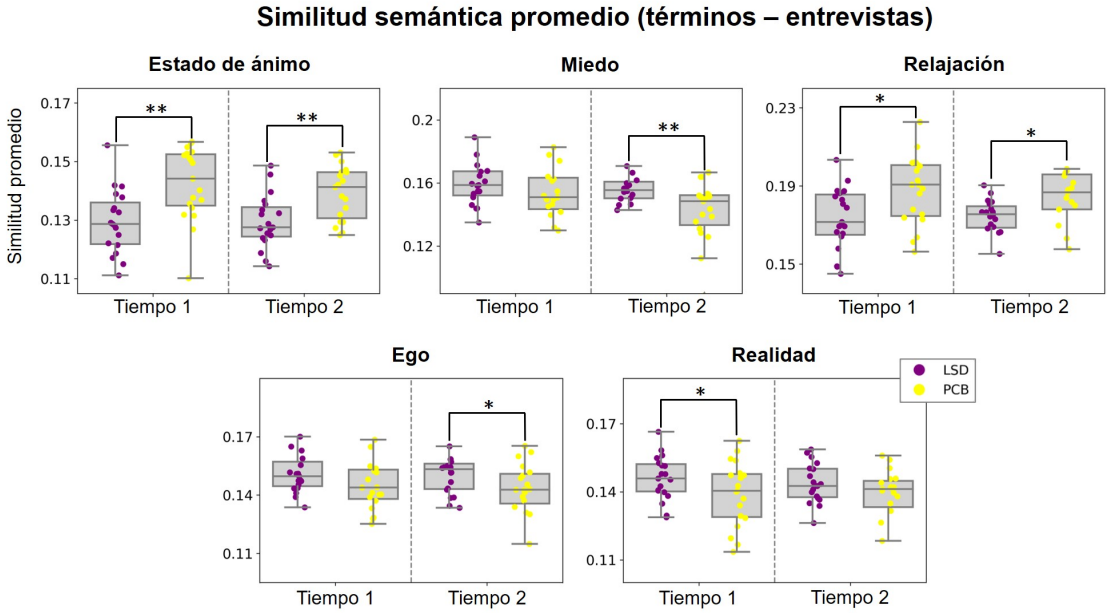

FIGURA 3.3: Gráficos de caja para las métricas semánticas que presentaron diferencias entre grupos para (como mínimo) uno de los Tiempos (* $p < 0.05$ sin correcciones y ** $p < 0.05$ con correcciones de Bonferroni; test de rango con signo de Wilcoxon).

ambos tiempos. En simultáneo, se observaron reducciones significativas en la cantidad de palabras diferentes relativas a la cantidad de palabras totales de cada entrevista ($\frac{N}{WC}$) tanto para el Tiempo 1 ($Z = 3.90$, $p = 1 \times 10^{-4}$) como para el 2 ($Z = 3.98$, $p = 7 \times 10^{-5}$). En la Figura 3.4.B, se muestran las distribuciones de cajas correspondientes a estas medidas para cada Tiempo y condición.

Las comparaciones de las métricas normalizadas por los grafos aleatorios (Sección 3.2.3), no mostraron diferencias significativas resultando en p-valores mayores a 0.07 (Tiempo 1) y 0.005 (Tiempo 2) que no superaron correcciones de Bonferroni. La tendencia general observada para la condición psicodélica consistió en incrementos para las métricas relacionadas con ciclos (L1, L2 y L3). Además, se obtuvieron aumentos en el CC y LSC, mostrando nuevamente una mayor conectividad en los grafos de discurso. En contra parte, se encontraron reducciones en el diámetro para ambos momentos y en el ASP sólo para el Tiempo 1.

3.3.3. Medidas de desorganización directa

Los participantes bajo los efectos del LSD exhibieron aumentos significativos de entropía (Tiempo 1: $Z = 3.24$, $p = 0.0012$; Tiempo 2: $Z = 4.04$, $p = 5 \times 10^{-5}$), variabilidad semántica (Tiempo 2: $Z = 2.05$, $p = 0.04$) y rango (Tiempo 2: $Z = 2.26$, $p = 0.02$)

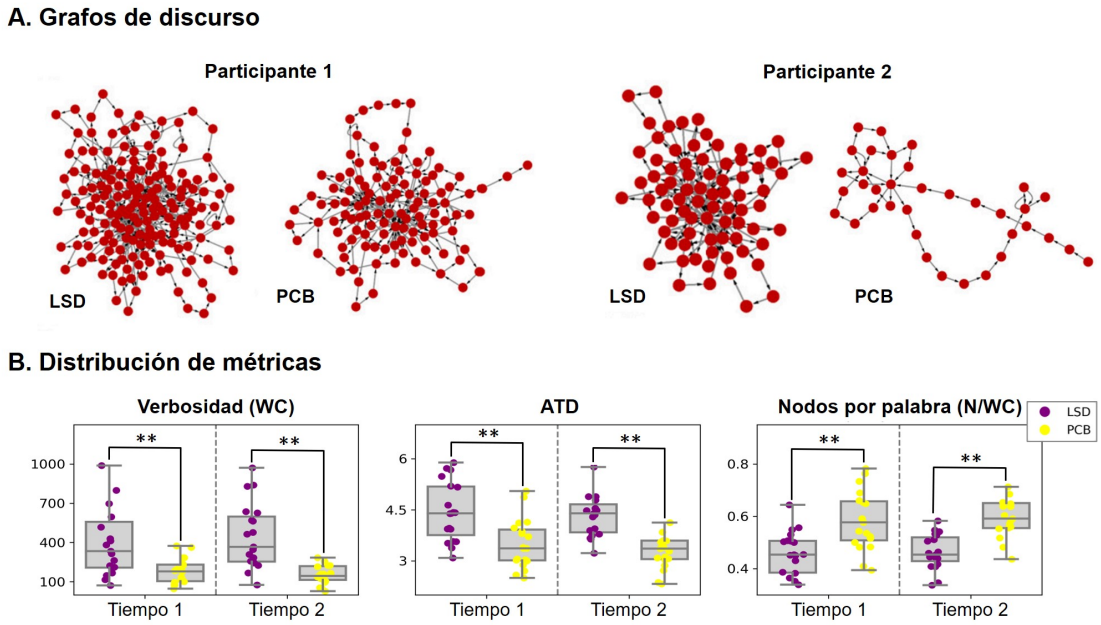

FIGURA 3.4: **A.** Grafos de discurso de dos participantes en el Tiempo 1 para la condición de LSD (izquierda) y PCB (derecha). **B.** Gráficos de caja para las métricas topológicas que presentaron diferencias significativas entre grupos (* $p < 0.05$ sin correcciones y ** $p < 0.05$ con correcciones de Bonferroni; test de rango con signo de Wilcoxon).

frente a la condición de PCB (Ver Figura 3.5). Sin embargo, solo la medida de entropía resultó significativa al aplicar correcciones de Bonferroni ($\alpha = 0.017$).

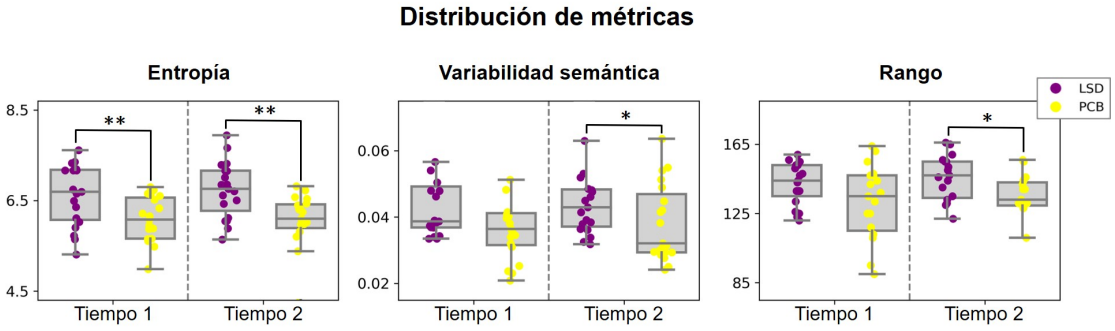

FIGURA 3.5: Gráficos de caja para métricas de desorganización directa: **A.** Entropía, **B.** Variabilidad semántica y **C.** Rango. (* $p < 0.05$ sin correcciones y ** $p < 0.05$ con correcciones de Bonferroni; test de rango con signo de Wilcoxon).

3.3.4. Clasificación

En la Figura 3.6, se presentan los histogramas normalizados de los valores AUC para las *features* semánticas (izquierda) y no semánticas (derecha) en el Tiempo 1. Las distribuciones en negro y azul corresponden a los resultados obtenidos con la etiqueta de la

condición real y aleatoria, respectivamente. Las *features* semánticas exhibieron un valor AUC medio de 0.757 ± 0.0003, en comparación con 0.507 ± 0.004 obtenido de la distribución azarosa ($p = 0.015$). Mientras que para las *features* no semánticas, se encontró un valor AUC medio de 0.742 ± 0.0002, en comparación con 0.499 ± 0.004 de la distribución aleatoria ($p = 0.012$).

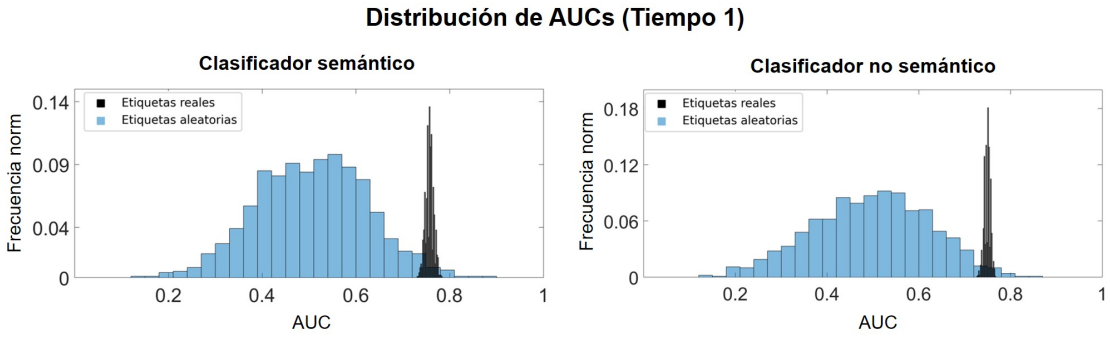

FIGURA 3.6: Histogramas normalizados de los valores AUC obtenidos con los clasificadores basados en las *features* semánticas (izquierda) y no semánticas (derecha), para el Tiempo 1. Los colores indican si la asignación de etiquetas correspondía a la real (negro) o la aleatoria (azul).

3.4. Discusión

En este trabajo, examinamos métricas semánticas y no semánticas del lenguaje en entrevistas conducidas bajo los efectos del LSD y el estado de vigilia (PCB). Ambos conjuntos de medidas discriminaron las condiciones con una exactitud comparable y estadísticamente diferenciable de los modelos nulos basados en la distribución aleatoria de clases. Además, los análisis relativos a la entropía, variabilidad semántica y rango, revelaron que el estado psicodélico se caracteriza por un discurso desorganizado de coherencia reducida y contenido diversificado.

El análisis semántico permitió efectuar un estudio del contenido presente en el discurso que excede las palabras específicas de las transcripciones. En efecto, la ventaja de considerar la similitud semántica en lugar de la ocurrencia de palabras radica en que incluso cuando el término particular no se utiliza con frecuencia o no se menciona explícitamente, la similitud semántica media entre el término y la entrevista se incrementará si el resto de las palabras presentan un significado similar entre sí (cercanas en el espacio del *embedding*) (Raskovsky et al., 2010). En consecuencia, términos que no fueron frecuentemente utilizados en las transcripciones bajo la condición de LSD (e.g. "ego", "miedo", "realidad") mostraron una similitud semántica promedio mayor que para la condición de PCB. Como se mencionó previamente, esta lista de palabras se extrajo de un estudio que examinaba un amplio corpus de experiencias psicodélicas en ambientes arbitrarios y cantidades de sustancia psicoactiva diversos (Zamberlan et al., 2018). Probablemente, estos términos no tuvieron una mención frecuente en el conjunto de transcripciones debido a las limitaciones impuestas en el experimento, es decir, ambientes de laboratorio controlado y tareas específicas a cumplir. Nuestros resultados coinciden con los conocidos efectos subjetivos de otros psicodélicos serotonérgicos (Nichols, 2016; Nour et al., 2016; Kometer and Vollenweider, 2016; Preller and Vollenweider, 2016) y son consistentes con estudios previos de asociación semántica (Spitzer et al., 1996; Family et al., 2016), así como con las investigaciones asociadas a los efectos neurofisiológicos del LSD (Carhart-Harris et al., 2016; Tagliazucchi et al., 2016). En cambio, la condición de PCB se caracterizó por palabras asociadas a un estado de adormecimiento, evidenciado por su similitud semántica con "relajación" y la alta frecuencia de palabras como "adormecido" y "soñar despierto". También mostraron una similitud semántica mayor con "estado de ánimo", suponemos que la caída del estado anímico de los sujetos bajo la condición de LSD se debe a que el experimento involucraba estudios estáticos prolongados con máquinas que pueden resultar arduos y agotadores bajo los efectos de la sustancia.

El análisis de los grafos de discurso, permitió estudiar métricas no semánticas de las transcripciones mostrando que el habla bajo los efectos psicodélicos se encuentra caracterizado tanto por la verbosidad excesiva como por la reducción de la cantidad de

palabras distintas relativas a la cantidad de palabras totales. La normalización de las métricas con grafos aleatorios permitió descartar efectos asociados a la cantidad de palabras y reveló incrementos en medidas locales relacionadas con ciclos, reflejando discursos repetitivos. También encontramos aumentos en medidas globales asociadas a la conectividad, resultado que presenta concordancia con investigaciones previas de configuraciones cerebrales y su expansión en el estado psicodélico (Tagliazucchi et al., 2014; Schartner et al., 2017; Atasoy et al., 2017; Lord et al., 2019).

Estudios anteriores basados en neuroimágenes (Carhart-Harris et al., 2014; Carhart-Harris and Friston, 2019) mostraron que el estado psicodélico se caracteriza por un aumento de entropía en la actividad cerebral. Según este modelo (Carhart-Harris, 2018), los incrementos de entropía deberían manifestarse tanto a nivel neurofisiológico como a nivel de experiencia subjetiva. En la medida en que el lenguaje desorganizado se interprete como un reflejo del desorden de procesos de pensamiento generados por el funcionamiento anormal del cerebro, los resultados reportados en este trabajo contribuyen y concuerdan con el modelo entrópico. Investigaciones relativas al aumento de la asociación libre con psilocibina (Spitzer et al., 1996) y pensamientos hiperasociativos (Family et al., 2016) que podrían potenciar la creatividad (Girn et al., 2020), concuerdan con los incrementos de variabilidad semántica y rango. Estas métricas reflejan que el discurso bajo los efectos del LSD contiene alteraciones del flujo del lenguaje tanto por "saltos" repentinos en el contenido semántico como por su falta de compresibilidad (i.e. diversificación de conceptos). Cabe señalar que las diferencias en variabilidad y rango fueron más notorias para el Tiempo 2, implicando que estas *features* podrían ser específicas de las últimas etapas de los efectos del LSD ($\sim$ 225 min después de la infusión) que presentan un componente dopaminérgico (y tal vez psicotomimético) más fuerte (Marona-Lewicka et al., 2005).

Los efectos del LSD en la conciencia humana sobrepasan los paradigmas conductuales clásicos. Sin embargo, el análisis del lenguaje presenta un enfoque particularmente útil y escalable para la cuantificación de las experiencias subjetivas. La flexibilidad de aplicación, interpretabilidad de resultados, automatización, objetividad y rentabilidad lo constituyen como una herramienta óptima para la detección y predicción del estado psicodélico. Teniendo en cuenta las ventajas de este enfoque, futuros experimentos podrían explorar los efectos del LSD en tareas de lenguaje asociadas a la descripción de multimedia (imágenes, fragmentos de películas, audios, etc) o narraciones de experiencias personales (recuerdos, sueños, motivaciones, etc). Además, el análisis de diversas sustancias permitiría verificar si las características inferidas en este estudio son exclusivas del estado psicodélico. No obstante, consideramos que nuestro trabajo funda una línea de análisis y una base necesaria para cuantificar relatos de experiencias subjetivas.

En resumen, caracterizamos el lenguaje natural bajo los efectos del LSD. Encontramos que esta sustancia genera un discurso desorganizado en contenido y estructura que, probablemente, refleje el desorden en los procesos de pensamientos. Investigaciones que sigan esta línea de trabajo deberían incrementar el tamaño de las muestras (i.e. cantidad de participantes) y variar la dosificación para caracterizar el estado psicoactivo en diferentes rangos. Nuestros resultados se acoplan al modelo del cerebro entrópico, que propone a los agonistas del receptor $5-HT_{2A}$ como agentes que aumentan la entropía de la actividad cerebral y los procesos de pensamiento asociativos (Carhart-Harris, 2018). Mientras que los cambios referidos al contenido semántico del discurso son esperables, el presente trabajo muestra que los rasgos no semánticos también son capaces de distinguir la condición de LSD frente a la de PCB con una exactitud y rendimiento comparables.

Capítulo 4

El lenguaje natural bajo los efectos de la microdosis con psilocibina

Motivados por los resultados anteriores, nos propusimos estudiar el lenguaje natural bajo los efectos de reducidas dosificaciones (i.e. microdosis) de psicodélicos serotonérgicos. Este capítulo corresponde a contenidos adaptados de "Natural language signatures of psilocybin microdosing" (Sanz et al., 2022a) en donde estudiamos estos efectos.

La microdosificación con psicodélicos es una práctica basada en el consumo de cantidades reducidas (próximas al umbral de percepción) de sustancias serotonérgicas, típicamente alrededor de 10 %/20 % de una dosis completa (Fadiman and Korb, 2019; Kuypers et al., 2019; Ona and Bouso, 2020; Polito and Stevenson, 2019). A diferencia de dosis medianas/altas, se cree que la microdosis produce efectos agudos mínimos que pueden sostenerse durante un período de tiempo. En consecuencia, esta práctica requiere un programa donde se intercalan días de consumo con días de descanso (Kuypers et al., 2019). A pesar del estatus ilegal de los psicodélicos en la mayoría de los países, la microdosificación ganó una creciente popularidad en los últimos años, siendo el LSD y la psilocibina (en forma de hongos psicoactivos) los compuestos más consumidos (Hutten et al., 2019a; Lea et al., 2020b; Polito and Stevenson, 2019; Szigeti et al., 2021). Diversos sitios web contienen tanto discusiones como sugerencias relacionadas a esta práctica; numerosos usuarios afirman que la miscrodosificación incentiva la productividad, estimula las funciones cognitivas y mejora tanto el estado de ánimo como la concentración mental (Anderson et al., 2019b; Lea et al., 2020a). Además, algunos consumidores utilizan la miscrodosis para auto-tratar trastornos mentales como la depresión o la ansiedad

(Hutten et al., 2019b; Lea et al., 2020b,c), y fue sugerido para el uso clínico (Kuypers, 2020).

Múltiples estudios observacionales basados en encuestas, respaldan los efectos positivos de la microdosificación (Anderson et al., 2019a; Hutten et al., 2019a; Johnstad, 2018; Lea et al., 2020b; Polito and Stevenson, 2019; Prochazkova et al., 2018; Rootman et al., 2021). Sin embargo estas investigaciones carecen de controles adecuados (e.g. programa implementado, cantidad y tipo de sustancia, etc), se limitan a informes de preguntas preseleccionadas e ignoran los efectos provenientes de las ideas previas (expectativas) que los encuestados tienen respecto a la microdosificación. Dado que los efectos de esta práctica son sutiles, los consumidores pueden estar sugestionados a experimentar determinadas sensaciones (Kaertner et al., 2021; Olson et al., 2020). Por el contrario, investigaciones con diseños experimentales doble-ciego y controlados por una condición de placebo, mostraron menor apoyo a los efectos positivos de la microdosificación (Bershad et al., 2019; Family et al., 2020; Hutten et al., 2020; Szigeti et al., 2021; van Elk et al., 2021; Yanakieva et al., 2019). No obstante, estos estudios suelen incluir tareas y cuestionarios validados para dosis psicodélicas medianas/altas, que podrían carecer de especificidad y sensibilidad para identificar efectos más sutiles. De estos inconvenientes, surge la necesidad de desarrollar métodos que permitan estudiar las experiencias subjetivas producidas por dosificaciones bajas de sustancias psicodélicas.

Como se mencionó anteriormente (Capítulo 1), el NLP se constituye como un enfoque alternativo capaz de identificar automáticamente *features* lingüísticas informativas del discurso que pueden utilizarse para distinguir condiciones experimentales (Agurto et al., 2020; Bedi et al., 2014; Corcoran et al., 2018; Norel et al., 2020; Sanz et al., 2021, 2022b). Además de informar sobre los contenidos de la experiencia provocada por la sustancia, el NLP permite investigar la modulación de la producción del lenguaje en sí, revelando información asociada a la acción de la droga que excede los relatos individuales (Tagliazucchi, 2022). Diversas investigaciones aplicaron esta herramienta para estudiar efectos subjetivos de compuestos (Bedi et al., 2014; Cox et al., 2021; Cox and Johnson, 2021; Sanz et al., 2018, 2021; Zamberlan et al., 2018), sin embargo (hasta la fecha de publicación de nuestro estudio) no se registraban investigaciones que utilicen técnicas de NLP para caracterizar la microdosificación.

En este trabajo analizamos el lenguaje natural bajo los efectos de hongos Psilocybe Cubensis ($0.5g$ de material seco) (Polito and Stevenson, 2019; Szigeti et al., 2021; van Elk et al., 2021). Los datos se extrajeron de un experimento doble-ciego controlado por una condición de placebo y de dos semanas de duración (Cavanna et al., 2022). Durante cada semana, los participantes recibieron la dosis activa o el placebo y fueron entrevistados sobre su experiencia. Basados en nuestros resultados anteriores (Sanz et al., 2021),

analizamos dos métricas del lenguaje: la verbosidad (longitud total de las entrevistas, en palabras) y la variabilidad semántica (ver Sección 2.2.1). También investigamos la polaridad sentimental (i.e uso de términos vinculados a sentimientos positivos, negativos o neutrales; ver Sección 2.3) para cuantificar los supuestos efectos de la microdosificación en el estado de ánimo (Cameron et al., 2020; Hutten et al., 2020; Lea et al., 2020b,a; Polito and Stevenson, 2019). Considerando investigaciones anteriores, hipotetizamos que la microdosis psicodélica disminuye la coherencia semántica y aumenta tanto la verbosidad como la positividad del discurso. Además, evaluamos si estas métricas presentaban diferencias entre los participantes que identificaron correctamente su condición experimental (grupo "no cegado") y los que no (grupo "cegado"). Finalmente, implementamos pruebas estadísticas y herramientas de ML para evaluar si las *features* lingüísticas consideradas podían utilizarse para distinguir la condición y el grupo de los participantes.

4.1. Participantes y protocolo

Treinta y cuatro voluntarios de lengua nativa hispana (11 mujeres: 32.09 ± 3.53 años; 23 hombres: 30.87 ± 4.64 años) fueron convocados para este estudio por medio de redes sociales y comunicación oral (de boca en boca). Los participantes constaban de 11 ± 14.9 experiencias previas con psicodélicos serotonérgicos, de las cuales 1.5 ± 2.3 fueron clasificadas como desafiantes. Los principales criterios de exclusión involucraban presentar: diagnósticos previos de trastornos psicóticos, bipolares (tipo 1 o 2), depresivos, obsesivos compulsivos, de ansiedad, de pánico, neurológicos, distimia, entre otros. También se excluyeron participantes que padecían bulimia, anorexia y/o abuso/dependencia de sustancias en los últimos 5 años (exceptuando la nicotina). Además, en el transcurso experimental, se solicitó evitar el consumo drogas psicoactivas, alcohol y cafeína.

El experimento duró dos semanas, los participantes recibieron cápsulas por vía oral de Psilocybe Cubensis seco ($0.5g$) y el mismo peso de un hongo comestible. Todos los participantes transitaron las dos condiciones experimentales, es decir, una de las semanas consumieron la microdosis de psilocibina (dosis activa) y la otra el placebo. La asignación de condiciones fue efectuada por un miembro del equipo ajeno al experimento.

Los miércoles y viernes de cada semana, los participantes acudieron a las instalaciones del laboratorio y consumieron alguna de las cápsulas. Los viernes, aproximadamente 2,30 hs después de la dosis, un profesional de salud mental miembro del equipo de investigación entrevistó a los participantes. Dado que el programa de los miércoles incluía efectuar numerosas tareas que no fueron analizadas en este estudio (Cavanna et al., 2022), las entrevistas se realizaron únicamente los viernes para evitar el agotamiento. Estas incluyeron las siguientes preguntas:"¿Cómo se siente en este momento?"; "¿Se siente de

acuerdo con lo que esperaba?"; "¿Siente cambios en su percepción?"; "¿Siente cambios en su estado de ánimo?"; "¿Siente cambios en su nivel de imaginación o creatividad?"; "¿Siente cambios en su nivel de atención, alerta o energía?". En los análisis posteriores, nos referimos a estas preguntas como: "sensación", "expectativa", "percepción", "estado de ánimo", "creatividad" y "estado de alerta", respectivamente. Al finalizar las entrevistas, los participantes debían identificar la condición correspondiente a esa semana (dosis activa o placebo). No se solicitaron justificaciones de respuesta y no se reveló la condición real. En la Figura 4.1 se muestra una representación esquemática del diseño experimental.

FIGURA 4.1: Diseño experimental doble-ciego controlado por una condición de placebo. Las condiciones experimentales de los participantes se distribuyeron aleatoriamente entre las semanas. Los investigadores ignoraron el contenido de las cápsulas hasta la fase de análisis. El siguiente esquema se repitió en ambas semanas: el miércoles y el viernes, los participantes consumían la dosis. Sólo el viernes se realizaban las entrevistabas y se pedía identificar la condición.

Este estudio se realizó de acuerdo con la declaración de Helsinki y fue aprobado por el Comité de Ética en Investigación de la Universidad Abierta Interamericana (UAI) de Buenos Aires, Argentina (Número de protocolo: 0-1054). La medición y el análisis del discurso formaban parte de un protocolo más amplio pre-registrado en www.clinicaltrials.gov, número de identificación: NCT05160220. Todos los participantes dieron su consentimiento informado por escrito y no recibieron compensaciones económicas por colaborar en el estudio.

4.2. Caracterización química de las muestras

La microdosis de Psilocybe Cubensis estaba compuesta por $0.5g$ de material seco molido proveniente de tres fuentes independientes. Para determinar la concentración de los cuatro alcaloides principales (psilocibina, psilocina, baeocistina y norbaeocistina), se aislaron $150mg$ de cada fuente y se enviaron al Laboratorio de Análisis Forense de

Sustancias Biológicamente Activas de la Universidad de Química y Tecnología de Praga, República Checa.

El análisis de las muestras reveló concentraciones de: $640.2\mu g/g$ (psilocibina), $950.7\mu g/g$ (psilocina), $50.4\mu g/g$ (baeocistina) y $12.5\mu g/g$ (norbaeocistina). En consecuencia, la dosis activa constaba de $0.32mg$ de psilocibina, $0.48mg$ de psilocina, $0.025mg$ de baeocistina y $0.0063mg$ de norbaeocistina. Dado que transcurrió al menos un mes entre el final del experimento y el análisis químico, a pesar de conservar las muestras en condiciones óptimas, parte del material psicoactivo (i.e. psilocibina y psilocina) puede haberse perdido durante este período de tiempo (Gotvaldova et al., 2021).

4.3. Métodos

Las entrevistas fueron grabadas y trascriptas manualmente por un técnico con formación lingüística que desconocía la condición experimental de los participantes. Posteriormente, un miembro del equipo de investigación realizó un control de calidad al examinar las transcripciones. Al igual que en el estudio anterior, se descartaron las preguntas e intervenciones del entrevistador y sólo se analizó el discurso de los participantes.

Las trascripciones se etiquetaron con dos categorías principales, la primera correspondía a la condición experimental (dosis activa o placebo) y la segunda dependía de si el participante había identificado la condición correctamente. Aquellos participantes que (en ambas semanas) distinguieron el contenido de la cápsula se etiquetaron como “no cegados”, mientras que el resto se categorizó como “cegados”. De esta forma, se obtuvieron cuatro subgrupos secundarios:

1. Dosis activa vs Placebo restringido al grupo cegado.
2. Dosis activa vs Placebo restringido al grupo no cegado.
3. Cegado vs No cegado restringido a dosis activa.
4. Cegado vs No cegado restringido a placebo.

Además, las respuestas dadas en las entrevistas se analizaron por separado para determinar qué tópicos particulares se veían afectados por la microdosificación.

4.3.1. Verbosidad

La verbosidad se calculó contando el número de palabras (incluidas repeticiones y términos de uso frecuente) de las repuestas dadas por cada participante en las entrevistas.

4.3.2. Variabilidad semántica

En el estudio anterior (Capítulo 3), observamos que una dosis media/alta de LSD generaban una reducción de coherencia en el discurso. Para verificar si este efecto era característico de la microdosificación con psilocibina, se computó la variabilidad semántica de las respuestas de los participantes (ver apéndice B.2). Dado que las entrevistas estaban en español, el pre-procesamieto se realizó con la librería TreeTagger de Python entrenado con la base de datos de AnCora (`http://clic.ub.edu/corpus/es/ancora`) y la lista de palabras de uso frecuente se extrajo de la librería NLTK. Para el *embedding* se utilizó el modelo *FastText* (`https://fasttext.cc/`) pre-entrenado con la base de datos de *Common Crawl* y Wikipedia (en español). Este modelo contiene alrededor de $2,000,000$ de términos únicos y un espacio vectorial de dimensión 300.

4.3.3. Análisis de sentimiento

La polaridad de las entrevistas (ver Sección 2.3) se determinó con el modelo $senti-py$ pre-entrenado (a nivel frase) para el español (`github.com/aylliote/senti-py`). En la primera instancia, el modelo realiza un pre-procesamiento estándar de texto que incluye: cambiar mayúsculas por minúsculas, eliminar acentos, convertir verbos conjugados a su infinitivo, etc. En la segunda instancia, aplica un selector de *features* univariadas y un clasificador *Multinomial Naive Bayes* entrenado con datos de páginas web que ofrecen servicios y permiten a los usuarios tanto comentar como puntuar su experiencia (e.g. "TripAdvisor", "Pedidos Ya", "Mercado Libre", entre otros). Los datos de entrenamiento incluían alrededor de 1 millón de muestras. Los parámetros e hiper-parámetros del clasificador se optimizaron con la técnica *Grid Search CV* donde la validación cruzada constaba de 10 subconjuntos (K-fold, K= 10). Como resultado, el modelo puntúa cada frase en un rango de 0 (negativa) a 1 (positiva), donde ~ 0.5 implica una frase de polaridad neutra.

La transcripción de cada respuesta contenida en las entrevistas se dividió en frases que fueron utilizadas como entradas en $senti-py$ y la polaridad de las respuestas se estimó con la puntuación sentimental promedio (MSS) de sus frases.

4.3.4. Análisis estadísticos

Para determinar diferencias en las distribuciones de verbosidad, variabilidad semántica y MSS se aplicaron pruebas no paramétricas contenidas en la librería scipy (`https://scipy.org`) de Python. Tanto para las comparaciones entre la dosis activa y el placebo como para sus correspondientes subgrupos, se aplicaron pruebas de rango con signo

de Wilcoxon. Mientras que se utilizaron pruebas U de Mann-Whitney tanto para la comparación entre cegado y no cegado como para sus subgrupos. Considerando el número de preguntas contenidas en las entrevistas ($N = 6$) se modificó el nivel de significancia estadística (0.05) con correcciones de Bonferroni ($\alpha = 0.0083$)

La relación entre el tamaño de muestra y la potencia se realizó de forma análoga al estudio anterior (ver Sección 3.2.5). Confirmamos que $n = 34$ constituía una potencia estadística superior a 0.8; equivalente a la probabilidad de error tipo 2 (β) < 0.2

4.3.5. Clasificadores

Cada entrevista se identificó con 12 *features* dadas por las puntuaciones de verbosidad y el MSS de sus preguntas, la variabilidad semántica se excluyó de los análisis por no presentar diferencias significativas entre condiciones (ver apéndice B.2). Utilizando la librería sckit-learn de Python y el procedimiento descripto en la Sección 2.6, se entrenaron clasificadores *Ramdom Forest* con validación cruzada de 3 subconjuntos. El rendimiento del modelo se determinó calculando el AUC y el p-valor para 100 iteraciones. En total, se implementaron seis modelos (con idénticos parámetros), dos de ellos entrenados para distinguir las condiciones experimentales (dosis activa frente a placebo) y cegado frente a no cegado. Los cuatro restantes fueron entrenados para realizar la misma clasificación pero restringida a los subgrupos.

4.4. Resultados

4.4.1. Dosis activa frente a placebo

Los participantes bajo la dosis activa mostraron una mayor verbosidad que para la condición de placebo en todas sus respuestas (Figura 4.2A). Las comparaciones estadísticas revelaron diferencias significativas en las preguntas asociadas a: “percepción” ($T = 67$, $p = 0.0034$, $r = 0.67$), “estado de ánimo” ($T = 51$, $p = 0.0016$, $r = 0.75$) y “estado de alerta” ($T = 79$, $p = 0.0047$, $r = 0.67$). El resto de las respuestas resultaron en p-valores mayores a $p = 0.026$ que no superaron correcciones de Bonferroni. El grupo bajo la dosis activa presentó valores de mediana y rango intercuartil de: $M = 63.5$, $IQR = 26.25$ (“percepción”); $M = 50$, $IQR = 85.5$ (“estado de ánimo”) y $M = 65$, $IQR = 60.5$ (“estado de alerta”). Mientras que para el grupo bajo la condición de placebo se obtuvo: $M = 10.5$, $IQR = 39.75$ (“percepción”); $M = 10$, $IQR = 19.25$ (“estado de ánimo”) y $M = 33$, $IQR = 63.5$ (“estado de alerta”).

Los valores medios de MSS para los participantes bajo la dosis activa oscilaron alrededor de 0.4, indicando una polaridad neutra (Figura 4.2 B). Sin embargo, los valores medios de la microdosificación fueron más elevados que bajo la condición de placebo para todas las respuestas. Únicamente la pregunta asociada a "percepción" mostró diferencias significativas ($T = 72$, $p = 0.0039$, $r = 0.65$) entre grupos, las demás exhibieron p-valores mayores a 0.031 que no superaron correcciones de Bonferroni. El valor de la mediana y rango intercuartil del grupo para la dosis activa fue de: $M = 0.37$, $IQR = 0.11$. Mientras que para los participantes bajo la condición de placebo se encontró: $M = 0.22$, $IQR = 0.19$.

En conjunto, estos resultados indican que una parte de los efectos de la dosis activa se reflejó en el discurso de los participantes generando respuestas de mayor longitud con una polaridad de tendencia positiva.

La clasificación de condiciones experimentales dio como resultado un AUC de 0.79 ± 0.04; que fue significativamente superior ($p < 0.01$) al obtenido con la asignación aleatoria de clases (0.52 ± 0.10). En la Figura 4.2.C se muestran histogramas con los valores AUC de todas las iteraciones.

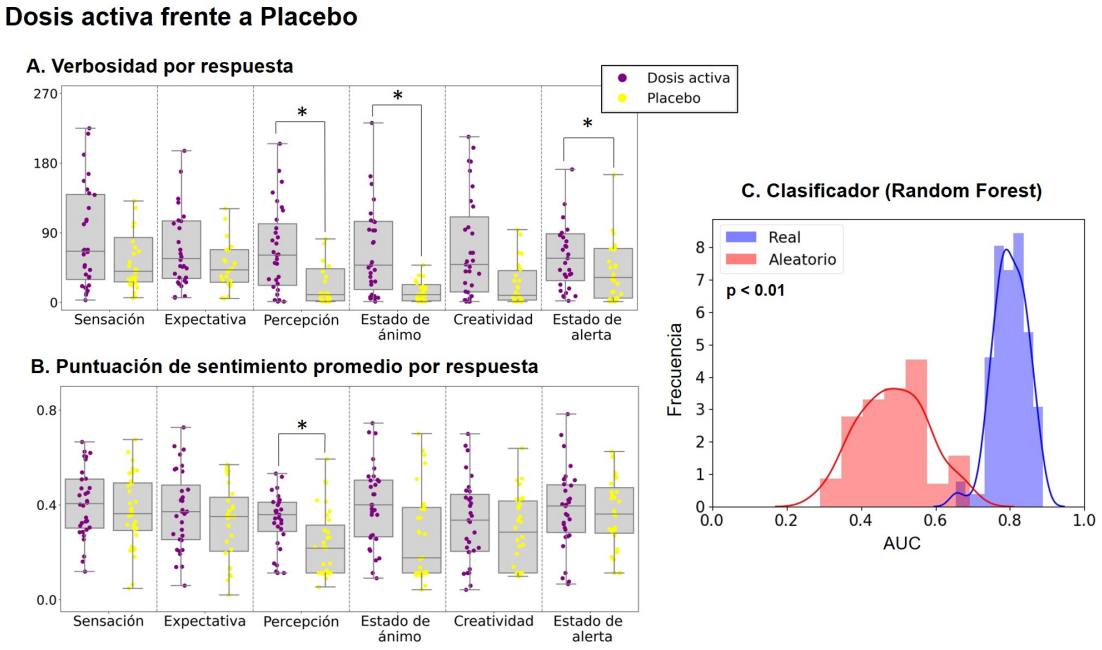

FIGURA 4.2: La dosis activa incrementó la verbosidad y la puntuación de sentimiento promedio (MSS). (A) Gráficos de cajas de la verbosidad para cada pregunta y condición (*$p < 0.05$, con correcciones de Bonferroni; prueba de rango con signo de Wilcoxon). (B) Misma información que en el panel A pero para el MSS. (C) Histogramas de los valores AUC obtenidos de la clasificación real (azul) y aleatoria (rojo).

4.4.2. Cegado frente a No cegado

En total, 18 participantes identificaron la condición experimental (grupo no cegado) y 16 fallaron en esta tarea (grupo cegado). Las comparaciones por pares no resultaron en diferencias significativas para ninguna de las respuestas. Se encontraron p-valores mayores a 0.013 y 0.10 para la verbosidad y el MSS, respectivamente. El grupo cegado obtuvo valores medios de verbosidad más elevados (Figura 4.3.A) y MSS más bajos (Figura 4.3.B) que los del no cegado.

Los resultados de los clasificadores (Figura 4.3.C) mostraron correspondencia con la ausencia de diferencias significativas entre grupos. La clasificación con las etiquetas reales resultó en un AUC de 0.53 ± 0.06 que no difirió ($p = 0.37$) del obtenido con la asignación aleatoria de clases (0.49 ± 0.11).

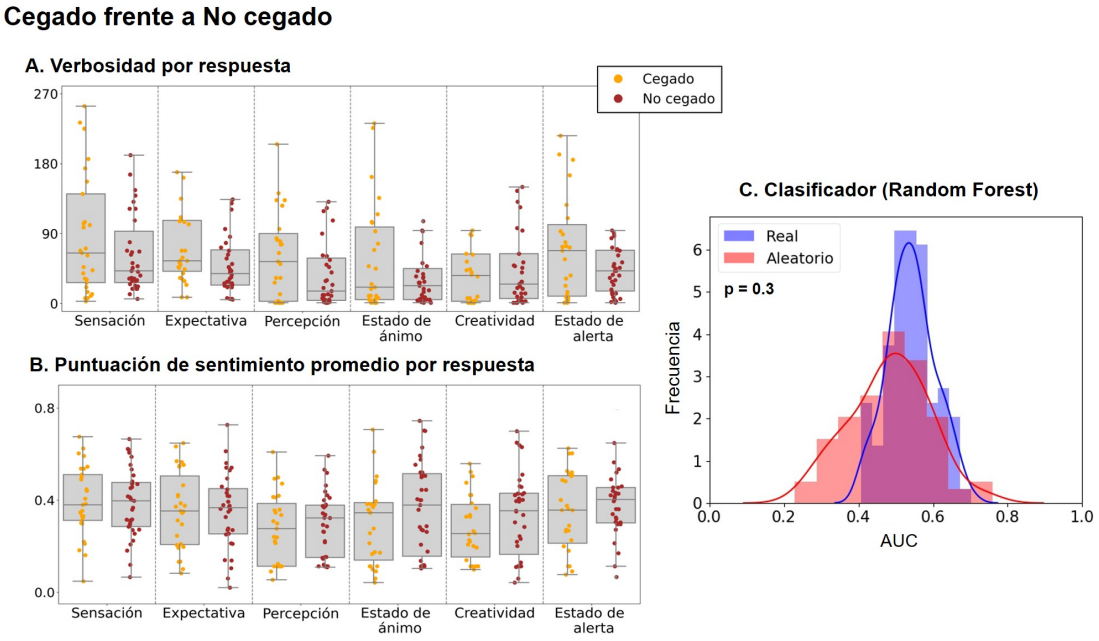

FIGURA 4.3: Los análisis de verbosidad y de MSS no mostraron diferencias significativas entre participantes cegados y no cegados. (A) Gráficos de cajas de la verbosidad para cada pregunta y condición ($^*p < 0.05$, con correcciones de Bonferroni; prueba U de Mann-Whitney). (B) Misma información que en el panel A pero para el MSS. (C) Histogramas de los valores AUC obtenidos de la clasificación real (azul) y aleatoria (rojo).

4.4.3. Subgrupos

Tanto los análisis estadísticos (apéndice B.1) como las clasificaciones de ML (Figura 4.4), se repitieron para cada subgrupo. El clasificador entrenado para distinguir la dosis activa frente a la condición de placebo, restringido a los participantes cegados (Figura 4.4.A), resultó en un valor AUC de 0.48 ± 0.13. El mismo no mostró diferencias significativas

($p = 0.56$) del valor obtenido con la asignación aleatoria de clases (0.52 ± 0.18). Por el contrario, al restringir la comparación de condiciones a participantes no cegados (Figura 4.4.B), la clasificación superó significativamente el nivel de azar ($p < 0.01$). Utilizando las etiquetas reales se obtuvo un valor AUC de 0.92 ± 0.05, mientras que con la asignación aleatoria de clases el valor fue de 0.51 ± 0.17.

Las clasificaciones conducidas para el grupo cegado frente a no cegado, restringido a las condiciones experimentales (Figuras 4.4.C y 4.4.D), no mostraron diferencias estadísticas significativas respecto a los resultados obtenidos con la asignación aleatoria de etiquetas ($p = 0.09$ para la dosis activa y $p = 0.22$ para el placebo).

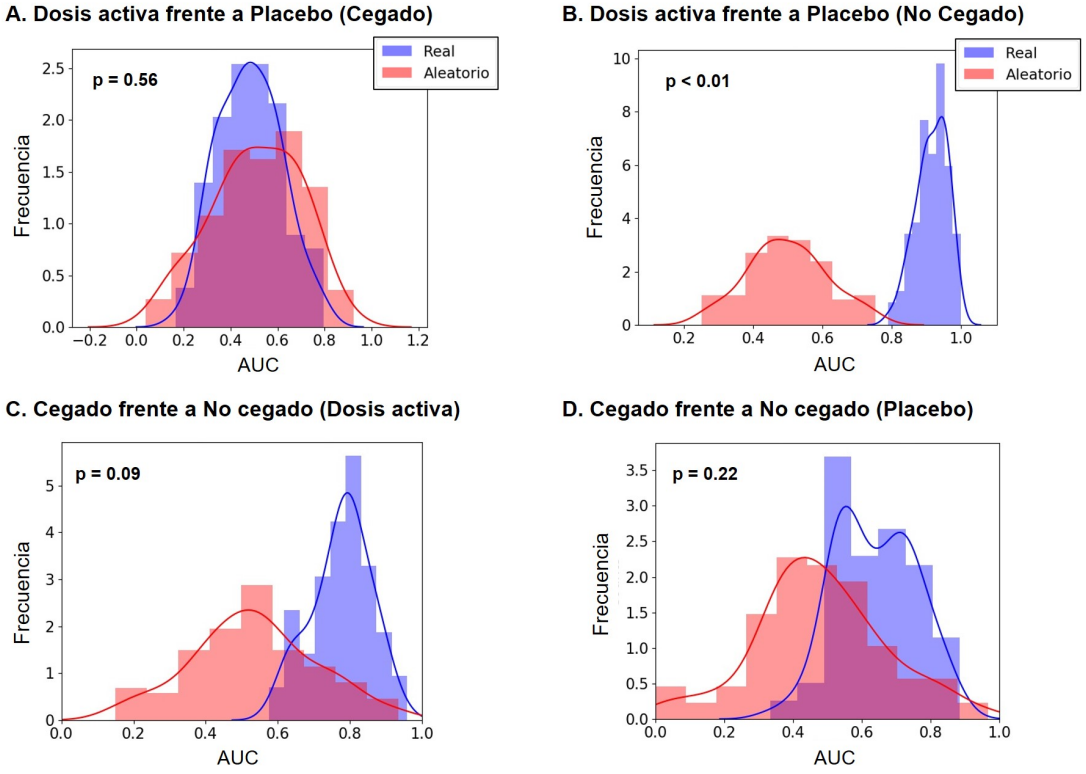

FIGURA 4.4: La clasificación de dosis activa frente a placebo restringido a participantes no cegados, fue la única en mostrar diferencias significativas ($p < 0.01$) respecto a la asignación aleatoria de etiquetas. Histogramas de los valores AUC obtenidos para la clasificación real (azul) y aleatoria (rojo) para los subgrupos: Dosis activa frente a placebo restringido a participantes (A) Cegados, (B) No cegados. Participantes cegados frente a no cegados restringido a la condición (C) Dosis activa, (D) Placebo.

4.5. Discusión

En este trabajo, se estudiaron los efecto de la microdosificación con psicodélicos en el lenguaje natural a través de tres métricas: la variabilidad semántica, la verbosidad y la

polaridad del sentimiento. Las dos últimas distinguieron significativamente determinadas respuestas dadas por los participantes bajo la dosis activa frente a la condición de placebo. Los clasificadores *Random Forest* discriminaron las condiciones experimentales con un AUC ~ 0.8 utilizando solo estas métricas como *features*. Por el contrario, no se encontraron diferencias significativas (en estas medidas) para los participantes que identificaron correctamente la condición experimental (no cegados) frente a los que no lo hicieron (cegados). Las clasificaciones mostraron concordancia con estos resultados dando valores de AUC cercanos al nivel de azar. Además, se encontraron diferencias estadísticamente significativas y una clasificación robusta para el subgrupo de dosis activa frente a la condición de placebo restringido a participantes no cegados.

En nuestra publicación anterior (Sanz et al., 2021) encontramos que dosis medias/altas de LSD aumentaban la verbosidad y reducían la coherencia del discurso. Asociamos estas características a un estado entrópico e hiper-asociativo que, a su vez, podría estimular la creatividad (Girn et al., 2020). Sin embargo, la variabilidad semántica no fue afectada por la microdosificación, cuestionando si la misma es capaz de potenciar aspectos de la cognición relacionados a la creatividad. Los gráficos de las Figura **??** mustran que los particioantes no cegados tendieron a hablar menos en la condición de placebo, principalmente en las preguntas asociadas a percepeción, estado de ánimo, creatividad y estado de alerta.

Como se mencionó previamente (ver Capítulo 1), uno de los efectos subjetivos frecuentemente reportados por usuarios de microdosis corresponde a mejoras en el estado anímico (Anderson et al., 2019a; Hutten et al., 2019a; Johnstad, 2018; Lea et al., 2020b; Polito and Stevenson, 2019; Prochazkova et al., 2018; Rootman et al., 2021). Nuestros resultados relativos al análisis de sentimiento presentan concordancia con este efecto. Más aún, el aumento del MSS no se observó solo en la respuesta asociada al "estado de ánimo" sino que resultó una característica general del lenguaje bajo los efectos de la dosis activa. Esto indica que la microdosificación es capaz de inducir un estado de ánimo positivo que se refleja en la expresión verbal.

Al etiquetar a los participantes como cegados y no cegados, encontramos que alrededor de la mitad identificó correctamente su condición experimental. Las comparaciones entre dosis activa frente a placebo restringido al grupo cegado (Figura 4.4.A), podrían indicar que las expectativas de los participantes desempeñan un papel importante en la experiencia percibida en la microdosificación. Sin embargo, si los efectos de la dosis activa fuesen un producto exclusivo de las expectativas, las comparaciones entre cegado frente a no cegado restringido a la dosis activa (Figura 4.4.C) deberían arrojar niveles de AUC significativamente diferentes que el azar. De los gráficos de la figura El problema de la condición experimental constituye una característica intrínseca del estudio de la

microdosificación psicodélica y, más en general, en la investigación de compuestos capaces de provocar alteraciones profundas en el estado de conciencia (Kuypers et al., 2019; Muthukumaraswamy et al., 2021; Schenberg, 2021; Szigeti et al., 2021; van Elk et al., 2021)

Nuestro estudio presenta limitaciones derivadas de su diseño experimental. Dado que consideramos los efectos de la microdosis durante una semana, los resultados a largo plazo asociados con efectos acumulativos no pueden ser captados por nuestros análisis. Aunque períodos experimentales de mayor duración impliquen una demanda de tiempo por parte de los participantes e investigadores, la flexibilidad del NLP mitiga estos inconvenientes ya que permite extraer datos a distancia (autograbados por parte de los participantes) y ejecutar análisis automatizados. También es posible que subyazca un sesgo en los resultados proveniente de que un participante aparezca tanto en el conjunto de entrenamiento como de testeo en algunas de las iteraciones. Si bien estimamos que las repetidas iteraciones mitigan este sesgo, no podemos descartarlo. En conjunto, estas observaciones y nuestros resultados, ponen en relieve la utilidad del NLP como herramienta capaz de monitorizar los efectos de la microdosificación con validez ecológica.

Capítulo 5

El lenguaje natural en enfermedades neurodegenerativas

Tanto investigaciones anteriores (Elvevåg et al., 2007; Mota et al., 2012; Carrillo et al., 2013; Mota et al., 2014; Bedi et al., 2014) como el conjunto de resultados presentados en los Capítulos 3 y 4, posicionan al NLP como una herramienta mínimamente invasiva, escalable, objetiva y económica que permite estudiar alteraciones en la cognición por medio del análisis del lenguaje. Debido a estas características, las técnicas de NLP comenzaron a adquirir interés en los ámbitos clínicos. En particular, muchos estudios focalizaron su atención en áreas que conllevan altos costes económicos y dependen de diagnósticos subjetivos (Nichols et al., 2019; Nakamura et al., 2015), como las enfermedades neurodegenerativas (König et al., 2015; Eyigoz et al., 2020b; Rentoumi et al., 2014) y los trastornos psicóticos (Carrillo et al., 2013; Bedi et al., 2014). Inspirados en estos proyectos, nos propusimos estudiar los efectos de las enfermedades neurodegenerativas en el lenguaje natural. Este capítulo abarca contenidos adaptados de "Automated text-level semantic markers of Alzheimer's disease" (Sanz et al., 2022b) donde estudiamos las alteraciones lingüísticas causadas por el deterioro cognitivo de la enfermedad de Alzheimer (EA).

Alrededor de 43 millones de individuos padecen la EA (ver Capítulo 1), un trastorno caracterizado por la atrofia progresiva del temporo-parieto-hipocampo y degeneraciones de la memoria semántica y episódica (Nichols et al., 2019; Ozel-Kizil et al., 2020; Scheltens et al., 2021). Por las razones previamente mencionadas, se pueden encontrar numerosas publicaciones que aplican métodos automatizados del habla en la detección de individuos con EA. Sin embargo, la mayoría carece de utilidad clínica (de la Fuente Garcia et al., 2020) debido a la falta de interpretabilidad. De hecho, múltiples estudios

basan sus resultados en dominios articulatorios (Fraser et al., 2016) o sintácticos (Orimaye et al., 2014) que no suelen considerarse como síntomas distintivos en las primeras etapas clínicas de la EA. Además, pocas investigaciones incluyen grupos de control con otras enfermedades neurodegenerativas, poniendo en duda la selectividad de los dominios analizados. Otros contratiempos frecuentes radican en la ausencia de emparejamiento demográfico, el análisis de tareas lingüísticas restringidas y la aplicación de enfoques de ML sub-óptimos (de la Fuente Garcia et al., 2020).

En este trabajo investigamos dos dominios que, según evaluaciones estándar, se encuentran sistemáticamente afectados en la EA (Bucks et al., 2000; Dijkstra et al., 2004; Giffard et al., 2001; Hodges et al., 1992): la granularidad semántica y la variabilidad semántica. En efecto, los pacientes con esta enfermedad se caracterizan por utilizar términos de menor granularidad (i.e. generales) que desencadenan en discursos con abundantes hiperónimos (e.g. "animal", "fruta") y escasos hipónimos (e.g. "gato", "arándano") (Giffard et al., 2001; Hodges et al., 1992). Por otro lado, las frecuentes interrupciones que desvían el relato de la temática principal (e.g. "¿Qué estaba diciendo?", "¿Cuál era la pregunta?", etc) provocan discontinuidades conceptuales (Dijkstra et al., 2004; Pistono et al., 2019) y cambios bruscos en el flujo del discurso. Para extraer estas características de forma automática, cuantificamos la granularidad semántica (ver Sección 2.5.1) empleando la taxonomía de WordNet (Bird et al., 2009) y medimos la variabilidad semántica entre palabras sucesivas (ver Sección 2.2) con el embedding *FastText*. Además, incluimos un grupo de pacientes con la enfermedad de Parkinson (EP), un trastorno que en etapas tempranas presenta alteraciones en conceptos principalmente asociados a la acción (Birba et al., 2017; García et al., 2018). Las comparaciones con este grupo permitieron verificar si las métricas consideradas eran marcadores selectivos de la EA.

En resumen, analizamos múltiples tareas de habla para capturar aspectos lingüísticos de pacientes con EA, EP y un grupo de control (GC). Los grupos presentaban un tamaño similar y fueron emparejados según edad, sexo y años de educación. Integramos herramientas de NLP para calcular la granularidad y variabilidad semántica, análisis estadísticos (ANOVAs) con comparaciones por pares HSD de Tukey y técnicas de clasificación basadas en algoritmos de ML (Gradient Boosting). Hipotetizamos que las métricas consideradas diferenciarán robustamente a los pacientes con EA frente al GC; pero no se encontrarán distinciones al comparar pacientes con EP frente al GC.

5.1. Participantes y protocolo

En colaboración con la Clínica de Memoria y Neuropsiquiatría de la Universidad de Chile y el Hospital del Salvador, se reclutaron 55 hispanohablantes nativos (con audición normal o corregida a normal). La muestra estaba compuesta por 21 pacientes con EA, 18 con EP y 16 sujetos sanos del GC (ver Apéndice C para la estimación de potencia). Los pacientes fueron diagnosticados por neurólogos expertos siguiendo los criterios clínicos del *National Institute of Neurological and Communicative Diseases and Stroke-Alzheimer's Disease and Related Disorders Association* para la EA, y los estándares de *United Kingdom Parkinson's Disease Society Brain Bank* para la EP (Hughes et al., 1992). En línea con investigaciones anteriores (García et al., 2018; Eyigoz et al., 2020a; García et al., 2016; Santamaría-García et al., 2017), los diagnósticos fueron respaldados por extensos exámenes neurológicos, neuropsicológicos y de neuroimagen. Además, ningún paciente presentó historial de otros desórdenes neurológicos, condiciones psiquiátricas, déficits primarios del lenguaje o abuso de sustancias.

Las puntuaciones medias de la Evaluación Cognitiva de Montreal (MoCA) se encontraban por debajo del umbral de demencia para el grupo con EA y de deterioro cognitivo leve para el grupo con EP (Delgado et al., 2019). Siguiendo la batería de pruebas *INECO Frontal Screening* (IFS) (Torralva et al., 2010), se determinó que los pacientes con EA presentaban disfunción ejecutiva. Los pacientes con EP no exhibían síntomas de Parkinson-plus y fueron evaluados en la fase activa de la medicación. Los participantes en el GC no contaban con antecedentes de enfermedades neuropsiquiátricas o abuso de sustancias, eran funcionalmente autónomos y obtuvieron puntuaciones estándar en el MoCA e IFS. Para descartar influencias demográficas, se emparejaron los grupos según sexo, edad y educación (ver Tabla 5.1).

Los participantes realizaron siete tareas de lenguaje, cuatro de ellas correspondían a narraciones espontáneas de experiencias personales. Específicamente, se pedían relatos de (1) su rutina diaria, (2) sus principales intereses, (3) un recuerdo agradable y (4) un recuerdo desagradable. Las restantes, correspondían a tareas de lenguaje semi-espontáneas que implicaban describir (5) la imagen *Cookie Theft* (CTP) del Examen Diagnóstico de Afasia de Boston, (6) una imagen de una familia trabajando en una cocina y (7) una película muda animada de un minuto de duración (inmediatamente después de su finalización). Las narraciones de los participantes se grabaron en salas silenciosas con micrófonos anti-ruido y se almacenaron en archivos ".wav" (44100 Hz, 16 bits) con la aplicación *Cool Edit Pro 2.0*.

Todos los participantes dieron su consentimiento informado por escrito de acuerdo con la Declaración de Helsinki. El estudio fue aprobado por el comité institucional de ética

	EA ($n = 21$)	EP ($n = 18$)	GC ($n = 16$)	Estadística (Grupos)	Comparaciones por pares		
					Grupos	MSE	p-valor
Datos demográficos							
Sexo (F:M)	13:8	10:8	13:3	$\chi^2 = 4.86$ $p = 0.1^a$	-----	-----	-----
Edad	77.24 (6.47)	76.50 (6.40)	75.94 (4.35)	$F = 0.21$ $p = 0.81^b$	-----	-----	-----
Años de educación	11.24 (3.78)	9.39 (5.11)	12.94 (4.28)	$F = 2.62$ $p = 0.08^b$	-----	-----	-----
Datos neuropsicológicos							
MoCA	13.90 (4.34)	20.33 (4.68)	25.07 (3.43)	F = 29.01 $p < 0.001^b$	EA vs GC EP vs GC EA vs EP	12.75 29.39 23.27	$< 0.001^c$ 0.006^c $< 0.001^c$
Batería IFS	11.07 (4.48)	17.08 (4.86)	18.90 (4.26)	$F = 14.30$ $p < 0.001^b$	EA vs GC EP vs GC EA vs EP	13.85 57.72 18.98	$< 0.001^c$ 0.51^c $< 0.001^c$

FIGURA 5.1: Los datos se presentan como media (desvío estándar), a excepción del sexo. **a.** P-valores calculados mediante la prueba chi-cuadrado (χ^2). **b.** P-valores calculados mediante ANOVAs de medidas independientes. **c.** P-valores calculados mediante pruebas post hoc HSD de Tukey. EA: Enfermedad de Alzheimer; EP: Enfermedad de Parkinson; GC: Grupo de Control; MoCA: Evaluación Cognitiva de Montreal; IFS: batería de pruebas INECO *Frontal Screening*.

del Centro de Memoria y Neuropsiquiatría (CMYN) del Hospital del Salvador (Departamento de Neurología) y la Facultad de Medicina de la Universidad de Chile.

5.2. Métodos

Las grabaciones se transcribieron utilizando el algoritmo de reconocimiento de voz de *Google Cloud Platform* junto con la API *Speech-to-Text* (`https://cloud.google.com/speech-to-text/docs`) incluida en la librería $speech_{v1}$ de Python. El código utilizado para la conversión de audio a texto puede encontrarse en: `https://github.com/ToyokoLabs/cloudscripts/blob/master/cgp/m4atotext.ipynb`. Las trascripciones fueron revisadas y corregidas manualmente por miembros del equipo de investigación.

5.2.1. Pre-procesamiento

Siguiendo los pasos del estudio anterior (Capítulo 4), las transcripciones se pre-procesaron con la librería TreeTagger de Python entrenado con la base de AnCora. Para optimizar la potencia estadística y abarcar múltiples escenarios lingüísticos, se agruparon las siete transcripciones de cada participante en un único documento. Las distribuciones de verbosidad de palabras lemmatizadas, no mostraron diferencias significativas entre grupos

($F[2, 52] = 0.64, p = 0.53, \eta_p^2 = 0.02$). Las medias y desvíos de verbosidad fueron de: $M = 1051$, $SD = 112$ (EA); $M = 1193$, $SD = 140$ (EP) y $M = 1239$, $SD = 124$ (GC).

5.2.2. Granularidad semántica

Siguiendo los pasos descriptos en la Sección 2.5.1, se computó la granularidad semántica de las transcripciones utilizando la taxonomía de sustantivos provista por la base de datos léxica *WordNet*. Aplicando la herramienta PoS *tagging* de la librería TreeTagger, se identificaron los sustantivos contenidos en las transcripciones y los errores de etiqueta fueron solventados manualmente. Las traducciones (al inglés) de los términos identificados se efectuaron con el diccionario de *WordNet*.

La granularidad de cada sustantivo vino dada por la cantidad de nodos contenidos en el camino más corto entre el mismo y la palabra "entidad" (ver Figura 2.3.C). Considerando el conjunto completo de transcripciones, alrededor de 5.68 % de sustantivos no estaban incluidos en el corpus de *WordNet*, tras verificar que la cantidad excluida no presentaba diferencias significativas entre grupos ($p = 0.93$), se optó por descartarlos. De esta forma, cada participante se identificó con 12 *features*, donde la *feature* n representa la cantidad de sustantivos con valor de granularidad n que el participante pronunció (incluyendo repeticiones) en el conjunto de tareas consideradas, normalizado por la cantidad de sustantivos totales. El valor de granularidad 1 se excluyó dado que en ninguna transcripción figuraba la palabra "entidad". Dentro de la *feature* 12, se incluyó un grupo de sustantivos ($\sim 0.18\,\%$) que presentaban un valor de granularidad mayor a 12.

5.2.3. Variabilidad semántica

Replicando los estudios anteriores, se computó la variabilidad semántica para determinar si esta métrica podía utilizarse en la distinción de enfermedades neurodegenerativas. Para evitar sesgos provenientes de disfluencias, vacilaciones o estrategias de búsqueda de palabras, se excluyeron los términos repetidos consecutivos (i.e. un texto compuesto por una sola palabra repetida presentaría una variabilidad nula). El modelo *FastText* pre-entrenado con la base de datos de Common Crawl y Wikipedia en español (ver Sección 4.3.2) se utilizó para obtener la representación vectorial de las transcripciones y calcular su variabilidad semántica (Sección 2.2.1).

5.2.4. Análisis estadísticos

Los análisis estadísticos se realizaron con la biblioteca *Pingouin* de Python (https://pingouin-stats.org/), aplicando pruebas ANOVA unidireccionales para las comparaciones entre todos los grupos y pruebas HSD de Tukey para las comparaciones por pares. Los niveles de significación estadística se fijaron en $\alpha = 0.05$. Los tamaños de efecto se estimaron con el cuadrado eta parcial (η_p^2) para los resultados del ANOVA y con la d de Cohen para las comparaciones por pares. Cada una de las 12 métricas (11 valores de granularidad y una de la variabilidad semántica) se consideraron como variables independientes.

5.2.5. Clasificadores

Para determinar si las métricas evaluadas eran marcadores distintivos y selectivos de la EA, se implementaron dos clasificadores *Gradient Boosting* con la librería scikit-learn (https://scikit-learn.org/). El primero se entrenó para diferenciar pacientes con la EA y el GC, el segundo para distinguir pacientes con la EP y el GC. Los modelos contaban de 5000 árboles de decisión entrenados con subgrupos de 3 *features*, una tasa de aprendizaje de 0.01 y validación cruzada de 3 subconjuntos. Aplicando el procedimiento descripto en la Sección 2.6, se determinó el rendimiento de los clasificadores computando el AUC y el p-valor para 1000 iteraciones. Además, en cada repetición se aplicó un selector de *features* univariado (en los subconjuntos de entrenamiento) para identificar las cinco *features* principales en función de su valor F ANOVA entre grupos. La importancia de cada *feature* se calculó contando la cantidad de veces que la misma fue seleccionada por su valor F ANOVA, dividido por el número de subconjuntos y multiplicado por el número de iteraciones (ver Apéndice C).

5.3. Resultados

5.3.1. Resultados estadísticos

Las comparaciones entre grupos mostraron que los pacientes con EA obtuvieron puntuaciones menores para valores de granularidad elevados (8 a 12, exceptuando el 10), indicando un escaso uso de hipónimos en su discurso (Figura 5.2.A). Los análisis estadísticos revelaron diferencias significativas entre grupos para granularidades de 5 ($F[2, 52] = 5.43$, $p = 0.007$, $\eta_p^2 = 0.17$) y 11 ($F[2, 52] = 4.71$, $p = 0.013$, $\eta_p^2 = 0.15$). Comparando por pares se encontró que los pacientes con EA presentaban aumentos significativos para el valor de granularidad 5 ($p = 0.072$, $d = 0.73$) y disminuciones para el valor

11 ($p = 0.008$, $d = 1$) respecto al GC. También se obtuvieron puntuaciones significativamente superiores en los pacientes con EP frente al GC en el valor de granularidad 5 ($p = 0.003$, $d = 1.11$). Para el resto de las comparaciones por pares se encontraron p-valores mayores a 0.05.

Los análisis de la variabilidad semántica (Figura 5.2.B), mostraron un efecto de grupo significativo ($F[2, 52] = 4.24$, $p = 0.02$, $\eta_p^2 = 0.14$). Las comparaciones por pares revelaron puntuaciones más altas para los pacientes con EA frente al GC ($p = 0.011$, $d = 0.97$). El resto de las pruebas HSD de Tukey resultaron en p-valores mayores a 0.05 (EP frente a GC: $p > 0.2$; EA frente a EP: $p > 0.4$).

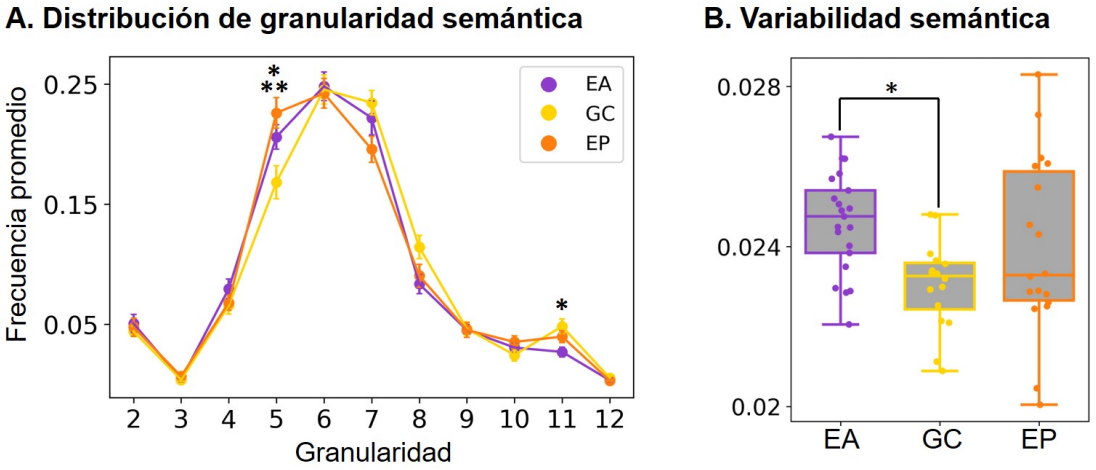

FIGURA 5.2: Distribuciones de granularidad y variabilidad semántica en el conjunto de tareas lingüísticas. **A.** Frecuencia de los valores de granularidad semántica promediada por grupo. Respecto al GC, los pacientes con EA mostraron aumentos en valores de granularidad más bajos y disminuciones en valores más altos. **B.** Representación de la variabilidad semántica con diagramas de caja, los pacientes con EA mostraron aumentos respecto al GC. Las diferencias significativas entre pares ($p < 0.05$), se indican con * (EA frente a GC) y ** (EP frente a GC).

5.3.2. Resultados de la Clasificación

La clasificación entre pacientes con EA y el GC (Figura 5.3.A) alcanzó un valor AUC de 0.80 ± 0.06 (exactitud: 0.71 ± 0.11; sensibilidad: 0.80 ± 0.15; precisión: 0.73 ± 0.12). Este valor fue significativamente mayor ($p = 0.022$) que el obtenido para la asignación aleatoria de clases, donde todas las medidas de rendimiento presentaron cercanías con el nivel de azar (AUC: 0.49 ± 0.13; exactitud: 0.52 ± 0.16; sensibilidad: 0.64 ± 0.21; precisión: 0.57 ± 0.18). Por el contrario, la clasificación entre pacientes con EP y el GC (Figura 5.3.B) resultó en un valor AUC de 0.65 ± 0.08 (exactitud: 0.60 ± 0.12; sensibilidad: 0.61 ± 0.20; precisión: 0.64 ± 0.15) que no difirió significativamente ($p = 0.16$) del obtenido

con la asignación aleatoria de clases (AUC: 0.50±0.13; exactitud: 0.50±0.15; sensibilidad: 0.56 ± 0.23; precisión: 0.53 ± 0.20).

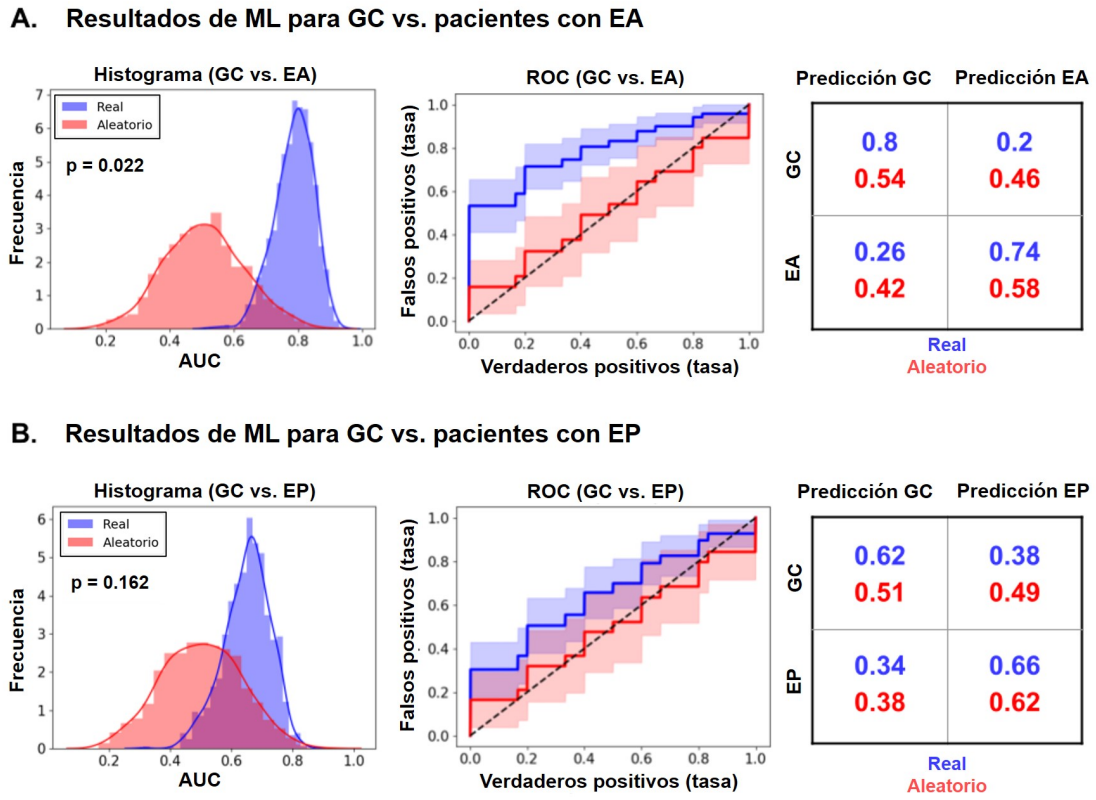

FIGURA 5.3: Clasificación entre el grupo de control frente a (**A**) pacientes con EA y (**B**) pacientes con EP; utilizando las métricas de granularidad y variabilidad semántica en el conjunto de tareas lingüísticas. Los paneles muestran los histogramas de AUC normalizados (izquierda), curvas ROC promedio (centro) y matrices de confusión normalizadas por fila y promediadas sobre las iteraciones (derecha). Los resultados reales se representan en azul y los obtenidos con la asignación aleatoria de clases en rojo.

5.4. Discusión

En este trabajo, examinamos dos posibles marcadores distintivos de la EA: la granularidad y variabilidad semántica. Ambas métricas diferenciaron los pacientes con EA del GC, según pruebas ANOVAs y HSD de Tukey. En concordancia, los clasificadores de ML alcanzaron predicciones robustas a nivel sujeto. Estas diferencias no se observaron en las comparaciones de pacientes con EP frente al GC.

Respecto al GC, los pacientes con EA presentaron un reducción de la granularidad semántica en las tareas lingüísticas. Investigaciones previas (Giffard et al., 2001; Ozel-Kizil et al., 2020), mostraron con medidas estandarizadas (número de repuestas correctas y tiempo de respuestas) que este fenómeno se encuentra presente en tareas de lenguaje controladas (e.g. fluidez de categorías, paradigmas de *priming*, etc). Nuestro estudio

sugiere que el habla natural de los pacientes con EA también se caracteriza por una dependencia de hiperónimos, posicionando a la granularidad como un marcador de enfermedades con alteraciones primarias de la memoria semántica (Patterson et al., 2007). De hecho, este fenómeno también se ha observado en pacientes con demencia semántica (Patterson et al., 2007), un trastorno que comparte las regiones centrales de atrofia con la EA (e.g. hipocampo y lóbulos temporales) (Chapleau et al., 2016). En consecuencia, aunque varios niveles de granularidad mostraron solapamiento entre los pacientes con EA y el GC, esta métrica contiene el potencial de capturar alteraciones sutiles pero informativas.

El análisis de la variabilidad semántica reveló que los pacientes con EA presentaban un discurso discontinuo. La reducción de coherencia y cohesión en esta enfermedad, ya fue estudiada con métodos frecuentistas basados en ocurrencias de frases desligadas de frases adyacentes o de la temática principal (Dijkstra et al., 2004; Pistono et al., 2019; Ash et al., 2007). Nuestro estudio muestra que las relaciones semánticas disímiles también se manifiestan palabra a palabra. La inspección manual de las trascripciones, reveló que el discurso de los pacientes con EA contenía numerosas interrupciones (e.g. “No sé”, “Me olvidé el nombre”, “No me acuerdo”, entre otras) que concuerdan con el excesivo uso de lenguaje “formulado” (palabras semánticamente vacías) (Zimmerer et al., 2016; Yuan et al., 2020) observado en este grupo.

Los clasificadores de ML corroboraron la robustez de ambas medidas. Utilizando como entrada las métricas de granularidad y variabilidad semántica, se encontró un valor AUC de 0.8 que permitió identificar correctamente el 80 % de los participantes en el GC y el 74 % de los pacientes con EA. Estos resultados superan el rendimiento alcanzado por estudios previos de NLP basados en dominios que no se encuentran notablemente afectados en la EA, como la articulación o la sintaxis (Ahmed et al., 2013; Kaprinis and Stavrakaki, 2007; Weiner et al., 2008). Además, las clasificaciones revelaron diferencias significativas respecto al nivel de azar (i.e. asignación aleatoria de categorías). Estos resultados muestran que la granularidad y variabilidad semántica pueden contribuir en la diferenciación clínica de pacientes con EA respecto a sujetos sanos.

Exceptuando un valor de granularidad, los resultados anteriores fueron selectivas de la EA. En efecto, la clasificación de pacientes con EP frente al GC, presentó un rendimiento cercano al nivel de azar y estadísticamente indistinguible de la asignación aleatoria de clases. Es importante destacar que otros dominios verbales frecuentemente evaluados en la EA, como la fluidez semántica y fonémica (Laws et al., 2010), también suelen verse comprometidos en la EP (Henry and Crawford, 2004), lo cual limita su uso para la diferenciación de enfermedades. Dado que las alteraciones semánticas en la EP son sensibles al estado de medicación (Norel et al., 2020), futuros estudios deberían explorar

las *features* analizadas para niveles variables de dopamina y corroborar o refutar si las medidas consideradas se mantienen selectivas de la EA.

Un enfoque frecuentemente utilizado en investigaciones del lenguaje consiste en extraer extensos grupos de métricas del habla y utilizar selectores de importancia de *features* junto con clasificadores de ML. Si bien este enfoque suele derivar en un rendimiento elevado, carece de interpretabilidad neuropsicológica al no poder alinearse con el conocimiento clínico (de la Fuente Garcia et al., 2020). De hecho, determinados rasgos fonológicos y sintácticos postulados como marcadores distintivos de la EA (Fraser et al., 2016; Orimaye et al., 2014), contradicen la evidencia neuropsicológica (Ahmed et al., 2013; Kaprinis and Stavrakaki, 2007; Weiner et al., 2008). Nosotros adoptamos el procedimiento inverso: primero identificamos los dominios lingüísticos que se ven afectados de forma sistemática por la enfermedad (según la literatura neuropsicológica) y luego, desarrollamos técnicas de NLP para cuantificarlos. De esta forma, nuestro estudio encamina la aplicación clínica de técnicas automatizadas del lenguaje como complemento de las evaluaciones habituales de la EA (Laske et al., 2015). Mientras que las pruebas neuropsicológicas estándar pueden resultar costosas, dependiendentes del examinador (Orimaye et al., 2018) e ignorar el comportamiento espontáneo (García et al., 2018); las técnicas del NLP conllevan costes mínimos y resulta en datos naturalistas objetivos.

Es importante destacar que nuestro trabajo respeta el equilibrio demográfico entre las muestras. Al analizar el lenguaje, es fundamental considerar factores como el nivel educativo que podría desembocar en un léxico de mayor complejidad. Además, utilizamos una serie de tareas espontáneas (autobiográficas) y semiesponténeas (basadas en estímulos) (Boschi et al., 2017), que permiten englobar diversos aspectos lingüísticos y descartar resultados provenientes de muestras breves de discurso. Sin embargo, aunque el número de muestra consideradas respeta los valores típicos del campo (Dijkstra et al., 2004), futuros estudios deberían incrementar el tamaño de los participantes. También resaltamos que este trabajo se centró exclusivamente en hispanohablantes, dado que la misma enfermedad puede desencadenar distintos patrones según el idioma (Canu et al., 2020), estudios multi-lingüísticos son necesarios para fundar enfoques globales.

En resumen, nuestro estudio muestra que las herramientas de NLP pueden utilizarse para cuantificar marcadores diferenciables e interpretables de la EA. Los pacientes con esta enfermedad parecen estar caracterizados por una granularidad reducida y un incremento de la variabilidad semántica, ambos patrones se encontraron ausentes en pacientes con la EP. Nuestros resultados impulsan la aplicación clínica de técnicas lingüísticas automatizadas en la identificación de marcadores de la demencia.

Capítulo 6

BERT y el lenguaje natural en la enfermedad de Alzheimer

El algoritmo BERT y sus variantes adquirieron una creciente popularidad en la resolución de (casi) cualquier problema de NLP (Kenton and Toutanova, 0189). Al igual que la mayoría de las redes neuronales profundas, el principal problema de estos algoritmos radica en la interpretabilidad. Es por ello que múltiples estudios enfocaron su atención en técnicas que permitan explicar (parcialmente) el funcionamiento de esta red, por ejemplo analizar los mecanismos de atención en las capas (Clark et al., 2019; Kovaleva et al., 2019) o buscar interpretaciones locales de los resultados (Ribeiro et al., 2016; Lundberg and Lee, 2017).

Diversas investigaciones comenzaron a aplicar algoritmos del tipo BERT para detectar enfermedades neurodegenerativas y regresionar pruebas neuropsicológicas numéricas con un alto rendimiento (Balagopalan et al., 2019; Yuan et al., 2020; Ilias and Askounis, 2022). Motivados por estos resultados, extrajimos dos bases de datos de la plataforma *DementiaBank* (`https://dementia.talkbank.org`) que contiene información multimedia (audios y transcripciones) de numerosos estudios de demencia. En particular, descargamos las transcripciones de la tarea lingüística semi-espontánea *Cookie Theft* (CTP) para pacientes con EA y un GC. Clasificamos los datos con el algoritmo BERT y aplicamos el modelo *Local Interpretable Model-agnostic Explanations* (LIME) (Ribeiro et al., 2016) para detectar las *features* de mayor importancia. Además, comparamos los resultados con un algoritmo de ML interpretable, la regresión logística.

6.1. Base de datos

Dos bases de datos se extrajeron de la plataforma *DementiaBank*, la primera correspondía a un estudio longitudinal (5 años) del programa *Alzheimer Research* (ARP) de la Universidad de Pittsburgh. El corpus contenía tanto pruebas neuropsicológicas como tareas lingüísticas (e.g. descripción de imagen, fluidez, recuerdo inmediato, etc) provenientes de individuos sanos (GC) y pacientes que presentaban algún tipo de demencia. La mayoría de las tareas consideradas por el programa eran estructuradas o de corta duración, exceptuando la descripción de la imagen *Cookie Theft* (CTP). La explicación detallada de la inclusión/exclusión de participantes, exámenes neurológicos y métodos experimentales se puede encontrar en (Becker et al., 1994). Inicialmente, el ARP reclutó 244 participantes, sin embargo, muchos cesaron su colaboración en el estudio. Dado este factor, las transcripciones disponibles en *DementiaBank*, la naturaleza de las tareas lingüísticas y el bajo porcentaje de datos ($\sim 16\,\%$) correspondientes a enfermedades que no fuesen Alzheimer (i.e. Demencia Vascular, Parkinson, Deterioro Cognitivo Leve, etc) se optó por considerar exclusivamente las muestras recolectadas en el primer año de estudio para la tarea CTP, provenientes del GC y pacientes con EA.

En total, se seleccionaron las transcripciones de la tarea CTP para 97 pacientes con EA y 58 participantes en el GC. Además, se extrajeron los datos demográficos disponibles (sexo, edad y años de educación) y las puntuaciones del Mini Examen del Estado Mental (MMSE). Para no descartar muestras por emparejamientos demográficos, se identificaron 39 participantes del estudio longitudinal de Wisconsin (WLS) que no presentaban historial clínico de enfermedades o trastornos degenerativos. El WLS se llevó a cabo durante un período de (aproximadamente) 60 años e incluía individuos graduados de secundarios ubicados en Winsconsin. Los participantes completaron múltiples evaluaciones cada ~ 10 años, en nuestro trabajo solo analizamos los datos correspondientes a la descripción de imagen CTP. La explicación detallada del estudio se puede encontrar en (Herd et al., 2014). De esta forma, se recolectaron las transcripciones de 194 participantes (97 con EA y 97 del GC), en la Figura 6.1 se muestran las comparaciones estadísticas asociadas a los datos demográficos y el MMSE del estudio neuropsicológico (efectuado solo para los participantes del ARP).

6.2. Métodos

Al igual que en los estudios anteriores, se descartaron las preguntas en intervenciones del entrevistador y solo se analizó el discurso de los participantes. Las distribuciones de

	EA (*n* = 97)	GC (*n* = 97)	Comparaciones por pares
Datos demográficos			
Sexo (F:M)	66:31	66:31	-----
Edad	70.4 (8.8)	70.5 (5.3)	$d = 0.16$ $p = 0.91$
Años de educación	12.91 (2.08)	12.90 (2.34)	$d = 0.0093$ $p = 0.94$
Datos neurológicos			
MMSE	19.39 (4.47)	29.02 (1.09)	$d = 2.66$ $p < 0.001$

FIGURA 6.1: Los datos se presentan como media (desvío estándar), a excepción del sexo. Todos los p-valores fueron calculados mediante pruebas t-test. EA: Enfermedad de Alzheimer; GC: Grupo Control; MMSE: Mini Examen del Estado Mental.

verbosidad, no mostraron diferencias significativas entre grupos ($T = -0.43$, $p = 0.67$). Las medias y desvíos fueron de: $M = 121$, $SD = 56$ (EA) y $M = 125$, $SD = 60$ (GC).

6.2.1. BERT y LIME

La clasificación entre los pacientes con EA frente al GC, se realizó con el algoritmo BERT (base) pre-entrenado para el inglés (https://huggingface.co/bert-base-uncased) que contiene un alrededor de 110 millones de parámetros (12 capas con 768 neuronas y 12 mecanismos de atención). Al final de la red, se agregó una capa lineal de dos neuronas completamente conectada con la posterior. Siguiendo las recomendaciones de la literatura (Kenton and Toutanova, 0189), se empleó el optimizador $AdamW$ con una tasa de aprendizaje de $5x10^{-5}$, un *batch* de 16 muestras y 8 épocas; para el resto de los hiperparámetros se conservaron los valores por defecto. Las probabilidades de clasificación se calcularon con la normalización *Softmax*: $p(z_1) = \frac{e^{z_1}}{\sum_j^2 e^{z_i}}$; donde z_i corresponde al resultado de la neurona i-ésima de la capa lineal.

El pre-procesamiento se realizó con el algoritmo *BertTokenizer*, entrenado en el mismo corpus que BERT. En cada transcripción, se aplicó el algoritmo para identificar los *tokens* (palabras o sub-palabras), la posición de los mismos y la frase a la que correspondían. Ninguna de las muestras excedió el límite de *tokens* permitidos por BERT (512). El modelo se entrenó aplicando el procedimiento descripto en la Sección 2.6 con validación cruzada estratificada de 5 subconjuntos y 100 iteraciones, para calcular tanto el AUC como el p-valor. Los *batches* se delimitaron de forma que contuvieran igual proporción de clases (50 % EA y 50 % GC).

Todos los análisis se efectuaron con la librería *PyTorch* (https://pytorch.org/) y *Transformers* (https://pypi.org/project/transformers/) de Python. Los modelos se entrenaron en GPU para reducir el tiempo computacional.

Los resultado de la clasificación se analizaron con LIME para construir una explicación local e interpretable (Ribeiro et al., 2016). Este algoritmo recibe como entrada una muestra de interés, crea N copias ligeramente modificadas y clasifica este corpus de $N+1$ datos utilizando la frontera de decisión del modelo a explicar. Las perturbaciones consisten en eliminar una o más *features* de la muestra original (e.g. píxeles de una imagen, palabras de un texto, segundo en un audio, etc). Además, se le asigna un peso a cada copia perturbada según la distancia (coseno) que presenten con la muestra (i.e. menor distancia implica mayor peso). Luego, el algoritmo entrena un modelo de ML interpretable (e.g. regresión logística, árbol de decisión, entre otros) priorizando la clasificación correcta de las copias más cercanas a la muestra. De esta forma, se atribuye la importancia de cada *feature* según el modelo interpretable como, por ejemplo, los pesos de los coeficientes en una regresión lineal o el índice de Gini en un árbol de decisión. La intuición detrás de LIME consiste en que las fronteras de decisión complejas puede descomponerse, localmente, por una sencilla (e.g. frontera lineal). En la Figura 6.2 se muestra una representación esquemática de algoritmo.

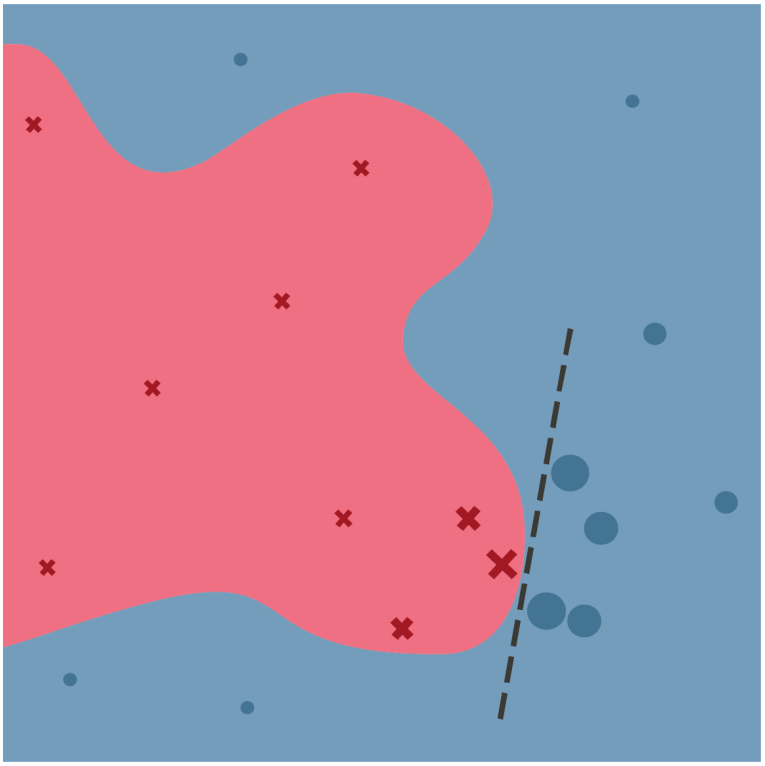

FIGURA 6.2: Representación de LIME donde se esquematizan: la muestra de interés (cruz roja de mayor tamaño), las N copias (clasificadas según la frontera de decisión del modelo complejo y con tamaño inversamente proporcional a la distancia de la muestra) y la aproximación del modelo interpretable (línea punteada).

Utilizando la librería ELI5 (*https://eli5.readthedocs.io*) de Python, se aplicó el algoritmo LIME con el modelo de regresión logística (Sección 6.2.2) regularizado, 3000 copias por muestra y perturbaciones de una (como mínimo) o más palabras individuales. Se seleccionaron aleatoriamente 5 iteraciones de las 100 realizadas, cada una contenía los parámetros de 5 modelos diferentes obtenidos de la validación cruzada. Para cada conjunto de parámetros, se aplicó LIME sobre los datos de testeo correspondientes. Dado que las copias perturbadas que genera LIME se realizan eliminando palabras y se clasifican como positivas o negativas, ELI5 construye con los coeficientes de la regresión lineal cuáles contribuyeron a que la copia se clasifique como positiva y cuáles contribuyeron a la clasificación negativa. De esta forma se calcularon los pesos de las palabras individuales y se realizaron los promedios de los pesos a nivel grupo para determinar determinar la importancia de los términos en la clasificación.

6.2.2. Regresión logística

Las clasificaciones obtenidas con BERT y los resultados de LIME se compararon con un modelo de regresión logística. Este algoritmo determina la probabilidad de que una instancia pertenezca a una categoría según la función sigmoidea (ecuación 6.1).

$$\begin{aligned} y = P(y|\vec{X},\vec{\beta}) &= \frac{1}{1+\exp^{-\vec{\beta}\vec{X}}} \\ P(y|\vec{X},\vec{\beta}) \geq p_c &\Rightarrow y = 1 \\ P(y|\vec{X},\vec{\beta}) < p_c &\Rightarrow y = 0 \end{aligned} \tag{6.1}$$

Donde $\vec{X}$ representa el conjunto de *features*, $\vec{\beta}$ corresponde a los parámetros del modelo y p_c denota la probabilidad de corte (0.5 para una clasificación binaria balanceada). Por tanto, la importancia de la *feature* i-ésima (X_i) puede determinarse con el valor del coeficiente asociado (β_i).

Para aplicar este algoritmo se construyó una matriz donde cada fila correspondía a un participante ($n = 194$) y las columnas representaban términos únicos del conjunto de todas las transcripciones ($m = 1131$). La celda ij de la matriz contenía el número de veces que el participante i-ésimo había empleado la palabra j. Además, se agregó una última columna a la matriz que contenía la etiqueta (CTR/EA) de cada participante. La librería *Sklearn* de Python se utilizó para estandarizar las frecuencias y aplicar el modelo, el entrenamiento siguió los pasos explicados en la Sección 2.6 con una validación cruzada de 5 subgrupos y 100 iteraciones para determinar tanto el AUC como el p-valor. A diferencia de los casos anteriores, en cada validación cruzada, se optimizó el parámetro

de regularización de la regresión logística. La importancia de las *features* se determinó promediando su coeficiente (β) asociado sobre las 100 iteraciones.

6.3. Resultados

La clasificación entre pacientes con EA frente al GC según el modelo BERT, resultó en un AUC medio de 0.87 ± 0.02 que fue significativamente mayor ($p < 0.01$) al obtenido con la asignación aleatoria de etiquetas (AUC $= 0.49 \pm 0.04$). En la Figura 6.3.A se presentan los histogramas de los AUCs obtenidos en cada iteración. El análisis de los términos más relevantes en la clasificación mostraron que los pacientes con EA se caracterizaban por el uso de onomatopeyas "Yeah", "Oh", "Uh", "Um", "Sh" y términos generales como "He" ("Él"), "Here" ("Aquí"), "Then" ("Entonces"), "Thing" ("Cosa"), entre otros. Por el contrario, los participantes en el GC utilizaron un léxico de mayor granularidad con términos descriptivos de la imagen *CTP* como "Action" ("Acción"), "Reaching" ("Alcanzando"), "Overflowing" ("Desbordando"), "Mother" ("Madre"), entre otras. La Figura 6.3.B esquematiza con nubes de palabras los términos de mayor relevancia, donde solo se incluyeron las palabras utilizadas en el 20 % de las transcripciones (considerando cada grupo por separado).

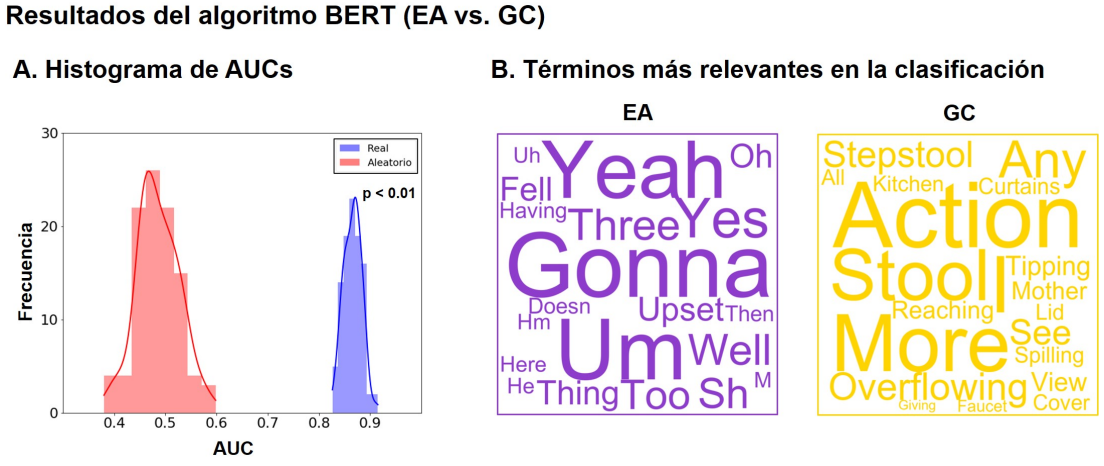

FIGURA 6.3: Clasificación entre los pacientes con EA y los participantes en el GC según el modelo de BERT. **A.** Histogramas de AUCs obtenidos según las etiquetas reales (azul) y la asignación aleatoria (rojo). **B.** Nubes de palabras de los términos más relevantes en la clasificación. El tamaño de cada palabra corresponde al valor promedio obtenido con el modelo LIME.

De forma análoga, la clasificación efectuada con el modelo de regresión logística resultó en un AUC de 0.82 ± 0.02 y 0.49 ± 0.05 ($p < 0.01$) para la clasificación con las etiquetas reales y aleatorias, respectivamente. En la Figura 6.4.A se presentan los histogramas de los AUCs obtenidos en cada iteración. El análisis de los términos más relevantes en

la clasificación mostró concordancia con los resultados de LIME. En la Figura 6.4.B se esquematizan las nubes de palabras donde se seleccionaron los 20 términos con coeficientes β de mayor valor (positivo) para representar a los pacientes con EA y menor valor (negativo) para los participantes en el GC.

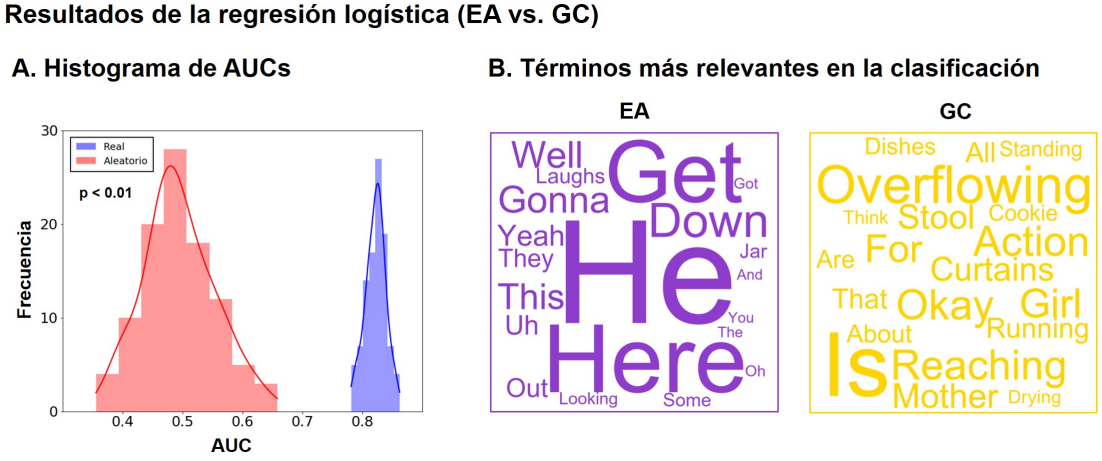

FIGURA 6.4: Clasificación entre los pacientes con EA y los participantes en el GC según el modelo de regresión logística. **A.** Histogramas de AUCs obtenidos según las etiquetas reales (azul) y la asignación aleatoria (rojo). **B.** Nubes de palabras de los términos más relevantes en la clasificación. El tamaño de cada palabra corresponde al valor absoluto promedio de su coeficiente β asociado.

6.4. Discusión

En este trabajo analizamos las transcripciones de la tarea lingüística descriptiva *CTP* para distinguir pacientes con EA frente a un GC. Dado que los datos se recolectaron de dos fuentes diferentes, no descartamos que subsista algún desparejamiento oculto. Dado que este análisis se basó en comparar dos métodos, cualquier posible sesgo presente debería manifestarse en ambos sin afectar significativamente los resultados. La clasificación se realizó con dos modelos de ML de diferente naturaleza: (1) BERT junto con las interpretaciones locales de LIME y (2) la regresión logística. A diferencia de los estudios anteriores donde se extrajeron o construyeron métricas del discurso que reflejaban aspectos particulares de los grupos a comparar (Sanz et al., 2021, 2022b,a), este trabajo se basó en la inspección de las transcripciones completas. Ambos modelos resultaron en clasificaciones significativamente distinguibles del azar dando valores de AUC medio similares ($> 0{,}8$) y comparables con investigaciones que analizaron bases de datos semejantes (Balagopalan et al., 2019; Ilias and Askounis, 2022).

La extracción de los términos de mayor relevancia para la distinción de clases mostró que los pacientes con EA se caracterizaban por emplear onomatopeyas. Estos resultados

reafirman la utilización de lenguaje formulado (Yuan et al., 2020; Zimmerer et al., 2016; Sanz et al., 2022b) previamente observado. En contraparte, los participantes en el GC se caracterizaron por utilizar verbos y sustantivos descriptivos de la imagen. Es importante destacar que el algoritmo LIME proporciona una interpretación local de los resultados que puede carecer de sentido global. En particular, cabe la posibilidad de que el modelo BERT no base sus resultados en la identificación de términos generales y onomatopeyas para lograr una clasificación de alto rendimiento.

La similitud presentada entre BERT y la regresión logística llevan a preguntarse cuándo es necesario aplicar un modelo complejo y poco interpretable. Basados en publicaciones previas (Yuan et al., 2020; Balagopalan et al., 2019; Ilias and Askounis, 2022), inferimos que la clasificación binaria de estas transcripciones no presenta la dificultad necesaria para la aplicación de un modelo complejo. Futuros estudios podrían verificar esta hipótesis con el agregado de otro grupo de pacientes con enfermedades neurodegenerativas, es decir, aumentar la complejidad de la clasificación agregando una o más clases.

Capítulo 7

Conclusiones

En el transcurso de esta tesis, el NLP nos permitió identificar características asociadas a estados alterados de conciencia producidos tanto por sustancias psicodélicas como por enfermedades neurodegenerativas. Aplicando análisis estadísticos y algoritmos de clasificación, determinamos qué características eran distintivas de estos estados. En resumen, la información extraída con métodos de NLP nos permitieron reflejar estado subjetivos y cognitivos de individuos de forma mínimamente invasiva. Entre las ventajas de esta herramienta se destacan su objetividad, escalabilidad y validez ecológica (asociada al habla cotidiana) que sobrepasa los análisis provenientes de cuestionarios estandarizados o basados en escalas. Además, permite la adquisición de datos a distancia y el análisis automático, economizando los posibles gastos de los analistas. Cabe destacar que ciertos aspectos del lenguaje no pueden ser captados con este método (e.g. pausas o tonalidades del habla) resultando en una pérdida de información asociado al discurso.

En nuestra publicación (Sanz et al., 2021), el NLP permitió plasmar tanto el contenido semántico (asociado a los tópicos discursivos) como no semántico (relacionado con la topología del lenguaje) del estado psicodélico. Los *Embeddings* de palabras habilitaron la posibilidad de analizar similitudes en el discurso que excedían las palabras específicas pronunciadas por los participantes. Mientras que los grafos de discurso plasmaron la estructura del habla de forma global y local. Además, las métricas de entropía, variabilidad semántica y rango mostraron aumentos significativos respecto al estado de placebo. En conjunto, nuestros resultados indican que el LSD produce la desorganización del discurso. Estos hallazgos se alinean tanto con conocimientos previos del estado psicodélico (Nichols, 2016; Nour et al., 2016; Kometer and Vollenweider, 2016; Preller and Vollenweider, 2016) como con la teoría del cerebro entrópico, que propone a los agonistas del receptor $5 - HT2_A$ como agentes que aumentan la entropía de la actividad cerebral y los procesos de pensamientos asociados (Carhart-Harris, 2018).

Como se mencionó previamente (ver Capítulo 4), múltiples estudios acerca de la microdosificación psicodélica basan sus resultados en cuestionarios y escalas diseñadas para identificar efectos de dosis completa que podrían no ser de utilidad para capturar cambios cognitivos próximos al umbral de percepción (Anderson et al., 2019b; Lea et al., 2020a; Polito and Stevenson, 2019). Por ejemplo, en nuestra publicación (Sanz et al., 2022a), mostramos que la variabilidad semántica no presenta diferencias significativas entre participantes bajo los efectos de la microdosis y la condición de placebo. Los resultados presentados en el Capítulo 4 muestran que el NLP permite reflejar diferencias en la experiencia subjetiva que tuvieron los participantes, por ejemplo en la emotividad del discurso. Estas diferencias pueden surgir tanto de los efectos agudos de la droga como de una consecuencia asociada a los efectos de cegado y sesgos generados por la expectativa. Tampoco se puede descartar que se deba a una combinación de ambos factores. Resulta complejo desambiguar entre las opciones ya que es probable que los sujetos no cegado fueron los que experimentaron efectos subjetivos más intensos. En el estudio del estado psicodélico se debe tener en cuenta que el ambiente y la naturaleza del experimento producirán sesgos en los efectos medibles. Por un lado, los experimentos basados en cuestionarios estandarizados o escalables resultarán en respuestas restrictivas y limitadas. Por otro lado, las investigaciones basadas en la performance de tareas específicas pueden resultar repetitivas y tediosas, creando una fuerte dependencia con la predisposición de los participantes al realizar la tarea. En contraparte, el NLP surge como una herramienta capaz de cuantificar características del lenguaje espontáneo y extraer métricas que exceden preguntas específicas. Además, abre un abanico de posibilidades de medición como la estimación de la dosis administrada, la inferencia del tipo de sustancia consumida, la predicción de efectos terapéuticos, entre otros.

En general, el estudio de enfermedades neurodegenerativas precisa del seguimiento continuo de los pacientes. Entre las principales desventajas de esta práctica se destaca el seguimiento de itinerarios clínicos (visitas médicas semanales, estimación de progresión, ajuste de dosis medicinales, entre otros) que puden resultar tediosas tanto para los pacientes como para los investigadores. Además, resulta importante considerar el presupuesto que conlleva realizar estudios clínicos y el costo de los profesionales encargados del seguimiento. Por otro lado, uno de los inconvenientes intrínsecos de la neuropsicología radica en la subjetividad de los examinadores, que puede variar según la formación o experiencia previa de los mismos (Orimaye et al., 2018). Es por ello que métodos como el NLP que conllevan análisis automatizados, escalables y objetivos surgen como una necesidad en el estudio de las enfermedades neurodegenerativas. Se debe tener en cuenta que las investigaciones basadas en este tipo de métodos deben presentar concordancia con los conocimientos clínicos previamente medidos y extensamente estudiados para que puedan aplicarse (de la Fuente Garcia et al., 2020).

En nuestra publicación (Sanz et al., 2022b) desarrollamos métodos de NLP para extraer métricas que reflejan déficits cognitivos de la EA estudiados en el ámbito clínico. En las investigaciones de enfermedades neurodegenerativas, la comparación entre grupos de pacientes con distintos trastornos es necesaria para determinar la especificidad de los resultados. Es importante considerar que ciertas métricas podrían estar asociadas al deterioro cognitivo general, mientras que otras podrían relacionarse con el deterioro típico producido por cada trastorno (e.g. Alzheimer - Déficits en la memoria y en la navegación espacial; Demencia Frontotemporal - Dificultades en la interacción social; Parkinson - Déficits motoros, etc).

Actualmente, los modelos complejos de lenguaje pre-entrenados para distinguir grupos de enfermedades neurodegenerativas, permiten lograr predicciones precisas sin introducir características específicamente diseñadas para cada problema (Yuan et al., 2020; Balagopalan et al., 2019). Sin embargo, estos métodos funcionan como "cajas negras" y conllevan a resultados poco interpretables. Además, no siempre arrojan resultados mejores que modelos sencillos e interpretables (ver Capítulo 6). Pensando a futuro, existe la posibilidad de que la compresión de estos algoritmos sea más extensa y permitan encontrar *features* distintivas que ignoramos hoy en día.

Apéndice A

Apéndice A

A.1. Categorías gramaticales

El análisis de las categorías gramaticales se realizó para descartar diferencias provenientes de computar similitudes semánticas entre el mismo o diferente tipo de palabra. Es decir, puede que la similitud sea mayor entre dos adjetivos que entre un adjetivo y un sustantivo solo por compartir la categoría gramatical. Para descartar este factor, se contó la ocurrencia normalizada de las principales partes del discurso (verbos, adjetivos, adverbios y sustantivos) de cada transcripción. Las comparaciones estadísticas entre condiciones mostraron diferencias en la cantidad de verbos ($Z = 2.05$; $p = 0.04$) para el Tiempo 1 y de adjetivos ($Z = 2.10$; $p = 0.036$) para el Tiempo 2. Sin embargo, estas diferencias no superaron las correcciones de Bonferroni ($\alpha = 0.013$). En la Figura A.1 se muestran gráficos de barra con el promedio de ocurrencia normalizada de las categorías para cada tiempo.

A.2. Resumen de métricas

A continuación se resumen las métricas utilizadas en los clasificadores junto con las medidas asociadas a la desorganización del discurso. En la Figura A.2, se muestran las matrices de correlación entre todas las variables.

Métricas utilizadas en el clasificador semántico:

Similitud semántica (i.e coseno) promedio entre cada palabra de la entrevista y la lista de 10 términos ("visuales", "patrón", "relajación", "escuchar", "estado de ánimo", "estimular", "normal", "ego", "miedo" y "realidad"). Una mayor puntuación entre un

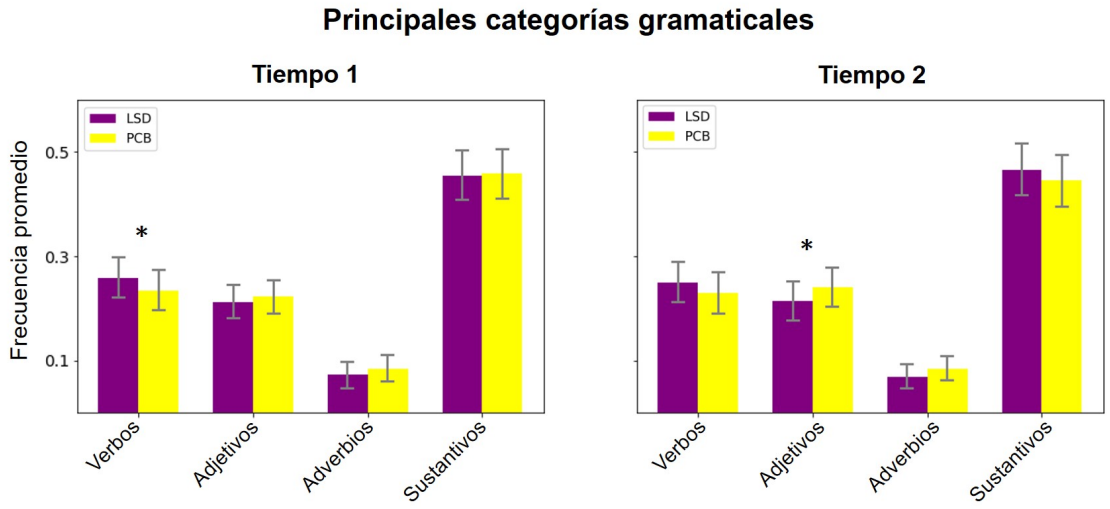

FIGURA A.1: Gráfico de barras de las principales categorías gramaticales. Las barras indican la frecuencia de ocurrencia promedio de la categoría, junto con su desvío estándar. (* $p < 0.05$, sin correcciones; test de rango con signo de Wilcoxon).

término y una entrevista implica que las palabras articuladas por el participante se encuentran cerca del término en el *embedding* de *Word2Vec*, por lo tanto presentan un significado similar.

Métricas utilizadas en el clasificador no semántico:

Medidas relacionadas a la conectividad de las entrevistas (ASP, LSC y Diámetro) y métricas asociadas a la recurrencia (L1, L2, L3, RE, PE y CC).

Métricas asociadas a la desorganización del discurso:

- Entropía de Shannon: Cuantifica la cantidad de incertidumbre (desorden) del discurso.
- Variabilidad semántica: Mide la discontinuidad del discurso (inversamente proporcional a la coherencia).
- Rango: Determina la compresibilidad del discurso. Un alto rango implica un mayor número de términos linealmente independientes (i.e. diferente contenido semántico).

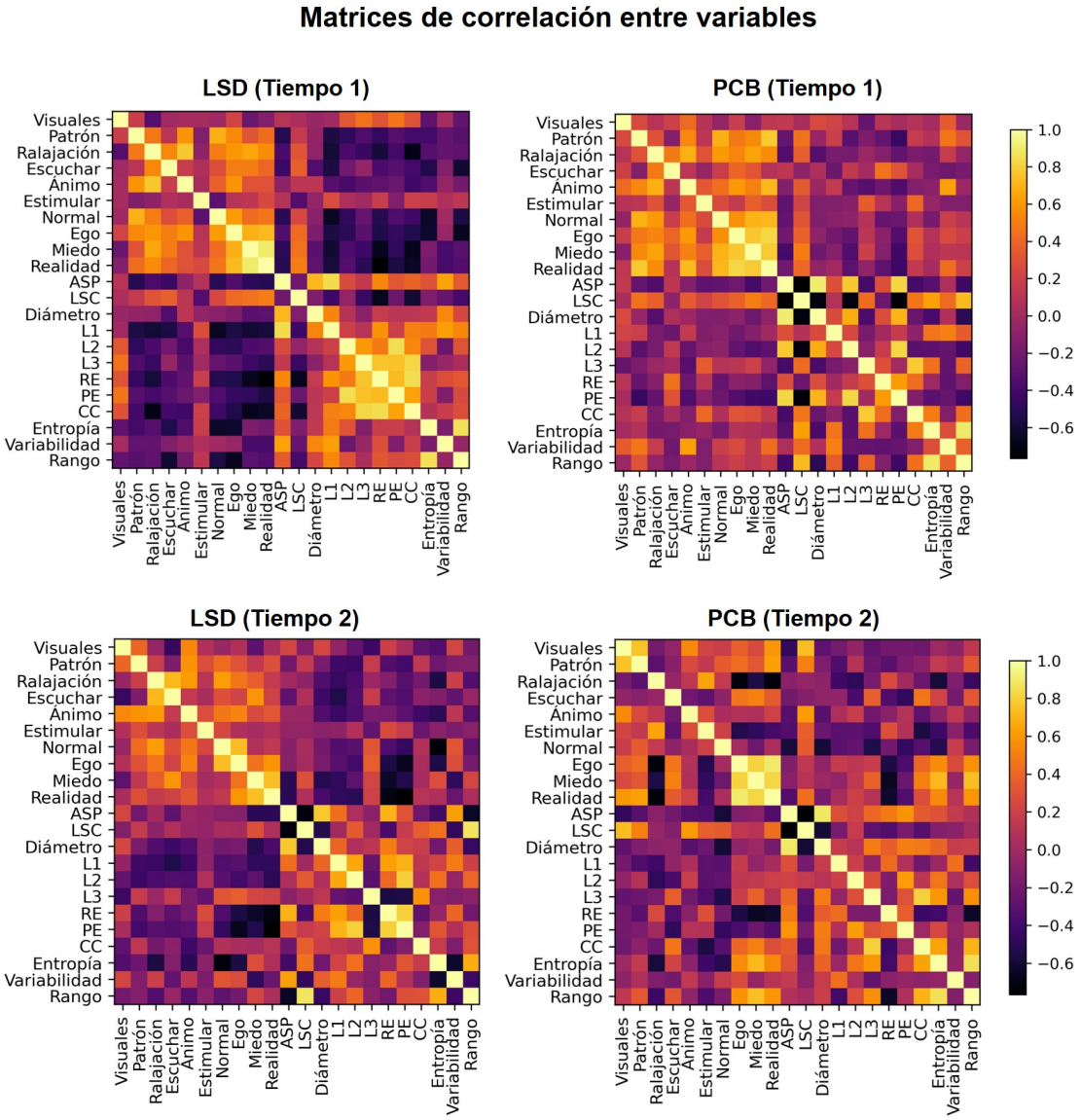

FIGURA A.2: Matrices de correlación de las métricas para cada condición y Tiempo.

Apéndice B

Apéndice B

B.1. Distribuciones para los subgrupos

En la Figura B.1, se presentan los gráficos de caja de verbosidad y MSS para todas las combinaciones posibles (dosis activa/placebo y cegado/no cegado). Los participantes bajo la dosis activa mostraron valores medios de verbosidad y MSS mayores que para el placebo. Sin embargo, solo se encontraron diferencias significativas en participantes no cegados; para el grupo cegado se obtuvieron p-valores mayores a 0.0099 (verbosidad) y 0.11 (MSS) que no superaron correcciones de Bonferroni. Las diferencias significativas en verbosidad se encontraron para las respuestas relacionadas a "percepción" ($T = 5$, $p = 7.1x10^{-4}$, $r = 0.93$), "estado de ánimo" ($T = 9$, $p = 0.0014$, $r = 0.88$) y "estado de alerta" ($T = 20$, $p = 0.0075$, $r = 0.74$); las demás respuestas dieron p-valores superiores a 0.01. Los valores de mediana y rango intercuartílico para los participantes bajo la dosis activa fueron $M = 58.5$, $IQR = 90.8$ ("percepción"); $M = 45$, $IQR = 66.8$ ("estado de ánimo"), y $M = 56.5$, $IQR = 50$ ("estado de alerta"). Mientras que para el grupo bajo la condición placebo se obtuvo $M = 5$, $IQR = 9$ ("percepción"); $M = 5$, $IQR = 15$ ("estado de ánimo"), y $M = 28$, $IQR = 40$ ("estado de alerta"). Para la medida MSS, se encontraron diferencias significativas en la respuesta relacionada con el "estado de ánimo" ($T = 17$, $p = 0.0049$, $r = 0.78$); las demás arrojaron valores p-valores mayores a 0.031. El valor de mediana e intervalo intercuartílico para participantes bajo la dosis activa fue: $M = 0.44$, $IQR = 0.15$. Mientras que bajo la condición placebo se obtuvo: $M = 0.14$, $IQR = 0.29$.

Las comparaciones de participantes cegados frente a no cegados restringidos a dosis activa/placebo no mostraron diferencias significativas. Para la restricción de dosis activa, se encontraron p-valores mayores a 0.023 (verbosidad) y 0.022 (MSS). Mientras que para la condición de placebo los p-valores fueron superiores a 0.023 (verbosidad) y 0.085 (MSS)

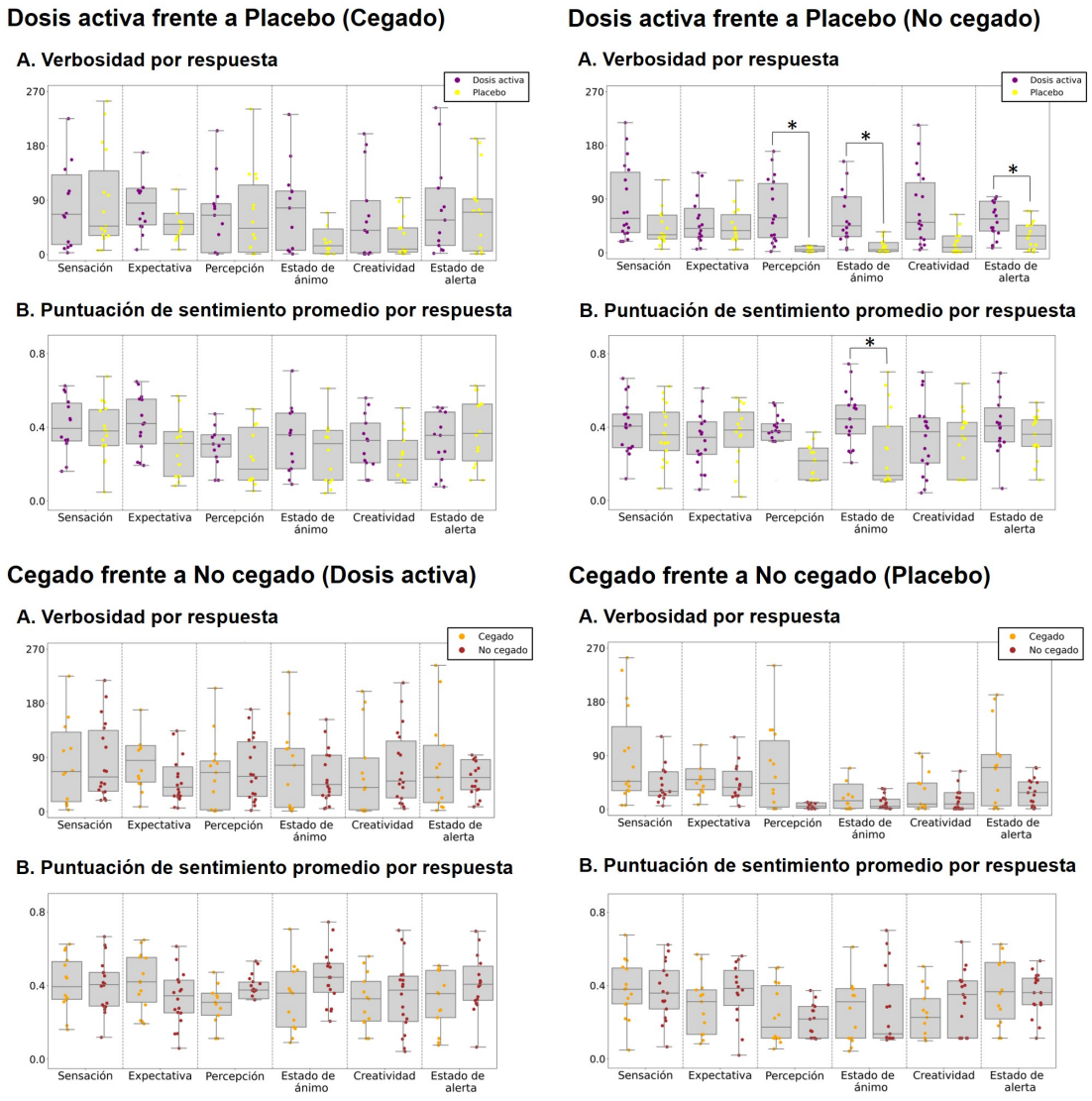

FIGURA B.1: Únicamente la comparación entre dosis activa frente a placebo, restringida a participantes no cegados, mostró diferencias significativas. Gráficos de caja de la verbosidad (A) y MSS (B) para: la dosis activa frente al placebo (panel superior) restringida a los participantes cegados (izquierda) y no cegados (derecha); participantes cegados frente a no cegados (panel inferior) restringidos a la dosis activa (izquierda) y el placebo (derecha) (* $p < 0.05$, con correcciones de Bonferroni. Superior: prueba de rango con signo de Wilcoxon; inferior: Prueba U de Mann-Whitney)

B.2. Variabilidad semántica

Los resultados del análisis de variabilidad semántica se muestran en la Figura B.2. Como se mencionó anteriormente (Ver Sección 2.2.1), esta medida refleja la coherencia del discurso. Frecuentes saltos en la distancia semántica implican que palabras sucesivas presentan significados disímiles, dando como resultado una variabilidad alta (disminución de la coherencia).

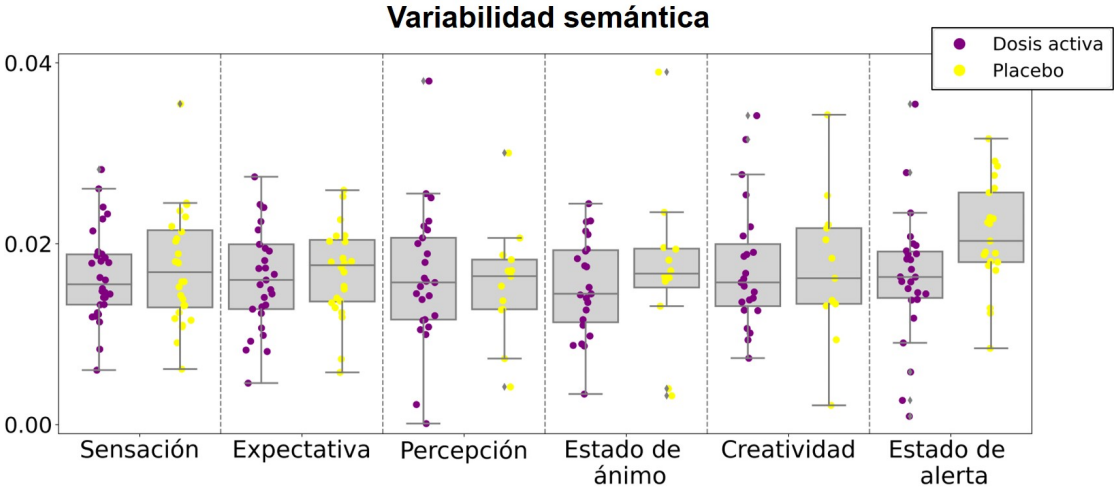

FIGURA B.2: La variabilidad semántica no mostró diferencias significativas al comparar la dosis activa frente a la condición de placebo. Todos los p-valores fueron mayores a 0.023 (Prueba de rango con signo de Wilcoxon). Gráficos de caja de la variabilidad semántica para cada pregunta y condición.

Apéndice C

Apéndice C

C.1. Estimación de potencia

Realizamos una estimación de potencia en G*Power 3.1.1. Basados en las pruebas ANOVAs de medidas independientes, establecimos (1) un nivel alfa de $p = 0.05$; (2) una potencia de 0.8; (3) un tamaño de efecto $\eta_p^2 = 0.25$ en línea con referencias de la literatura actual. Encontramos que un tamaño muestral de 33 resultaba en un potencia de 0.96.

C.2. Importancia de *features*

Para identificar las *features* que presentaban mayor relevancia en la clasificación entre pacientes con EA frente al GC, calculamos la importancia de la forma descripta en la Sección 5.2.5. En la Figura C.1, se muestran los resultados obtenidos para la clasificación real (azul) y aleatoria (rojo). Los valores de granularidad 6, 9 y 12; junto con la variabilidad semántica resultaron más informativas que el resto. En la clasificación aleatoria, las *features* mostraron un nivel de importancia comparable con el azar (50 %).

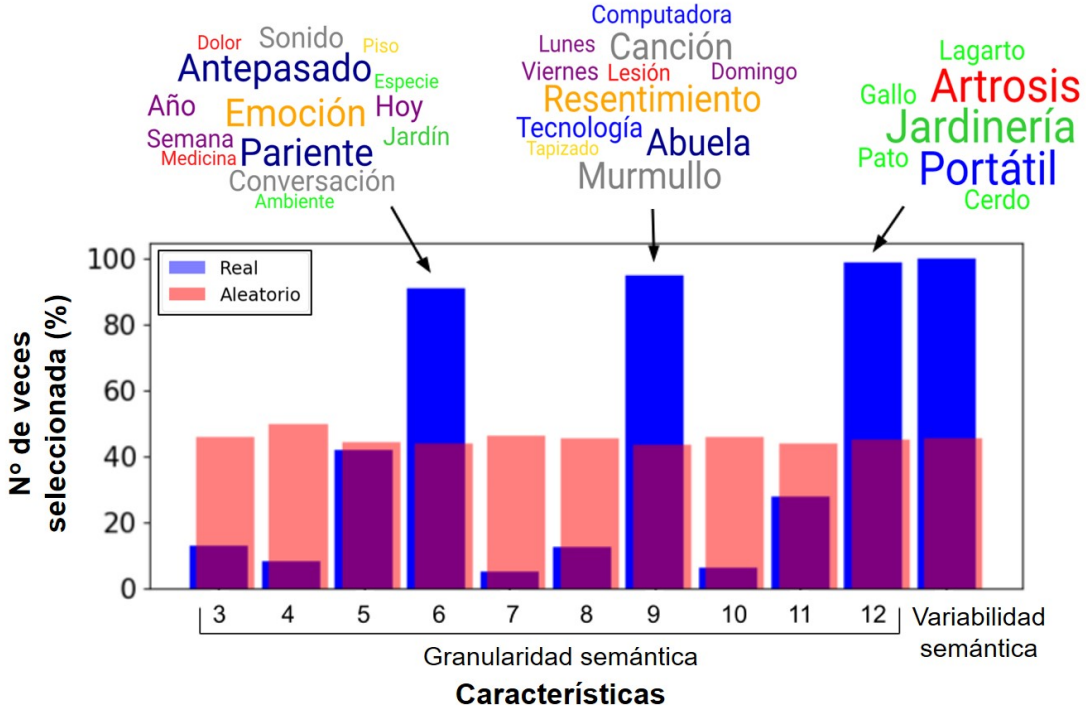

FIGURA C.1: Porcentaje de casos en los que se seleccionó cada *feature* en función de su puntuación F del ANOVA. Las barras azules representan los resultados reales, mientras que las rojas denotan los resultados obtenidos de la asignación aleatoria de clases. Los intervalos de granularidad correspondientes a las tres *features* más relevantes, se emparejaron con nubes de palabras que incluían sustantivos de la granularidad correspondiente. Los colores representan que los términos se encuentran semánticamente relacionados.

Bibliografía

Agurto, C., Cecchi, G. A., Norel, R., Ostrand, R., Kirkpatrick, M., Baggott, M. J., Wardle, M. C., Wit, H., and Bedi, G. (2020). Detection of acute 3,4-methylenedioxymethamphetamine (MDMA) effects across protocols using automated natural language processing. *Neuropsychopharmacology, 45(5).*

Ahmed, S., Haigh, A. M. F., de Jager, C. A., and Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain, 136(12), 3727-3737.*

Anandarajan, M., Hill, C., and Nolan, T. (2019). Text preprocessing. *In Practical Text Analytics (pp. 45-59). Springer, Cham.*

Anderson, T., Petranker, R., Christopher, A., Rosenbaum, D., Weissman, C., Dinh-Williams, L. A., Hui, K., and Hapke, E. (2019a). Psychedelic microdosing benefits and challenges: an empirical codebook. *Harm reduction journal, 16(1), 1-10.*

Anderson, T., Petranker, R., Rosenbaum, D., Weissman, C. R., Dinh-Williams, L. A., Hui, K., Hapke, E., and Farb, N. A. S. (2019b). Microdosing psychedelics: personality, mental health, and creativity differences in microdosers. *Psychopharmacology, 236(2), 731-740.*

Ash, S., Moore, P., Vesely, L., and Grossman, M. (2007). The decline of narrative discourse in Alzheimer's disease. *Brain and Language, 1(103), 181-182.*

Atasoy, S., Roseman, L., Kaelen, M., Kringelbach, M. L., Deco, G., and Carhart-Harris, R. L. (2017). Connectome-harmonic decomposition of human brain activity reveals dynamical repertoire re-organization under LSD. *Scientific reports, 7(1), 1-18.*

Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2019). To BERT or not to BERT: comparing speech and language-based approaches for Alzheimer's disease detection. *arXiv preprint arXiv:2008.01551.*

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology, 51(6), 585-594.*

Bedi, G., Cecchi, G. A., Slezak, D. F., Carrillo, F., Sigman, M., and De Wit, H. (2014). A window into the intoxicated mind? Speech as an index of psychoactive drug effects. *Neuropsychopharmacology, 39(10), 2340.*

Bershad, A. K., Schepers, S. T., Bremmer, M. P., Lee, R., and de Wit, H. (2019). Acute subjective and behavioral effects of microdoses of lysergic acid diethylamide in healthy human volunteers. *Biological psychiatry, 86(10), 792-800.*

Birba, A., García-Cordero, I., Kozono, G., Legaz, A., Ibáñez, A., Sedeño, L., and García, A. M. (2017). Losing ground: Frontostriatal atrophy disrupts language embodiment in parkinson's and huntington's disease. *Neuroscience & Biobehavioral Reviews, 80, 673-687.*

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., and Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology, 8, 269.*

Breiman, L. (1996). Bias, variance, and arcing classifiers. *Tech. Rep. 460, Statistics Department, University of California, Berkeley, CA, USA.*

Breiman, L. (2001). Random forests. *Machine learning, 45(1), 5-32.*

Breiman, L. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis.*

Bucks, R. S., Singh, S., Cuerden, J. M., and Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology, 14(1), 71-91.*

Cameron, L. P., Nazarian, A., and Olson, D. E. (2020). Psychedelic microdosing: Prevalence and subjective effects. *Journal of psychoactive drugs, 52(2), 113-122.*

Canu, E., Agosta, F., Battistella, G., Spinelli, E. G., DeLeon, J., Welch, A. E., et al. (2020). Speech production differences in english and italian speakers with nonfluent variant PPA. *Neurology, 94(10), e1062-e1072.*

Carhart-Harris, R. L. (2018). The entropic brain-revisited. *Neuropharmacology, 142, 167–178.*

Carhart-Harris, R. L. and Friston, K. (2019). Rebus and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews, 71(3), 316-344.*

Carhart-Harris, R. L., Kaelen, M., Whalley, M. G., Bolstridge, M., Feilding, A., and Nutt, D. J. (2015). LSD enhances suggestibility in healthy volunteers. *Psychopharmacology.*

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Nutt, D., et al. (2014). The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience, 20.*

Carhart-Harris, R. L., Muthukumaraswamy, S., Roseman, L., Kaelen, M., Droog, W., Murphy, K., Nutt, D. J., et al. (2016). Neural correlates of the LSD experience revealed by multimodal neuroimaging. *Proceedings of the National Academy of Sciences.*

Carrillo, F., Mota, N., Copelli, M., Ribeiro, S., Sigman, M., Cecchi, G., et al. (2013). Automated speech analysis for psychosis evaluation. *In Machine learning and interpretation in neuroimaging (pp. 31–39). Cham: Springer.*

Cavanna, F., Muller, S., de la Fuente, L. A., Zamberlan, F., Palmucci, M., Janeckova, L., Kuchar, M., Pallavicini, C., and Tagliazucchi, E. R. (2022). Microevidence for microdosing with psilocybin mushrooms: a double-blind placebo-controlled study of subjective effects, behavior, creativity, perception, cognition, and brain activity. *Biorxiv.*

Chapleau, M., Aldebert, J., Montembeault, M., and Brambati, S. M. (2016). Atrophy in Alzheimer's disease and semantic dementia: an ALE meta-analysis of voxel-based morphometry studies. *Journal of Alzheimer's disease, 54(3), 941-955.*

Chen, G. F., Xu, T. H., Yan, Y., Zhou, Y. R., Jiang, Y., Melcher, K., and Xu, H. E. (2017). Amyloid beta: structure, biology and structure-based therapeutic development. *Acta Pharmacologica Sinica, 38(9), 1205-1235.*

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. *In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (pp. 276-286).*

Corcoran, C. M., Carrillo, F., Fernandez-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., and Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry, 17(1), 67-75.*

Cox, D. J., Garcia-Romeu, A., and Johnson, M. W. (2021). Predicting changes in substance use following psychedelic experiences: natural language processing of psychedelic session narratives. *The American journal of drug and alcohol abuse, 47(4), 444-454.*

Cox, D. J. and Johnson, M. W. (2021). Verbal behavior related to drug reinforcement in polysubstance cannabis users: Comparison across drugs. *Experimental and clinical psychopharmacology, 30(2), 172.*

de la Fuente Garcia, S., Ritchie, C. W., and Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease, 78(4), 1547-1574.*

De Saussure, F., Bally, C., Sechehaye, A., Riedlinger, A., Alonso, A., and Sechehaye, A. (1987). *Curso de lingüística general.*

Delgado, C., Araneda, A., and Behrens, M. I. (2019). Validación del instrumento montreal cognitive assessment en español en adultos mayores de 60 años. *Neurología, 34(6), 376-385.*

Dickson, D. W. et al. (2009). Neuropathological assessment of parkinson's disease: refining the diagnostic criteria. *The Lancet Neurology, 8(12), 1150-1157.*

Dijkstra, K., Bourgeois, M. S., Allen, R. S., and Burgio, L. D. (2004). Conversational coherence: Discourse analysis of older adults with and without dementia. *Journal of Neurolinguistics, 17(4), 263-283.*

Dolder, P. C., Schmid, Y., Müller, F., B. S., and Liechti, M. E. (2016). LSD acutely impairs fear recognition and enhances emotional empathy and sociality. *Neuropsychopharmacology, 41(11), 2638.*

Dugger, B. N. and Dickson, D. W. (2017). Pathology of neurodegenerative diseases. *Cold Spring Harbor perspectives in biology, 9(7), a028035.*

Duong, S., Patel, T., and Chang, F. (2017). Dementia: What pharmacists need to know. *Canadian Pharmacists Journal, 150(2), 118-129.*

Elvevåg, B., Foltz, P. W., Weinberger, D. R., and Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research, 93(1-3), 304-316.*

Erkkinen, M. G., Kim, M. O., and Geschwind, M. D. (2018). Clinical neurology and epidemiology of the major neurodegenerative diseases. *Cold Spring Harbor perspectives in biology, 10(4), a033118.*

Eyigoz, E., Courson, M., Sedeño, L., Rogg, K., Orozco-Arroyave, J. R., Nöth, E., et al. (2020a). From discourse to pathology: automatic identification of parkinson's disease patients via morphological measures across three languages. *Cortex, 132, 191-205.*

Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., and Naylor, M. (2020b). Linguistic markers predict onset of alzheimer's disease. *EClinicalMedicine, 28, 100583.*

Fadiman, J. and Korb, S. (2019). Might microdosing psychedelics be safe and beneficial? an initial exploration. *Journal of psychoactive drugs, 51(2), 118-122.*

Family, N., Maillet, E. L., Williams, L. T. J., Krediet, E., Carhart-Harris, R. L., Williams, T. M., Nichols, C. D., Goble, D. J., and Raz, S. (2020). Safety, tolerability, pharmacokinetics, and pharmacodynamics of low dose lysergic acid diethylamide (lsd) in healthy older volunteers. *Psychopharmacology, 237(3), 841-853.*

Family, N., Vinson, D., Vigliocco, G., Kaelen, M., Bolstridge, M., Nutt, D. J., and Carhart-Harris, R. L. (2016). Semantic activation in LSD: evidence from picture naming. *Language, Cognition and Neuroscience, 1(10), 1320-1327.*

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease, 49(2), 407-422.*

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics data analysis.*

García, A. M., Bocanegra, Y., Herrera, E., Moreno, L., Carmona, J., Baena, A., et al. (2018). Parkinson's disease compromises the appraisal of action meanings evoked by naturalistic texts. *Cortex, 100, 111-126.*

García, A. M., Carrillo, F., Orozco-Arroyave, J. R., Trujillo, N., Bonilla, J. F. V., Fittipaldi, S., et al. (2016). How language flows when movements don't: an automated analysis of spontaneous discourse in parkinson's disease. *Brain and language, 162, 19-28.*

Ghannay, S., Favre, B., Esteve, Y., and Camelin, N. (2016). Word embedding evaluation and combination. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).*

Giffard, B., Desgranges, B., Nore-Mary, F., Lalevée, C., de la Sayette, V., Pasquier, F., and Eustache, F. (2001). The nature of semantic memory deficits in alzheimer's disease: new insights from hyperpriming effects. *Brain, 124(8), 1522-1532.*

Girn, M., Mills, C., Roseman, L., Carhart-Harris, R. L., and Christoff, K. (2020). Updating the dynamic framework of thought: Creativity and psychedelics. *Neuroimage, 213, 116726.*

Gotvaldova, K., Hajkova, K., Borovicka, J., Jurok, R., Cihlarova, P., and Kuchar, M. (2021). Stability of psilocybin and its four analogs in the biomass of the psychotropic mushroom psilocybe cubensis. *Drug testing and analysis, 13(2), 439-446.*

Gupta, B., Rawat, A., Jain, A., Arora, A., and Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications.*

Hefferon, J. (2020). *Linear Algebra (4th ed.).* Orthogonal Publishing L3C.

Henry, J. D. and Crawford, J. R. (2004). Verbal fluency deficits in Parkinson's disease: a meta-analysis. *Journal of the International Neuropsychological Society, 10(4), 608-622.*

Herd, P., Carr, D., and Roan, C. (2014). Cohort profile: Wisconsin Longitudinal Study (WLS). *International Journal of Epidemiology 43, 34-41.*

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science, 349(6245), 261-266.*

Hodges, J. R., Salmon, D. P., and Butters, N. (1992). Semantic memory impairment in alzheimer's disease: failure of access or degraded knowledge?. *Neuropsychologia, 30(4), 301-314.*

Hoehn, M. M. and Yahr, M. D. (1998). Parkinsonism: onset, progression, and mortality. *Neurology, 50(2), 318-318.*

Huang, J. and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering.*

Hughes, A. J., Daniel, S. E., Kilford, L., and Lees, A. J. (1992). Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases. *Journal of neurology, neurosurgery & psychiatry, 55(3), 181-184.*

Hutten, N., Mason, N. L., Dolder, P. C., and Kuypers, K. (2019a). Motives and side-effects of microdosing with psychedelics among users. *International Journal of Neuropsychopharmacology, 22(7), 426-434.*

Hutten, N., Mason, N. L., Dolder, P. C., and Kuypers, K. P. C. (2019b). Self-rated effectiveness of microdosing with psychedelics for mental and physical health problems among microdosers. *Frontiers in psychiatry, 10, 672.*

Hutten, N., Mason, N. L., Dolder, P. C., Theunissen, E. L., Holze, F., Liechti, M. E., Feilding, A., Ramaekers, J. G., and Kuypers, K. P. C. (2020). Mood and cognition after administration of low lsd doses in healthy volunteers: A placebo controlled dose-effect finding study. *European Neuropsychopharmacology, 41, 81-91.*

Ilias, L. and Askounis, D. (2022). Explainable identification of dementia from transcripts using transformer networks. *IEEE Journal of Biomedical and Health Informatics, 26(8), 4153-4164.*

Iversen, L., Iversen, S., Bloom, F. E., and Roth, R. H. (2017). Introduction to neuropsychopharmacology. *Oxford University Press.*

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R (2nd edition), (pp.327-360).* New York: Springer.

Johnstad, P. G. (2018). Powerful substances in tiny amounts: An interview study of psychedelic microdosing. *Nordic Studies on Alcohol and Drugs, 35(1), 39-51.*

Kaelen, M., Barrett, F. S., Roseman, L., Lorenz, R., Family, N., Bolstridge, M., et al. (2015). LSD enhances the emotional response to music. *Psychopharmacology, 232(19), 3607–3614.*

Kaertner, L. S., Steinborn, M. B., Kettner, H., Spriggs, M. J., Roseman, L., Buchborn, T., Balaet, M., Timmermann, C., Erritzoe, D., and Carhart-Harris, R. L. (2021). Positive expectations predict improved mental-health outcomes linked to psychedelic microdosing. *Scientific reports, 11(1), 1-11.*

Kaprinis, S. and Stavrakaki, S. (2007). Morphological and syntactic abilities in patients with alzheimer's disease. *Brain and Language, 1(103), 59-60.*

Kenton, J. D. M. W. C. and Toutanova, L. K. (20189,). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of naacL-HLT (Vol. 1, p. 2).*

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications, 1-32.*

Kometer, M. and Vollenweider, F. X. (2016). Serotonergic hallucinogen-induced visual perceptual alterations. *In Behavioral neurobiology of psychedelic drugs (pp. 257–282). Berlin, Heidelberg: Springer.*

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of BERT. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 4365-4374).*

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden markov model. *Computer Speech Language.*

Kuypers, K. P. (2020). The therapeutic potential of microdosing psychedelics in depression. *Therapeutic advances in psychopharmacology, 10, 2045125320950567.*

Kuypers, K. P., Ng, L., Erritzoe, D., et al. (2019). Microdosing psychedelics: More questions than answers? an overview and suggestions for future research. *Journal of Psychopharmacology, 33(9), 1039-1057.*

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., et al. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1(1), 112-124.*

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. discourse processes. *Discourse processes.*

Laske, C., Sohrabi, H. R., Frost, S. M., López-de Ipiña, K., Garrard, P., Buscema, M., et al. (2015). Innovative diagnostic tools for early detection of alzheimer's disease. *Alzheimer's & Dementia, 11(5), 561-578.*

Laws, K. R., Duncan, A., and Gale, T. M. (2010). "normal" semantic–phonemic fluency discrepancy in Alzheimer's disease? a meta-analytic study. *Cortex, 46(5), 595-601.*

Lea, T., Amada, N., and Jungaberle, H. (2020a). Psychedelic microdosing: A subreddit analysis. *Journal of Psychoactive Drugs, 52(2), 101-112.*

Lea, T., Amada, N., Jungaberle, H., Schecke, H., and Klein, M. (2020b). Microdosing psychedelics: Motivations, subjective effects and harm reduction. *International Journal of Drug Policy, 75, 102600.*

Lea, T., Amada, N., Jungaberle, H., Schecke, H., Scherbaum, N., and Klein, M. (2020c). Perceived outcomes of psychedelic microdosing as self-managed therapies for mental and substance use disorders. *Psychopharmacology, 237(5), 1521-1532.*

Lewis, D. (1975). Languages and language.

Liddy, E. D. (2001). Natural language processing.

Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. *AI Open.*

Lord, L. D., Expert, P., Atasoy, S., Roseman, L., Rapuano, K., Lambiotte, R., Cabral, J., et al. (2019). Dynamical exploration of the repertoire of brain networks at rest is modulated by psilocybin. *NeuroImage, 199, 127-142.*

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems.*

Luong, M. T., Sutskever, I., Le, Q. V., Vinyals, O., and Zaremba, W. (2015). Addressing the rare word problem in neural machine translation. *In Proceedings of the*

53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.

Madhoushi, Z., Hamdan, A., and Zainudin, S. (2015). Sentiment analysis techniques in recent works. *In 2015 science and information conference (SAI), IEEE.*

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).*

Marona-Lewicka, D., Thisted, R. A., and Nichols, D. E. (2005). Distinct temporal phases in the behavioral pharmacology of LSD: Dopamine D 2 receptor-mediated effects in the rat and implications for psychosis. *Psychopharmacology, 180(3), 427–435.*

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal.*

Mikolov, T., Yih, W. T., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. *In Hlt-naacl, volume 13, pages 746– 751.*

Mota, N. B., Furtado, R., Maia, P. P., Copelli, M., and Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis. *Scientific reports.*

Mota, N. B., Vasconcelos, N. A. P., Lemos, N., Pieretti, A. C., and Kinouchi, O. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS ONE.*

Muthukumaraswamy, S. D., Forsyth, A., and Lumley, T. (2021). Blinding and expectancy confounds in psychedelic randomized controlled trials. *Expert review of clinical pharmacology, 14(9), 1133-1152.*

Nakamura, A. E., Opaleye, D., Tani, G., and Ferri, C. P. (2015). Dementia underdiagnosis in brazil. *The Lancet, 385(9966), 418-419.*

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b.*

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics.*

Newman, M. E. J. (2010). *Networks: An Introduction.* Oxford University Press, page 9-10.

Nichols, D. E. (2016). Psychedelics. *Pharmacological reviews, 68(2), 264-355.*

Nichols, E., Szoeke, C. E., Vollset, S. E., Abbasi, N., Abd-Allah, F., Abdela, J., et al. (2019). Global, regional, and national burden of alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology, 18(1), 88-106.*

Norel, R., Agurto, C., Heisig, S., Rice, J. J., Zhang, H., Ostrand, R., Wacnik, P. W., Ho, B. K., Ramos, V. L., and Cecchi, G. A. (2020). Speech-based characterization of dopamine replacement therapy in people with parkinson's disease. *npj Parkinson's Disease, 6(1), 1-8.*

Nour, M. M., Evans, L., Nutt, D., and Carhart-Harris, R. L. (2016). Ego-dissolution and psychedelics: validation of the ego-dissolution inventory (EDI). *Frontiers in human neuroscience, 10, 269.*

Nutt, D. J., King, L. A., and Nichols, D. E. (2013a). Effects of schedule I drug laws on neuroscience research and treatment innovation. *Nature Reviews Neuroscience, 14(8), 577-585.*

Nutt, D. J., King, L. A., and Nichols, D. E. (2013b). New victims of current drug laws. *Nature Reviews Neuroscience, 14(12), 877-877.*

Olson, J. A., Suissa-Rocheleau, L., Lifshitz, M., Raz, A., and Veissiere, S. P. L. (2020). Tripping on nothing: placebo psychedelics and contextual factors. *Psychopharmacology, 237(5), 1371-1382.*

Ona, G. and Bouso, J. C. (2020). Potential safety, benefits, and influence of the placebo effect in microdosing psychedelic drugs: A systematic review. *Neuroscience & Biobehavioral Reviews, 119, 194-203.*

Orimaye, S. O., Wong, J. S. M., and Golden, K. J. (2014). Learning predictive linguistic features for alzheimer's disease and related dementias using verbal utterances. *In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality (pp. 78-87).*

Orimaye, S. O., Wong, J. S. M., and Wong, C. P. (2018). Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia. *PloS one, 13(11), e0205636.*

Ozel-Kizil, E. T., Bastug, G., and Kirici, S. (2020). Semantic and episodic memory performances of patients with alzheimer's disease and minor neurocognitive disorder: Neuropsychiatry and behavioral neurology/mild cognitive impairment/early symptomatic disease. *Alzheimer's & Dementia, 16, e039310.*

Patterson, K., Nestor, P. J., and Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature reviews neuroscience, 8(12), 976-987.*

Pistono, A., Jucla, M., Bézy, C., Lemesle, B., Le Men, J., and Pariente, J. (2019). Discourse macrolinguistic impairment as a marker of linguistic and extralinguistic functions decline in early alzheimer's disease. *International Journal of Language Communication Disorders, 54(3), 390-400.*

Poewe, W. et al. (2017). Parkinson disease. *Nature reviews Disease primers, 3(1), 1-21.*

Polito, V. and Stevenson, R. J. (2019). A systematic study of microdosing psychedelics. *PLoS One, 14(2), e0211023.*

Preller, K. H. and Vollenweider, F. X. (2016). Phenomenology, structure, and dynamic of psychedelic states. *Behavioral neurobiology of psychedelic drugs, 221-256.*

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes (3rd edition).* Cambridge University Press, page 795.

Prochazkova, L., Lippelt, D. P., Colzato, L. S., Kuchar, M., Sjoerds, Z., and Hommel, B. (2018). Exploring the effect of microdosing psychedelics on creativity in an open-label natural setting. *Psychopharmacology, 235(12), 3401-3413.*

Raskovsky, I., Slezak, D. F., Diuk, C., and Cecchi, G. A. (2010). The emergence of the modern concept of introspection: a quantitative linguistic analysis. *In Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas (pp. 68-75).*

Rentoumi, V., Raoufian, L., Ahmed, S., de Jager, C. A., and Garrard, P. (2014). Features and machine learning classification of connected speech samples from patients with autopsy proven alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease, 42(s3), S3-S17.*

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research, 11, 95-130.*

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?" Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).*

Rogers, S. L. and Friedman, R. B. (2008). The underlying mechanisms of semantic memory loss in Alzheimer's disease and semantic dementia. *Neuropsychologia, 46(1), 12-21.*

Rong, X. (2014). Word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738.*

Rootman, J. M., Kryskow, P., Harvey, K., Stamets, P., Santos-Brault, E., Kuypers, K. P. C., Polito, V., Bourzat, F., and Walsh, Z. (2021). Adults who microdose psychedelics report health related motivations and lower levels of anxiety and depression compared to non-microdosers. *Scientific reports, 11(1), 1-11.*

Rosner, B. (2015). *Fundamentals of biostatistics.* Cengage learning.

Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. ieee transactions on systems, man, and cybernetics. *IEEE transactions on systems, man, and cybernetics.*

Sankar, H. and Subramaniyaswamy, V. (2017). Investigating sentiment analysis using machine learning approach. *In 2017 International Conference on Intelligent Sustainable Systems (ICISS), IEEE.*

Santamaría-García, H., Baez, S., Reyes, P., Santamaría-García, J. A., Santacruz-Escudero, J. M., Matallana, D., et al. (2017). A lesion model of envy and schadenfreude: legal, deservingness and moral dimensions as revealed by neurodegeneration. *Brain, 140(12), 3357-3377.*

Sanz, C., Cavanna, F., M. S., de la Fuente, L., Zamberlan, F., Palmucci, M., Janeckova, L., Kuchar, M., García, A. M., and Tagliazucchi, E. (2022a). Natural language signatures of psilocybin microdosing. *Psychopharmacology.*

Sanz, C., Pallavicini, C., Carrillo, F., Zamberlan, F., S. M., Mota, N., Copelli, M., Ribeiro, S., Nutt, D., Carhart-Harris, R., and Tagliazucchi, E. (2021). The entropic tongue: disorganization of natural language under lsd. *Consciousness and cognition.*

Sanz, C., Zamberlan, F., Erowid, E., Erowid, F., and Tagliazucchi, E. (2018). The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports. *Frontiers in neuroscience, 12, 7.*

Sanz, C., C., F., Slachevsky, A., Forno, G., Gorno Tempini, M. L., Villagra, R., Ibáñez, A., Tagliazucchi, E., and García, A. M. (2022b). Automated text-level semantic markers of alzheimer's disease. *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring.*

Schartner, M. M., Carhart-Harris, R. L., Barrett, A. B., Seth, A. K., and Muthukumaraswamy, S. D. (2017). Increased spontaneous meg signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Scientific reports, 7(1), 1-12.*

Scheltens, P. et al. (2021). Alzheimer's disease. *The Lancet, 397(10284), 1577-1590.*

Schenberg, E. E. (2021). Who is blind in psychedelic research? letter to the editor regarding: blinding and expectancy confounds in psychedelic randomized controlled trials. *Expert review of clinical pharmacology, 14(10), 1317-1319.*

Schmid, H. (1994). Part-of-speech tagging with neural networks. *In Proceedings of the 15th conference on Computational linguistics - Volume 1.*

Silverman, L. (2021). Verbal semantic coherence. *In: Volkmar, F.R. (eds) Encyclopedia of Autism Spectrum Disorders. Springer, Cham.*

Spitzer, M., Thimm, M., Hermle, L., Holzmann, P., Kovar, K. A., Heimann, H., Gouzoulis-Mayfrank, E., Kischka, U., and Schneider, F. (1996). Increased activation of indirect semantic associations under psilocybin. *Biological psychiatry, 39(12), 1055-1057.*

Symeonidis, S., Effrosynidis, D., and Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications.*

Szigeti, B., Kartner, L., Blemings, A., Rosas, F., Feilding, A., Nutt, D. J., Carhart-Harris, R. L., and Erritzoe, D. (2021). Self-blinding citizen science to explore psychedelic microdosing. *Elife, 10, e62878.*

Tagliazucchi, E. (2022). Natural language as a window into the subjective effects and neurochemistry of psychedelic drugs. *Preprint.*

Tagliazucchi, E., Carhart-Harris, R., Leech, R., Nutt, D., and Chialvo, D. R. (2014). Enhanced repertoire of brain dynamical states during the psychedelic experience. *Human brain mapping, 35(11), 5442-5456.*

Tagliazucchi, E. et al. (2017). *Un libro sobre drogas.* El gato y la caja.

Tagliazucchi, E., Roseman, L., Kaelen, M., Orban, C., Muthukumaraswamy, S. D., Murphy, K., et al. (2016). Increased global functional connectivity correlates with LSD-induced ego dissolution. *Current Biology, 26(8), 1043–1050.*

Torralva, T., Roca, M., Gleichgerrcht, E., Lopez, P., and Manes, F. (2010). Ineco frontal screening (ifs): A brief, sensitive, and specific tool to assess executive functions in dementia–erratum. *Journal of the International Neuropsychological Society, 16(5), 737-747.*

van Elk, M., Fejer, G., Lempe, P., Prochazckova, L., Kuchar, M., Hajkova, K., and Marschall, J. (2021). Effects of psilocybin microdosing on awe and aesthetic experiences: a preregistered field and lab-based study. *Psychopharmacology, 239(6), 1705-1720.*

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

Wang, B., Wang, A., Chen, F., Wang, Y., and Kuo, C. C. J. (2019). Evaluating word embedding models: methods and experimental results. *APSIPA transactions on signal and information processing.*

Weiner, M. F., Neubecker, K. E., Bret, M. E., and Hynan, L. S. (2008). Language in alzheimer's disease. *Journal of Clinical Psychiatry, 69(8), 1223-1227.*

Wiese, G., Weissenborn, D., and Neves, M. (2017). Neural domain adaptation for biomedical question answering. *arXiv:1706.03610.*

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144.*

Yanakieva, S., Polychroni, N., Family, N., Williams, L. T. J., Luke, D. P., and Terhune, D. B. (2019). The effects of microdose lsd on time perception: a randomised, double-blind, placebo-controlled trial. *Psychopharmacology, 236(4), 1159-1170.*

Yang, D. and Powers, D. M. (2005). Measuring semantic similarity in the taxonomy of wordnet. *Australian Computer Society.*

Yin, Z. and Shen, Y. (2018). On the dimensionality of word embedding. advances in neural information processing systems. *Advances in neural information processing systems.*

Yu, S., Indurthi, S. R., Back, S., and Lee, H. (2018). A multi-stage memory augmented neural network for machine reading comprehension. in proceedings of the workshop on machine reading for question answering. *In Proceedings of the workshop on machine reading for question answering (pp. 21-30).*

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., and Church, K. (2020). Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. *In INTERSPEECH (Vol. 2020, pp. 2162-6).*

Zamberlan, F., Sanz, C., Martínez Vivot, R., Pallavicini, C., Erowid, F., Erowid, E., and Tagliazucchi, E. (2018). The varieties of the psychedelic experience: a preliminary study of the association between the reported subjective effects and the binding affinity profiles of substituted phenethylamines and tryptamines. *Frontiers in integrative neuroscience.*

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *In Proceedings of the IEEE international conference on computer vision (pp. 19-27).*

Zimmerer, V. C., Wibrow, M., and Varley, R. A. (2016). Formulaic language in people with probable Alzheimer's disease: A frequency-based approach. *Journal of Alzheimer's Disease, 53(3), 1145-1160.*