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

Caracterización automática del lenguaje natural en sujetos con alteraciones mentales

Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Facundo Carrillo

Director: Dr Diego Fernandez Slezak

Buenos Aires, 2012

CARACTERIZACIÓN AUTOMÁTICA DEL LENGUAJE NATURAL EN SUJETOS CON ALTERACIONES MENTALES

La psiquiatría computacional tiene como parte de sus objetivos caracterizar las disfunciones mentales analizando el comportamiento de las personas a través de la resolución de cómputos que permiten identificar estados mentales. Para ello, se introduce el modelo computacional que busca identificar rastros computacionales de la actividad cognitiva y neural. En este trabajo empleamos técnicas del procesamiento del lenguaje natural para la caracterización de estados mentales. Entre las técnicas empleadas podemos destacar el uso de Latent Semantic Analysis, POS-tagging, Lematizacón y análisis topológicos en grafos inducidos, entre otras. Como caso de estudio tomamos discursos en pacientes psiquiátricos y en sujetos bajo el efecto de la droga Ecstasy. Para el primer caso contamos con reportes de tres grupos de sujetos: esquizofrénicos, maníacos y control. Tomando los reportes armamos grafos inducidos a partir de ellos y medimos características topológicas. Estas medidas son usadas para armar clasificadores que luego contrastamos contra el diagnostico clínico. Los resultados muestran que el diagnóstico automático de sujetos basado en sus discursos coinciden con el diagnóstico manual hecho por psiquiatras experimentados. Para el segundo caso de estudio (Ecstasy) tomamos reportes de sujetos en 4 condiciones distintas: habiendo consumido un placebo, MDMA (dosis baja), MDMA (dosis alta) y Metanfetamina. Los resultados indican que podemos clasificar los relatos de un sujeto en cada una de las cuatro condiciones. Esta tesis presenta evidencia de que los métodos automáticos de diagnóstico son posibles con el uso de herramientas de interpretación del lenguaje natural.

Palabras claves: psiquiatría computacional, análisis automático, lenguaje natural, esquizofrenia, manía, Ecstasy, clasificadores.

AGRADECIMIENTOS

A mi familia que fomentó desde siempre mis curiosidades dando respuestas y sobre todo más motivos para hacer preguntas, soportando siempre mi incansable motivación por transmitirles lo que tanto me apasiona: Mamá, Papá, Yiyi, Kike, Marta, Michel, Laura y mis hermanos Juli y Thiago.

A Luna por las incansables caminatas llenas de momentos de reflexión compartidos.

A mi gran grupo de TPs y amigos, *Cocomielsort*, por haberme enseñado y compartido tanto todos estos años: Doc, Fede y Brian.

A mis otros compañeros y amigos la facu, que conocí tan al principio y tan al final, pero fueron partes de mi carrera: Ale Mataloni, Agus Ciracco, Emi Mancusco, Martin De Micheli, Curtu, Romi, JuanMa, Mariano Bianchi, Pablo Brusco, Marianito Semelman, Kevin Allekotte, Juli Peller, Nico Echebarrena, Pablo Zivic, Dardo Marasca, Felipe Schargorosdky, Adrian Tamburri, Karina Borgna, Blas Couto, a los Divus y muchos más que me estoy olvidando pero que también estuvieron.

A mis compañeros de Orga2 por enseñarme a ser mejor docente y divertirnos dando clases.

A mis amigos de la vida por bancarse lo pesado que fui siempre con la facultad: Gonza, Bimbo, Chris, Jorgito, Aro y Andy.

A todos los del LNI por hacerme sentir tan cómodo ni bien llegué.

A mis compañeros de LIAA por abrirme las puertas en la forma en que lo hicieron.

Al jurado, por tomarse el trabajo de leer y evaluar mi tesis.

A Guillermo Cecchi por colaborar con esta tesis.

A Diego Fernández Slezak por ser un gran director, dedicarme todo el tiempo que necesité y enseñarme tanto.

A la Universidad de Buenos Aires y toda las personas que la componen por darme la posibilidad de estudiar en un lugar como Exactas y hacerme sentir orgulloso por ello.

Y sobre todo, a Lu, mi compañera de vida, por ser protagonista de esta carrera y tesis tanto como yo, por soportar mis cansancios, mal humor y lo pesado que soy con lo que me apasiona, por dar todo y bancar mis ritmos para enseñarme a ser más feliz.

¡Gracias!

A mis hermanos Juli y Thiago.

Somebody calls you, you answer quite slowly
A girl with kaleidoscope eyes

Cellophane flowers of yellow and green
Towering over your head
Look for the girl with the sun in her eyes
And she's gone

Lucy in the sky with diamonds
Lucy in the sky with diamonds
Lucy in the sky with diamonds

A Lu.

Índice general

1. INTRODUCCIÓN

La psiquiatría es la especialidad médica dedicada al estudio de las enfermedades mentales, cuyo objetivo es prevenir, diagnosticar y tratar a las personas con trastornos mentales. A grandes rasgos existen dos tipos de tratamientos psiquiátricos: el biológico y el psicoterapéutico. Los tratamientos del tipo biológico tienen como uno de sus objetivos encontrar las bases fisiológicas de las enfermedades mentales, estudiando aspectos biológicos del cerebro, resultando en varios casos exitoso a través del diseño y desarrollo de fármacos anti-depresivos y anti-psicóticos, entre otros. Sin embargo, existe una gran brecha en el entendimiento entre el abordaje molecular y la expresión clínica que se manifiesta en enfermedades como esquizofrenia, depresión y ansiedad.

La *Psiquiatría computacional* tiene como uno de sus objetivos caracterizar las disfunciones mentales analizando el comportamiento de las personas a través de la resolución de cómputos que permitan identificar los estados mentales [1]. Para ello, se introduce el *modelo computacional* que busca identificar rastros computacionales de la actividad cognitiva y neural, tal como proponía Alan Turing y su concepción de las funciones mentales como módulos de procesamiento de información en una plataforma de hardware muy particular: el cerebro [2]. Un ejemplo de estos modelos computacionales es el aprendizaje por refuerzos que plantea que a través de paradigmas experimentales sobre predicciones de refuerzo pueden estudiarse aspectos importantes de enfermedades mentales asociadas a alteraciones dopamínicas [3] o asociando errores en la predicción de valor a depresión o esquizofrenia [4].

Recientemente, utilizando la disponibilidad de repositorios públicos digitales de estudios del cerebro, se han llevado a cabo estudios masivos buscando estructuras en las imágenes asociadas a funciones mentales [5]. A través de técnicas de *datamining*, se ha podido establecer una relación entre los *tópicos* tratados en artículos científicos sobre imágenes del cerebro y algunas enfermedades mentales, como por ejemplo esquizofrenia. Estos artículos muestran que a partir de técnicas de extracción automática de información es posible evaluar (y diagnosticar) estados mentales. Sin embargo, ninguno de ellos aborda el problema con el repositorio de productos del pensamiento más vasto hoy en día: el texto escrito.

La forma en que nos expresamos nos permite entender cómo el cerebro organiza los conceptos confeccionando ideas y nos permite tanto identificar como clasificar procesos cognitivos de más alto nivel a través del estudio de características del discurso.

En esta tesis nos proponemos explotar diversas técnicas de análisis sintáctico y semántico con el objetivo de generar un analizador automático para el diagnóstico no supervisado de enfermedades psiquiátricas e identificación de estados inducidos por drogas.

1.1. Enfermedades Psiquiátricas

La psiquiatría describe la psicosis como una pérdida de contacto con la realidad causada por diversos factores, como la esquizofrenia y la manía. En la práctica clínica, existe un *Manual diagnóstico y estadístico de los trastornos mentales* (DSM) [6], elaborado a partir de datos empíricos con una metodología descriptiva. Tiene como objetivo mejorar la comunicación entre clínicos no pretendiendo explicar las diversas patologías, ni propo-

ner líneas de tratamiento farmacológico o psicoterapéutico, por lo que es usado como un lineamiento y no como un método de diagnóstico. Una característica del DSM, relevante a este trabajo, es que siempre debe ser utilizado por personas con experiencia clínica, ya que se usa como una guía que debe ser acompañada de juicio clínico, además de los conocimientos profesionales, lo que imposibilita la automatización del DSM como método de diagnóstico.

Recientemente, un programa de investigación dirigido por Sidarta Ribeiro propone abordar el análisis cuantitativo de textos provenientes de pacientes diagnosticados con enfermedades psiquiátricas [7]. Dicho análisis se basa en el registro de relatos sobre sueños recientes en pacientes esquizofrénicos y maníacos, así como en sujetos control. Luego, se construyen manualmente grafos inducidos a partir de los relatos registrados, y se cuantifican distintas medidas sobre estos grafos permitiendo discriminar en dos grupos de sujetos enfermos (esquizofrénicos y maníacos) y el grupo control.

Partiendo de este trabajo como base, nos propusimos desarrollar técnicas de procesamiento de texto que permitirían automatizar el proceso. En el trabajo original, cada reporte es transcripto a partir de relatos orales y procesado manualmente en elementos gramaticales canónicos. En primer lugar, se implementará un mecanismo automático de filtrado para luego comparar con los resultados obtenidos del limpiado manual. El éxito de la clasificación depende del conjunto de *features* que logre separar bien a los sujetos de los distintos grupos. Con la automatización del proceso es posible generar distintos tipos de grafos inducidos, ampliando el espectro de *features* a considerar para la clasificación. Para la evaluación de los clasificadores usaremos WEKA [8] que nos permitirá comparar distintos métodos. Dado que tenemos los reportes originales, probaremos el desempeño de nuestro modelo en comparación con el trabajo antes mencionado.

Este trabajo sirve como un caso de prueba para la evaluación de técnicas automáticas para la extracción de información en textos y su aplicación para el diagnóstico de estados mentales.

A partir de la evidencia de éxito en la automatización del proceso se esta diseñando, en colaboración con el grupo de Sidarta Ribeiro en Brasil, una aplicación orientada al médico: PsychoGraph. La motivación de este proyecto es brindarle al médico psiquiatra una herramienta clínica de diagnóstico basada en el análisis automático del discurso.

1.2. Ecstasy

El *Ecstasy* es el nombre coloquial de la droga MDMA (3,4-methylenedioxy-N-methy lamphetamine). El MDMA es una droga psicoactiva sintética, químicamente similar a la *methamphetamine* y al alucinógeno *mescaline*. Según diversos estudios [9–14], MDMA produce sentimientos de euforia, calidez emocional, bienestar, felicidad asociada a la extroversión y la sociabilidad, distorsión del tiempo, entre otros efectos. El MDMA tiene su principal efecto en el cerebro, en particular en las neuronas que usan la serotonina como neurotransmisor. El sistema serotoninérgico juega un rol importante en la regulación del humor, agresión, actividad sexual, sueño y sensibilidad al dolor. En un nivel molecular, el MDMA se acopla a los transportadores serotoninérgicos (responsables de remover la serotonina de la sinapsis) prolongando la señal entre neuronas. También promueve la liberación de dopamina pero en mucha menor medida. Puede producir confusión, depresión, problemas de sueño y ansiedad. El efecto puede ocurrir después de tomar la droga o incluso

tiempo después. Uno de los riesgos de la ingesta de ecstasy es la psicosis inducida, cuyos síntomas pueden confundirse con esquizofrenia.

Teniendo en cuenta las alteraciones mentales bajo el efecto de esta droga, es esperable que los productos del pensamiento se vean afectados. La detección de estos cambios en la producción cognitiva permiten la descripción de marcadores estereotipados del pensamiento, abriendo la posibilidad de utilizar estas condiciones como ventanas para observar la mente y ampliar nuestra comprensión sobre el funcionamiento del cerebro. En particular, nos interesa estudiar los cambios en la producción de discursos o relatos en sujetos bajo el efecto de MDMA. A partir de técnicas de extracción automática de información y análisis semántico en textos, nos proponemos estudiar la estructura de los relatos de sujetos en distintas condiciones de estados mentales. Analizaremos distintas medidas intentando cuantificar los cambios en la generación de ideas producto de los cambios anímicos y los cambios de la percepción producidos por la ingesta de drogas.

2. TRABAJOS RELACIONADOS

En este capítulo se introducen distintas técnicas de procesamiento del lenguaje natural y medidas en grafos de bibliografía que fueron usadas en distintas partes de esta tesis. Con respecto a técnicas de procesamiento del lenguaje natural se introducirán herramientas como: Part of Speech tagging, Lematización, Latent Semantic Analysis y The Google Similarity Distance. En cuanto a las medidas en grafos, se enumeraran que medidas fueron usadas. Los algoritmos para resolver estas últimas fueron tomados de la literatura.

2.1. Part-of-speech tagging

Part-of-speech tagging (POS-Tag) es una técnica de etiquetado gramatical automática que logra asignar a cada palabra de un texto su categoría gramatical o categoría morfosintáctica. Es una técnica difícil de implementar dado que no es decidible la categoría de una palabra analizándola sola sin su contexto ya que la misma puede tener distintas funciones gramaticales según su uso. Para resolver este inconveniente se recurre a poner en contexto la palabra con toda la dificultad que ésto significa al ser un proceso automático.

Para establecer una categorización , los métodos que se usan son estadísticos y dependen fuertemente del corpus. A su vez, la cantidad de categorías morfosintácticas varían según implementación-corpus usados. Por ejemplo, *Brown Corpus* [15] propone 87 categorías y *Penn Treebank corpus*[1] [16] 45-46.

En la tabla 2.1 se presenta una categorización simplificada propuesta por Loper y Bird [17] usada en la herramienta de procesamiento del lenguaje natural NLTK.

Como toda técnica de procesamiento del lenguaje natural basada en corpus, el éxito depende en parte de él. Los corpus necesarios para las técnicas de POS-tagging requieren mucho trabajo pues cuentan con etiquetado manual. El éxito también depende del paquete de algoritmos probabilísticos atrás de estas herramientas. En este caso, existen diferentes variantes de implementación, muchas basadas en Modelos Ocultos de Marcov [18]. Es importante destacar que como toda técnica basada en corpus, no es independiente al idioma.

Por ejemplo, usando la herramienta de POS-tagging de NLTK. La frase:

> "The present lineage of dogs was domesticated from gray wolves about 15,000 years ago."[2]

es etiquetada como se ve en la tabla 2.2.

2.2. Lematización

La lematización es un proceso lingüístico que consiste en la transformación de una forma flexionada (una palabra conjugada, con genero, plural/singular, etc) en su lema correspondiente. El lema de una palabra es el representante de las palabras con las variaciones mencionadas. Por ejemplo, las palabras: jugar, jugando, juego, comparten el mismo

[1] `http://www.cis.upenn.edu/~treebank/`

[2] Frase tomada de Wikipedia en inglés, del artículo *dog*. `http://en.wikipedia.org/wiki/Dog`

Tag	Significado	Ejemplos
ADJ	adjective	new, good, high, special, big, local
ADV	adverb	really, already, still, early, now
CNJ	conjunction	and, or, but, if, while, although
DET	determiner	the, a, some, most, every, no
EX	existential	there, there's
FW	foreign word	dolce, ersatz, esprit, quo, maitre
MOD	modal verb	will, can, would, may, must, should
N	noun	year, home, costs, time, education
NP	proper noun	Alison, Africa, April, Washington
NUM	number	twenty-four, fourth, 1991, 14:24
PRO	pronoun	he, their, her, its, my, I, us
P	preposition	on, of, at, with, by, into, under
TO	the word to	to
UH	interjection	ah, bang, ha, whee, hmpf, oops
V	verb	is, has, get, do, make, see, run
VD	past tense	said, took, told, made, asked
VG	present participle	making, going, playing, working
VN	past participle	given, taken, begun, sung
WH	wh determiner	who, which, when, what, where, how

Tab. 2.1: Categorización de TAGs de POS-tagging de Loper y Bird.

lema que es jugar. Automatizar este proceso es difícil y costoso computacionalmente. En muchos casos la lematización no solo depende de la forma flexionada, sino que está determinada por su contexto también. Las palabras con más de una acepción tienen que ser desambiguadas con información adicional de su entorno. Inicialmente, por limitaciones tecnológicas, los lematizadores funcionaban con heurísticas no basadas en corpus de datos sino en reglas de construcción de palabras. A medida que la tecnología lo permitió, los lematizadores se fueron volcando más a una base estadística, mejorando el desempeño con técnicas ajenas, como el POS-tagging para entender mejor el contexto de las frases y poder tomar decisiones con más datos.

Teniendo en cuenta estas técnicas presentadas, utilizando el POS-Tagging y luego lematizando el texto, nos proponemos automatizar el procedimiento de traducción a elementos gramaticales canónicos. Luego, con la secuencia de lemas generada (o elementos gramaticales canónicos) podemos construir el grafo inducido, para luego tomar las medidas que nos permitan caracterizar los discursos.

2.3. Medidas en grafos

Las medidas usadas son clásicas de la literatura de teoría de grafos [19].

- **Nodes:** La cantidad de nodos del grafo.
- **Edgs:** La cantidad de aristas del grafo.
- **PE:** La suma entre la cantidad de ejes paralelos entre un mismo par de nodos, para cualquiera del grafo.

Palabra	Etiqueta
The	DT
present	NN
lineage	NN
of	IN
dogs	NNS
was	VBD
domesticated	VBN
from	IN
gray	NN
wolves	NNS
about	IN
15000	CD
years	NNS
ago	RB
.	.

Tab. 2.2: Tabla de etiquetado de POS-tagging para las frase "The present lineage of dogs was domesticated from gray wolves about 15,000 years ago.üsando la herramienta de NLTK.

- **LCC:** Una componente conexa de un grafo es un subconjunto de nodos tal que para todo par de nodos en él existe un camino. También debe cumplir que no puede existir otro nodo del grafo que esté conectado a algún nodo de la componente. LCC es la cantidad de nodos en la máxima componente conexa. Para computar esta medida, computamos todas las componentes conexas y tomamos la de mayor cantidad de nodos.

- **LSC:** Una componente fuertemente conexa en un grafo dirigido es un subconjunto de nodos tal que para todo par de nodos a y b en él existe un camino entre a y b y entre b y a. LSC es la cantidad de nodos en la máxima componente fuertemente conexa. Para computar esta medida, computamos todas las componentes fuertemente conexas y tomamos la de mayor cantidad de nodos. Para calcular las componentes fuertemente conexas usamos el algoritmo de Tarjan [20].

- **ATD:** El grado de un nodo de un grafo es la cantidad de aristas que entran o salen de él. ATD es el promedio total de los grados para todos los nodos.

- **L1:** Cantidad de ciclos de un nodo. Para computar esta medida, tomamos la traza de la matriz de adyacencia del grafo.

- **L2:** Cantidad de ciclos de dos nodos. Para computar esta medida, tomamos la traza de la matriz de adyacencia del grafo elevada al cuadrado.

- **L3:** Cantidad de ciclos de tres nodos. Para computar esta medida, tomamos la traza de la matriz de adyacencia del grafo elevada al cubo.

Cada uno de estos parámetros será utilizado para caracterizar los discursos, y buscaremos clasificadores que a partir de estas medidas separen entre los distintos grupos.

Estas técnicas presentadas permiten evaluar la forma del discurso, no así su contenido. A continuación presentamos otros métodos que permiten extraer de forma automática elementos característicos del contenido de los discursos, sirviendo con parámetros adicionales para la clasificación.

2.4. Latent semantic analysis

Latent semantic analysis [21] (LSA) es una técnica de procesamiento del lenguaje natural que propone una noción espacial de las palabras. A partir de ésto se pueden definir distintas métricas, métodos de comparación y similitud entre otras medidas.

LSA esa una técnica fuertemente basada en corpus. Consiste en capturar la relación entre un conjunto de textos y las palabras contenidas en ellos asumiendo que dos términos son cercanos en significado si ocurren de una manera *similar* en los textos.

Para definir la similitud, se crea una matriz donde las filas representan palabras y las columnas los textos. El valor en la intersección representa la cantidad de ocurrencias que tiene la palabra en dicho texto. Como la matriz es muy grande se emplea una técnica de factorización de matriz, *Singular value decomposition*(SVD) para luego, reducir la dimensionalidad.

Dada una matriz M de $m \times n$, SVD propone una factorización $M = USV$ donde U es de $m \times m$ y es una *matriz unitaria* (i.e $U^t \times U = U \times U^t = I$), S es de $m \times n$ y es una matriz diagonal no negativa y V es de $n \times n$ y es una *matriz unitaria*. Los valores de la diagonal en S son *valores singulares*, ordenados decrecientemente.

Ésta técnica puede usarse para reducir el número de columnas preservando la similitud entre las filas. Una vez que se tiene la matriz reducida, la comparación de palabras consiste en tomar alguna medida (típicamente el coseno del ángulo) entre los dos vectores formados por las dos filas de las palabras. En el caso del coseno, los valores cercanos a 1 representan palabras similares y los valores cercanos a 0 representan palabras disímiles.

La Universidad de Colorado ofrece un sistema online [3] que permite hacer diversas consultas basándose en su implementación de LSA. Por ejemplo, si le pedimos los 10 términos más cercanos a *dog* usando el corpus *General_Reading_up_to_1st_year_college* obtenemos una tabla con dos columnas. La segunda es la palabra similar y la primera es su valor de similitud (donde 1 es cercano y 0 lejano), el resultado se puede consultar en la tabla 2.3. Podemos ver que las palabras más cercanas a *dog* son: ella misma, *barked* y *dogs*.

La efectividad del método depende del corpus de textos elegido, la dimensionalidad en la reducción y la medida de comparación vectorial.

Para este tesis usamos el corpus TASA. TASA es corpus de texto en inglés, compilado por Touchstone Applied Science Associates. Este corpus consiste en más de 37.000 documentos y más de 12 millones de palabras de un vocabulario de más de 90.000 palabras distintas. Los documentos consisten en textos generales para las educación desde el 1er año de *collage* en Estados Unidos, incluyendo partes de textos tomados de novelas, diarios, artículos entre otros. Es un corpus usado habitualmente para LSA.

La elección de dimensionalidad (esto es, cuántas componentes de la factorización de SVD vamos a usar) es otro factor importante para tener éxito en la medición de conceptos. Landauer y Dumais [21] estudiaron el efecto de cambiar este valor concluyendo un óptimo en 300 dimensiones.

[3] `http://lsa.colorado.edu/`

Similitud	Término
0.99	dog
0.86	barked
0.86	dogs
0.81	wagging
0.80	collie
0.79	leash
0.75	manilak
0.75	moops
0.75	moop
0.74	barking

Tab. 2.3: Ejemplo de 10 términos más cercanos a la palabra *dog* provisto por `http://lsa.colorado.edu/`.

En cuanto a la comparación vectorial usamos la medida coseno, dado que es la más aceptada y usada en la bibliografía.

2.5. The Google Similarity Distance

The Google Similarity Distance (GDS) [22] es una técnica que permite comparar la similitud entre dos palabras a partir de la cantidad de webs indexadas donde figuran cada palabra y el conjunto. Puede implementarse usando distintos motores de búsqueda.

Formalmente puede ser definida como: sea $f(x)$ la cantidad de web indexadas donde aparece la palabra x en un motor de búsqueda particular y $f(x, y)$ la cantidad en las que aparecen la palabra x e y, se propone *Normalized Google distance* como:

$$NGD(x, y) = \frac{max(logf(x), logf(y)) - logf(x, y)}{logM - min(logf(x), logf(y))}$$

donde M representa el numero total de web indexadas por el motor de búsqueda. De esta forma se obtiene un numero cercano a 1 para dos palabras disímiles, 0 para aquellas similares.

En principio parece difícil poder estimar M, pero se puede ver corriendo ejemplos que los resultados, para valores grandes de M, son invariantes. Un aspecto interesante de esta técnica es que usa como corpus toda la web, por lo que se puede aplicar en cualquier idioma, sin tener que pasar por un cambio de corpus y la calibración que ésto requiere.

Realizamos diversas pruebas en los buscadores Bing y Google y verificamos que éstos no dicen exactamente cuantas páginas indexadas tienen para una búsqueda determinada. Ésto genera resultados muy poco confiables y muy imprecisos. Sumado a eso, los buscadores tienen políticas que prohíben las búsquedas automáticas por lo que después de varios intentos bloquean la *IP* por tiempo indeterminado, o solicitan el ingreso de un *captcha* o devuelven resultados falsos a las búsquedas inutilizando este método.

Por las razones mencionadas este método fue descartado como medida semántica para esta tesis. Sin embargo, es presentado ya que propone una alternativa muy interesante para investigar a futuro, especialmente teniendo en cuenta la internacionalización de los métodos.

3. CASO DE ESTUDIO: DISCURSOS EN PACIENTES PSIQUIÁTRICOS

En este capítulo se aborda el desarrollo de técnicas automáticas de procesamiento de texto que permitan clasificar los relatos de acuerdo a la patología del paciente. Este procedimiento se basa en los datos generados por Mota y colaboradores [7]. En primer lugar, describimos el experimento clínico, el proceso de transcripción y la cuantificación de atributos del trabajo original. Luego, presentamos un método automático para algunos de estos procedimientos y verificamos que esta nueva cuantificación preserva las características pertinentes del trabajo original. Por último, probamos clasificadores usando los *features* provenientes de los métodos desarrollados.

3.1. Trabajo Original

La esquizofrenia y la manía son patologías psiquiátricas caracterizadas por alteraciones de la percepción y la expresión de la realidad, por lo que la producción de ideas puede tener claras particularidades, entendiendo al discurso como una ventana hacia la mente. En Mota et.al [7] se propone un método para cuantificar medidas del desorden del pensamiento en sujetos psicóticos analizando la morfología del discurso.

En ese trabajo, se entrevistaron 8 pacientes esquizofrénicos, 8 pacientes maníacos y 8 sujetos control. Las entrevistas tuvieron lugar en el Hospital Onofre Lopes, Federal University of Rio Grande do Norte, Natal, Brasil, en el año 2011 y 2012. Se entrevistaron sujetos que estuvieran tranquilos de modo que pudieran responder de forma comprometida. Durante las entrevistas, se midieron signos de conciencia y cambios de estado para saber si se encontraban aptos para responder las preguntas. Las entrevistas fueron realizadas y grabadas por Natalia Mota, una psiquiatra distinta al que diagnosticó originalmente a los pacientes. La entrevista constaba de una única consigna: reportar un sueño reciente. Una vez registrada la entrevista, éstas fueron transcriptas.

Una vez obtenidas las transcripciones se procedió a procesar el texto por medio de un mecanismo de interpretación manual, agrupando conjuntos de palabras en sintagmas. Un sintagma es un constituyente sintáctico que posee una función sintáctica específica dentro de la oración. Luego, a partir de la secuencias de sintagmas de texto se arma un grafo donde cada nodo representaba un sintagma y una arista entre dos de ellos representa la precedencia inmediata en la secuencia. A su vez, se identificó a los nodos como nodos dormidos (SN) o nodos despiertos (WN), según el sintagma asociado se refiera al relato del sueño o a un comentario por fuera del sueño respectivamente. Con el grafo armado, se tomaron medidas clásicas de la literatura de grafos, armando un conjunto de *features* de los relatos de cada sujeto para luego probar un clasificador automático el rendimiento del método.

A modo de ejemplo, a continuación se muestra una transcripción de un relato de un sujeto esquizofrénico:

eu lembro sim foi fiquei ruim mesmo fui internado só estava ouvindo muitas vozes estava muito fraco morava o lugar que eu morava era muito ruim e emprego também

Por otro lado, a continuación una transcripción de un relato de un sujeto maníaco:

tomei um banho de chuva o dia todinho foi, aí fui dormir tomei remédio aí deus me mostrou o guarda chuva

A partir de estos reportes, se construyeron los **grafos de sintagmas** para cada uno de los sujetos según el procedimiento de creación de grafos descripto anteriormente. En la figura 3.1 se puede ver un ejemplo de un reporte y su grafo del trabajo original. A la izquierda se ve un relato de un sujeto traducido manualmente al inglés donde las barras definen los sintagmas. En azul se ven los sintagmas que hacen referencia al relato del sueño y en rojo se ven los sintagmas referidos a comentarios que no son del sueño. A la derecha se ve el grafo inducido del relato de la izquierda. Los aristas se encuentran numeradas arbitrariamente. El grafo es construido a partir del reporte en portugués pero para su visualización los nodos fueron traducidos al inglés.

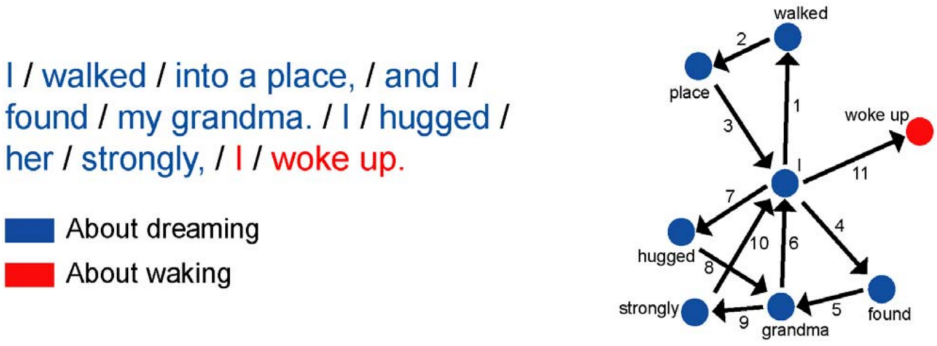

Fig. 3.1: Ejemplo de relato y grafo del trabajo de Mota et al. [7] . A la izquierda se ve un relato de un sujeto traducido manualmente al inglés donde las barras definen los sintagmas. En azul se ven los sintagmas que hacen referencia al relato del sueño y en rojo se ven los sintagmas referidos a comentarios que no son del sueño. A la derecha se ve el grafo inducido del relato de la izquierda. Los aristas se encuentran numeradas arbitrariamente. El grafo es construido a partir del reporte en portugués pero para su visualización los nodos fueron traducidos al inglés. La figura esta tomada del trabajo original.

Se calcularon 11 medidas de los grafos construidos con los textos originales en portugués para evaluar cada uno de los reportes: cantidad de nodos, cantidad de ejes, tamaño de la mayor componente conexa, cantidad de ciclo de 1 y 2 elementos, entre otros. En la tabla 3.1 se resumen todas las medidas obtenidas para lo reportes de los tres grupos de sujetos.

Cada uno de estos parámetros puede considerarse un feature a ser utilizado para entrenar un clasificador para separar entre los grupos. En primer lugar, se realizaron tests de hipótesis utilizando los parámetros tomados individualmente para verificar si existe diferencia significativa en alguna de las medidas que permita separar los distintos grupos. Para comparar entre los 3 grupos (esquizofrénico, maníaco y control) se utilizó el test KruskallWallis [23], y para testear los grupos de a 2 se utilizó el test de Wilcoxon [24]. En la tabla 3.2 se resumen estos tests, indicando para cada feature la significancia del test para la discriminación entre grupos. En rojo se indican aquellos test que dan significativos evidenciando que ese parámetro permite separar entre los diferentes grupos. Se observa que en todas las comparaciones se encuentra al menos un parámetro significativo mostrando diferencia poblacional.

Luego, utilizando estos parámetros, se construyeron clasificadores que obtuvieron un desempeño comparable con el de las evaluaciones estandarizadas sugeridas por el manual de diagnóstico DSM, tanto en su especificidad como su sensitividad. Sin embargo, como

Group	Initials	Nodes	Edges	PE	LCC	LSC	ATD	WN	WE	L1	L2	L3	Words per Report
Schizophr.	JA	27	42	1	27	24	3.111	1	2	1	5	19	67
Schizophr.	LF	18	23	0	18	9	2.555	0	0	1	3	7	49
Schizophr.	IG	21	34	3	21	13	3.238	11	19	1	3	7	58
Schizophr.	LL	13	19	1	13	12	2.923	0	0	1	3	4	32
Schizophr.	FV	14	13	0	9	3	1.857	1	1	3	3	6	43
Schizophr.	PS	11	13	0	11	9	2.364	0	0	0	2	0	25
Schizophr.	EL	14	18	0	14	9	2.571	1	1	2	2	5	41
Schizophr.	MG	10	12	0	7	5	2.400	3	2	0	2	0	24
Mania	AB	47	97	15	47	36	4.128	11	21	2	16	41	199
Mania	JB	28	57	6	28	25	4.071	16	30	2	4	20	118
Mania	DP	17	32	4	17	11	3.768	5	5	2	8	17	69
Mania	JM	25	29	0	19	13	2.320	9	10	1	3	4	50
Mania	FC	29	73	12	29	22	5.034	1	8	5	15	41	135
Mania	IS	28	58	9	28	18	4.143	26	54	3	11	21	109
Mania	OF	21	40	6	21	18	3.809	8	14	2	4	20	82
Mania	JG	44	86	20	37	29	3.909	26	43	5	15	20	185
Control	ML	24	31	1	24	4	2.583	5	5	0	0	0	71
Control	AA	28	56	6	28	23	4.000	11	23	2	10	26	116
Control	HF	24	46	4	24	22	3.833	7	11	1	9	22	89
Control	MJ	19	43	7	19	14	4.526	5	9	1	11	19	69
Control	JO	8	11	0	8	6	2.750	0	0	0	0	9	20
Control	OR	14	18	1	14	12	2.571	0	0	0	0	0	27
Control	FC	24	53	7	24	19	4.417	11	21	1	1	22	108
Control	JS	44	125	34	44	37	5.682	3	7	2	22	80	226

Tab. 3.1: Tabla de features de Mota et al. [7].

ya mencionamos, este método depende del procesamiento *manual* de la transcripción, convirtiendo el texto a secuencia de sintagmas. En consecuencia, en vistas de poder escalar estos análisis y construir herramientas de software que puedan asistir al psiquiatra, nos planteamos el desarrollo de técnicas de procesamiento automático de texto con el fin de obtener resultados similares de una forma completamente no supervisada.

3.2. Análisis automático

Como se mencionó anteriormente, un paso importante en el éxito del método propuesto por Mota et. al. es el procesamiento de texto a sintagmas. Automatizar este mecanismo no exacto resulta desafiante por lo que en esta sección nos focalizaremos en intentar replicar los resultados con métodos automáticos.

Como sistema *baseline* proponemos un análisis trivial, es decir sin procesamiento previo, al que llamaremos *naive*. Consiste en la creación de un grafo y la evaluación de medidas topológicas en él. Como siguiente paso intentaremos capturar las particularidades de la creación de sintagmas con métodos de lematización. A este análisis lo llamaremos *lematizado* donde también generaremos un grafo y tomaremos diversas medidas topológicas.

Atributos	**SxMxC** KruskallWallis	**SxM** Wilcoxon	**SxC** Wilcoxon	**MxC** Wilcoxon
N	0.0138	0.0028	0.1276	0.1354
E	0.0131	0.0030	0.0690	0.2786
LCC	0.0205	0.0050	0.0875	0.2875
LSC	0.0482	0.0078	0.2034	0.4564
ATD	0.0145	0.0070	0.0160	1
PE	0.0070	0.0031	0.0143	0.3641
L1	0.0129	0.0200	0.6892	0.0107
L2	0.0513	0.0025	1	0.2429
L3	0.0309	0.0050	0.0838	0.7765

Tab. 3.2: Tabla de test de Mota et al. [7], en rojo se ven las medidas separadoras significativas.

Automatizar el proceso de identificación de grupos requirió el uso de técnicas de procesamiento del lenguaje natural. Como se explicó en el capítulo *Trabajos relacionados*, muchas de las técnicas usadas están fuertemente basadas en corpus. Esto implica que su uso esta ligado a un idioma en particular. Los proceso de lematización y de POS-tagging no escapan a este regla. Como usaremos fuertemente estas herramientas para el análisis, decidimos trabajar con los textos en idioma inglés. Tomamos esta decisión porque el desempeño en los sistemas en inglés es bueno y bien conocido. Por lo tanto, como paso inicial del análisis tradujimos automáticamente los relatos del portugués al inglés usando Google Translate [1]. En este paso fue importante no realizar correcciones manuales dado que la intención fue automatizar el proceso íntegramente.

Por último intentaremos corroborar que los métodos automáticos que proponemos capturan las singularidades rescatadas en el trabajo original.

3.2.1. Grafo Naive

Como explicamos antes, en el primer análisis propusimos crear un grafo sin ninguna transformación del texto luego de su traducción. A este grafo lo llamamos *naive*. Por ejemplo, como se puede ver en la figura 3.2 para el reporte de un sujeto esquizofrénico se observa que el grafo inducido colapsa las dos apariciones de la palabra *crying* en un solo nodo.

A partir de este grafo, medimos y obtuvimos los mismos parámetros del trabajo original. Luego, realizamos los tests de hipótesis correspondientes para ver cuan separables eran los grupos. En la tabla 3.3 se muestran en rojo aquellas medidas separadoras para todas las comparaciones de los grupos. A diferencia de la tabla del trabajo original (tabla 3.2), no se obtiene ningún parámetro significativo para la separación entre los grupos Maníacos y Control. A simple vista observamos que se parecen, por lo decidimos medirlo. Teniendo calculados la significancia de cada uno de los tests, podemos medir cuánto se asemejan a los calculados originalmente (ver tabla 3.2).

Consideramos ambas tablas (tablas 3.2 y 3.3) como matrices de 9×4, digitalizadas a valores 1 y 0, siendo 1 para features significativos ($p < 0{,}05$) y 0 para valores no significativos ($p > 0{,}05$). Para ilustrar esta transformación, la digitalización de estas matrices pueden verse en la figura 3.3.

[1] http://translate.google.com/

With my evangelic daughter. She is crying. No, I only saw Jesus. She sometimes it appears to me laughing, crying and sometimes appears.

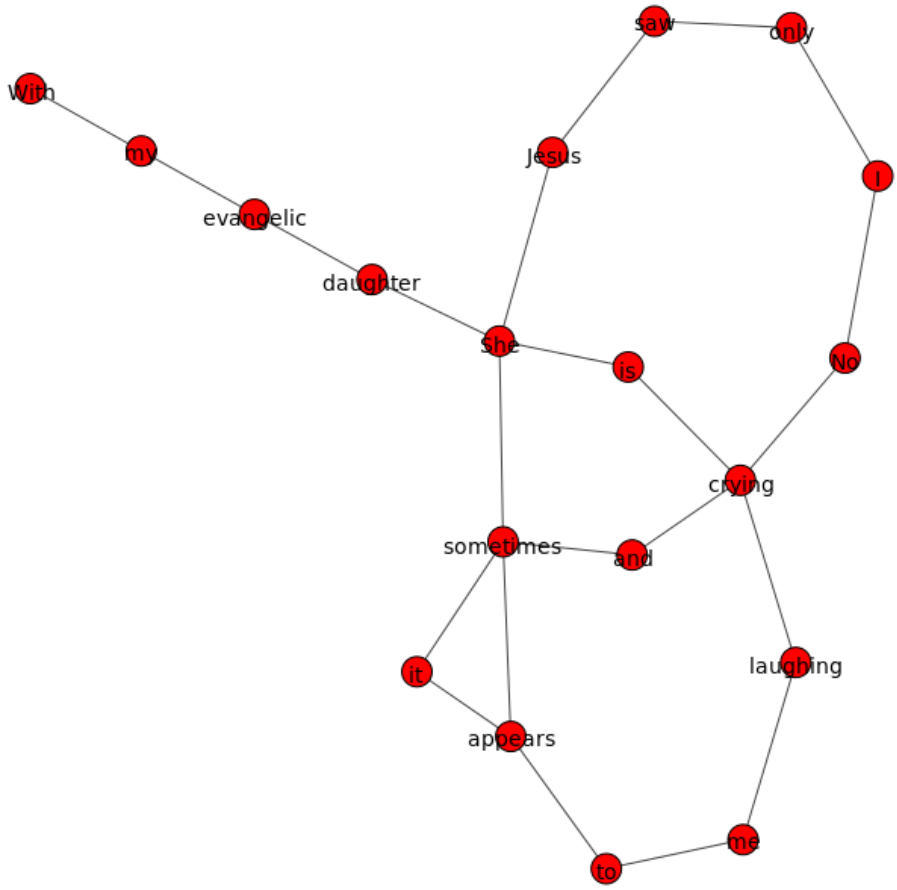

Fig. 3.2: Reporte de paciente esquizofrénico y su grafo *naive* inducido.

A partir de las matrices, podemos tratar de definir la significancia que hay entre estas dos. Para eso, simulamos 100000 matrices de 9×4 con valores 0 o 1 de forma aleatoria. Luego, tomamos la matrices del trabajo original y calculamos la cantidad de elementos iguales que tienen en el mismo lugar con cada una de las generadas al azar. De esta forma, obtenemos una distribución de valores correspondientes entre matrices. Nos interesa saber si la diferencia de valores correspondientes entre la matriz original y la automatizada se diferencia respecto a las matrices al azar. En la figura 3.4 se observa en azul el histograma de esta distribución. La posición de la barra roja representa la cantidad de valores comunes entre las matrices evaluadas, en este caso 32.

Ésto nos indica que la similitud entre ambas matrices está muy por encima de una distribución de matrices al azar, indicando que el mecanismo de generación de esta matriz automatizada produce un resultado parecido y robusto con respecto al original. En otras palabras, se puede concluir que la probabilidad de que la matriz de test del grafo *naive* haya sido similar producto del azar es muy baja, dado que se encuentra a más de dos veces del desvío con respecto a la media.

Sin embargo, a simple vista se observa que muchos de los elementos significativos de la

Atributos	SxMxC KruskallWallis	SxM Willcoxon	SxC Willcoxon	MxC Willcoxon
N	0.0132	0.0017	0.1234	0.2345
E	0.0128	0.0016	0.0985	0.2786
LCC	0.0017	0.0002	0.0058	0.0769
LSC	0.0464	0.0281	0.0789	0.3282
ATD	0.0213	0.0148	0.0207	0.7209
PE	0.0017	0.0002	0.0580	0.0769
L1	0.3166	1	0.4667	0.4667
L2	0.4941	0.2513	0.4104	0.9310
L3	0.2810	0.1958	0.7273	0.1935

Tab. 3.3: Tests sobre el grafo *naive*, los valores en rojo son p-values significativos ($p < 0{,}05$).

1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	1	0
1	1	1	0
1	0	0	1
0	0	0	0
1	1	0	0

trabajo original

1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	1	0
1	1	0	0
0	0	0	0
0	0	0	0
0	0	0	0

naive

Fig. 3.3: Matrices generadas a partir de tests trabajo original y tests sobre grafo *naive*.

matriz original no se encuentran presentes en la matriz generada de forma automática. Motivados por la transformación manual de los reportes en el trabajo original, automatizamos el proceso pre-procesando el reporte lematizando cada una de las palabras.

3.2.2. Grafo Lematizado

Basados particularmente en el proceso manual de creación de sintagmas propuesto en el trabajo original, proponemos construir un grafo que refleje el comportamiento capturado en la versión manual. Para ello, usamos la herramienta de lematización NLTK [17] que para esta funcionalidad NLTK usa *WordNet Lemmatizer*. Tomando el reporte original, usamos la herramienta en su versión más simple lo que significa que el lematizador solo lematizará las palabras que reconoce como sustantivos. La opción completa de lematización requiere contar con la información de POS-tagging de las oraciones. Dado que esta operación es costosa en performance y más compleja, decidimos para este procedimiento usar la versión más simple del lematizador.

A partir del texto lematizado, construimos el grafo de la misma manera que el grafo *naive*. Para el mismo reporte de un sujeto esquizofrénico 3.2 , obtenemos el grafo de la figura 3.5 donde podemos ver, por ejemplo, como la palabra *crying* es representada por el nodo con la etiqueta *cry*. En este caso, solamente cambio el nombre del nodo, pero para otros discursos puede juntar dos formas flexionadas en un solo nodo (ver Lematización en capítulo *Trabajos relacionados*).

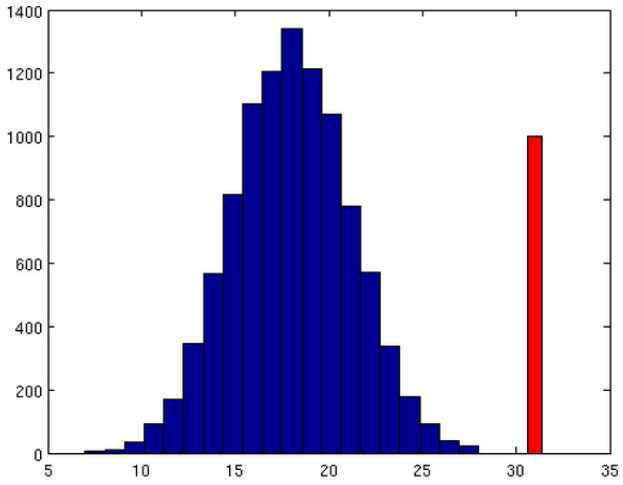

Fig. 3.4: En azul: histograma de similitud entre 100 mil matrices al azar con la matriz del trabajo original. En rojo: valor de similitud entre matriz trabajo original y la producida por el grafo *naive*.

Atributos	**SxMxC** KruskallWallis	**SxM** Willcoxon	**SxC** Willcoxon	**MxC** Willcoxon
N	0.0164	0.0022	0.1529	0.2345
E	0.0140	0.0019	0.1049	0.3282
LCC	0.0005	0.0002	0.1483	0.1674
LSC	0.0263	0.0095	0.1102	0.2073
ATD	0.0122	0.0047	0.0207	0.7209
PE	0.0050	0.0002	0.1483	0.1674
L1	0.0272	0.3016	0.2000	0.0256
L2	0.0358	0.0065	0.8959	0.0751
L3	0.1517	0.0455	0.7047	0.2583

Tab. 3.4: Test sobre grafo *lematizado*. En rojo tenemos los *p-values* significativos ($p<0.05$).

Nuevamente, podemos medir y obtener los parámetros del grafo, y posteriormente realizar los tests de hipótesis correspondientes. En la tabla 3.4 se muestran la significancia de los tests, indicando en rojo aquellos parámetros significativos. A diferencia del grafo *naive*, esta vez se observa que existe al menos un parámetro que separa a los grupos Maníacos y Control.

Entonces, nos preguntamos si esta transformación es significativamente mejor que la del grafo naive. Haciendo la misma transformación usada con el grafo *naive*, transformamos esta tabla en una matriz digitalizada y repetimos el procedimiento mencionado en la sección anterior.

En la figura 3.6 se muestra el resultado de la distribución de distancias con las matrices al azar y las matrices generadas a partir del grafo *naive* (barra roja) y el grafo lematizado (barra verde).

Nuevamente se repite el fenómeno, concluyendo que la probabilidad de que el grafo *lematizado* haya capturado las mismas singularidades que las del trabajo original por azar

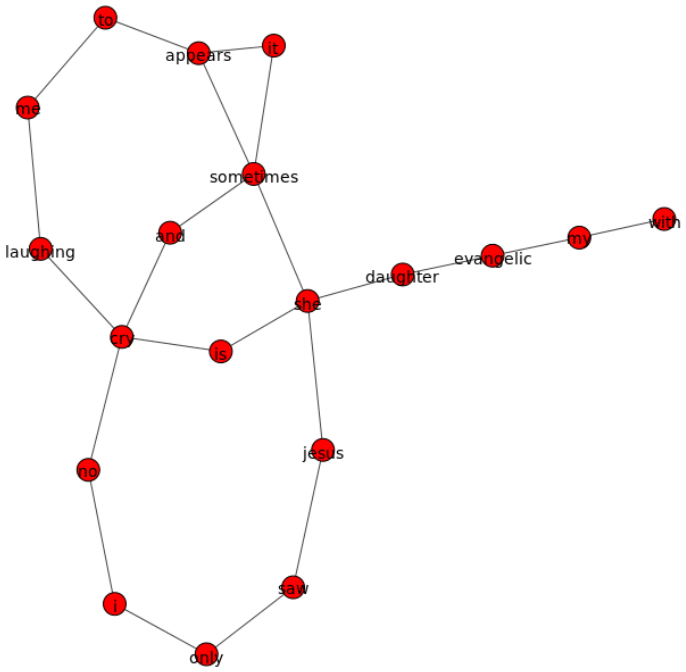

Fig. 3.5: Grafo *lematizado* inducido a a partir de reporte de sujeto esquizofrénico de la figura 3.2.

es ínfima, encontrándose nuevamente a más de dos veces del desvío con respecto a la media. Además, puede observarse que la similitud con respecto a la matriz original es aún mayor que la del grafo naive.

3.3. Clasificación

Habiendo verificado mediante tests que los métodos automáticos propuestos conservan las singularidades del método del trabajo original para esta muestra de datos, procedemos a probar la clasificación automática. Para ésto usamos el clasificador Naive Bayes (NB) [25] de Weka [8] (clasificador usado en el trabajo original). NB es un clasificador probabilístico basado en la aplicación del Teorema de Bayes al que se le suma la hipótesis de que las variables predictoras son independientes. Como tenemos los reportes del trabajo original (24 en total) vamos a testar con NB la performance de la clasificación con los mismos datos. La muestra es de 8 pacientes esquizofrénicos, 8 pacientes maníacos y 8 sujetos control. Usaremos cross-validation de *10 folds* para entrenar y testear el clasificador.

Cross-validation es una técnica de validación que consiste en tomar la muestra original, partirla en k *folds*, entrenar el algoritmo con $k-1$ sub grupos y testearlo con el restante. De esta forma se va separando cada uno y se testea con los demás, consiguiendo un valor de *rendimiento* para luego quedarse con la media de éstos.

Para poder corroborar que el clasificador no esta haciendo *overfitting* y también analizar la significancia de nuestro resultado, se procedió a generar 100000 muestras nuevas permutando las clases de los sujetos al azar, que fueron testeadas por el mismo algoritmo. Para el armado de estas muestras, se tomaron todos los features producidos por los algoritmos automáticos para cada sujeto (por ejemplo, $f_1, ..., f_6$) y para cada nueva muestra se eligió al azar 2 features (en el ejemplo, f_2 y f_6). Una vez seleccionados los features, se

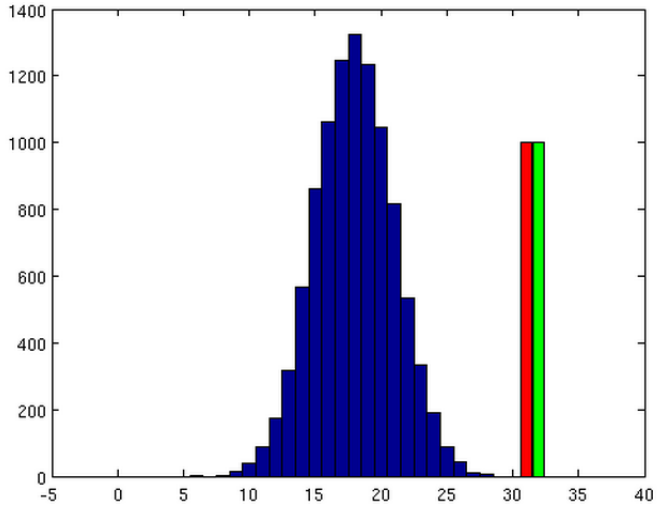

Fig. 3.6: En azul: histograma de similitud entre 100 mil matrices al azar con la matriz del trabajo original. En rojo: valor de similitud entre matriz trabajo original y la producida por el grafo *naive*. En verde: valor de similitud entre matriz trabajo original y la producida por el grafo *lematizado*.

procedió a permutar las clases del grupo al cual pertenecen esas muestras (conservando la cantidad de sujetos por grupo).

Para ilustrar este procedimiento, mostramos una tabla de sujetos (con su respectiva clase) y features, y una transformación al azar (figura 3.7). Estos datos no representan ningún dato real ya que fueron armados para el ejemplo.

Luego, cada muestra al azar es clasificada usando NB. Repitiendo este procedimiento 100000 veces, podemos estimar el promedio y el desvío estándar de la performance del algoritmo. De este modo, medimos la significancia del resultado real. Estos dos valores, la media y desvío, los denotamos como Random μ y Random *std* y los utilizaremos para evaluar cada una de las clasificaciones.

3.3.1. Clasificación de 3 grupos: Esquizofrénicos, Maníacos y Control

Para clasificar los relatos de los sujetos en las tres categorías, permitimos un máximo de 2 features para entrenar al clasificador y luego testear. En la figura 3.8 se observa la comparación de performance de clasificadores para la clasificación de 3 grupos (esquizofrénico, maníaco y control) simultáneamente. En primer lugar, comparamos la clasificación de los relatos entre el trabajo original y la técnica automatizada presentada. Podemos ver que la clasificación basada en los features que aporta el grafo *naive* tiene exactamente la misma performance que la clasificación del trabajo original. En este caso, el clasificador utilizó los parámetros ATD (promedio de grados en los nodos) y L2 (cantidad de loops de tamaño dos).

Para verificar si esta clasificación es significativa, comparamos contra la versión Random del clasificador, siendo μ un valor aproximado al valor que debería clasificar el azar. Teniendo en cuenta este valor y Random *std*, concluimos que el valor obtenido por el clasificador para este caso es significativo.

Para el clasificador al que se le permite utilizar los parámetros tanto del grafo *naive*

f1	f2	f3	f4	f5	f6	grupo
1.88	2.18	3.76	6.03	0.49	1.15	A
7.42	0.65	7.97	1.19	2.25	0.34	A
8.36	1.66	7.46	10.00	2.83	1.61	B
3.09	6.80	9.21	7.96	8.43	9.79	B
1.80	3.24	9.57	9.52	4.55	0.59	C
8.30	8.67	0.38	4.34	5.79	4.79	C

f2	f6	grupo
2.18	1.15	A
0.65	0.34	A
1.66	1.61	B
6.80	9.79	B
3.24	0.59	C
8.67	4.79	C

$\rightarrow$

f2	f6	grupo
2.18	1.15	C
0.65	0.34	B
1.66	1.61	B
6.80	9.79	A
3.24	0.59	A
8.67	4.79	C

Fig. 3.7: Ejemplo de procedimiento de generación de muestra al azar.

	SxMxC	Random μ	Random *std*
Trabajo Original	0.6250	-	-
Naive naive_ATD naive_L2	0.6250	0.3165	0.1111
Naive & Lematizado lematizado_ATD lematizado_L1	0.7500	0.3220	0.1172

Fig. 3.8: Tabla de comparación de performance de clasificadores para la clasificación de 3 grupos (esquizofrénico, maníaco y control) para el trabajo original, la automatización del grafo *naive* y *naive & lematizado*.

como *lematizado*, la performance sube a 0,75, mejorando el obtenido por el trabajo original. En este caso, los parámetros seleccionados son el promedio de grados en los nodos (ATD) del grafo lematizado y la cantidad de loops de tamaño 1 (L1) del mismo grafo. Nuevamente, comparando contra Random μ y *std*, se concluye que la clasificación es significativa.

A modo ilustrativo, mostramos las tablas 3.5 donde pueden verse las matrices de confusión. En la primera se observa que los esquizofrénicos son los grupos mejor clasificados (es decir, los sujetos esquizofrénicos son raramente clasificados en otra categoría), seguidos por los maníacos y por último los control. En la segunda, se repite el mismo orden.

3.3.2. Clasificación de 2 grupos: Esquizofrénicos y Maníacos

Para este clasificación tomamos hasta un máximo de 2 features para entrenar al clasificador y luego testear. En la figura 3.9 se observa la comparación de performance de clasificadores para la clasificación de 2 grupos (esquizofrénico y maníaco) simultáneamente. Podemos ver que la clasificación basada en los features que aporta el grafo naive tiene prácticamente el mismo rendimiento que la clasificación del trabajo original. Los features seleccionados para la clasificación fueron la cantidad de nodos y la cantidad de ejes paralelos. Es interesante notar que el Random μ es un valor cercano al valor que debería

maníaco	control	esquizofrénico	$<-$ clasificados como
5	1	2	maníaco
3	3	2	control
1	0	7	esquizofrénico

maníaco	control	esquizofrénico	$<-$ clasificados como
5	2	1	maníaco
1	6	1	control
1	0	7	esquizofrénico

Tab. 3.5: Matriz de confusión para la clasificación de 3 grupos (esquizofrénico, maníaco y control) usando NB, (arriba) features naive_ATD y naive_L2 y (abajo) features lematizado_ATD y lematizado_L1.

	SxM	Random μ	Random *std*
Trabajo Original	0.9380	-	-
Naive naive_Nodes naive_PE	0.9375	0.4788	0.1675
Naive & Lematizado lematizado_ATD naive_PE	1.0000	0.4766	0.1592

Fig. 3.9: Tabla de comparación de performance de clasificadores para la clasificación de 2 grupos (esquizofrénico y maníaco) para el trabajo original, la automatización del grafo *naive* y *naive & lematizado.*

clasificar tomando muestras al azar. Teniendo en cuenta el Random *std* podemos ver que el valor obtenido por el clasificador para este caso es significativo.

Para el caso del clasificador que usa tanto las features de los grafos *naive* como los del grafo *lematizado*, la clasificación es perfecta, superando el resultado original. En este caso, los features seleccionados son obtenidos a partir de ambos grafos: la cantidad de ejes paralelos (PE) del grafo *naive* y el promedio de grados de los nodos (ATD) del grafo lematizado.

3.3.3. Clasificación de 2 grupos: Esquizofrénicos y Control

Para este clasificación tomamos hasta un máximo de 2 features para entrenar al clasificador y luego testear. En la figura 3.10 se observa la comparación de performance de clasificadores para la clasificación de 2 grupos (esquizofrénico y control) simultáneamente. Podemos ver que la clasificación basada en los features que aporta el grafo naive tiene exactamente la misma performance que la clasificación del trabajo original. Para clasificar usamos el promedio del grado de los nodos (ATD) y la cantidad de loop de tamaño 1 (L1). También podemos ver que el Random μ es un valor cercano al valor que debería clasificar el azar. Teniendo en cuenta el Random *std* podemos ver que el valor obtenido por el clasificador para este caso es significativo. Para el caso del clasificador que usa las features de los grafos *naive & lematizado*, la performance se mantiene igual, no aportando mejora. En este caso usamos la cantidad de loops de tamaño 1 (L1) del grafo lematizado y el promedio del grado de los nodos (ATD) del grafo *naive*. También se puede ver que es un dato significativo pues nuevamente la comparación con Random μ y el desvío así lo

manifiestan.

	SxC	Random μ	Random *std*
Trabajo Original	0.8750	-	-
Naive naive_ATD naive_L1	0.8750	0.4946	0.1334
Naive & Lematizado lematizado_L1 naive_ATD	0.8750	0.4855	0.1281

Fig. 3.10: Tabla de comparación de performance de clasificadores para la clasificación de 2 grupos (esquizofrénico y control) para el trabajo original, la automatización del grafo *naive* y *naive & lematizado*.

3.3.4. Clasificación de 2 grupos: Maníacos y Control

Para este clasificación tomamos hasta un máximo de 2 features para entrenar al clasificador y luego testear. En la figura 3.11 se observa la comparación de performance de clasificadores para la clasificación de 2 grupos (maníaco y control) simultáneamente. Podemos ver que la clasificación basada en los features que aporta el grafo naive tiene una performance similar que la clasificación del trabajo original, a penas 0.0005 por debajo. En este caso usamos como features, la cantidad de loops de tamaño 1 (L1) y la cantidad de loops de tamaño 2 (L2). Sin embargo, en este caso podemos ver que Random μ (que es un valor cercano al valor que debería clasificar el azar) más el Random *std* es un poco menor a la performance de nuestro clasificador por lo que concluimos que en este caso la clasificación resulta ser a penas significativa.

Para el caso del clasificador que usa las features de los grafos *naive & lematizado*, la performance mejora considerablemente, llegando a 0.8125. En este caso, basado en el análisis de Random μ y el Random *std*, podemos concluir que esta clasificación es significativa. Para realizar esta clasificación usamos la cantidad de loops de tamaño 1 (L1) del grafo lematizado y la cantidad de ejes paralelos (PE) del grafo naive.

	MxC	Random μ	Random *std*
Trabajo Original	0.6880	-	-
Naive naive_L1 naive_L2	0.6875	0.4946	0.1599
Naive & Lematizado lematizado_L1 naive_PE	0.8125	0.4636	0.1598

Fig. 3.11: Tabla de comparación de performance de clasificadores para la clasificación de 2 grupos (maníacos y control) para el trabajo original, la automatización del grafo *naive* y *naive & lematizado*.

3.4. Extensión Morfosintáctica

Basados en la hipótesis cognitiva de las conclusiones de Mota et. al [7]:

Overall, the results point to automated psychiatric diagnosis based not on what is said, but on how it is said.

buscamos técnicas que nos permitieran modelar estas características y abstraernos del significado de las palabras. Por lo tanto, nos propusimos usar la función que cumple una palabra en una determinada frase. Para ésto tenemos la herramienta POS-tagging a partir de la cual proponemos el grafo *morfosintáctico*.

3.4.1. Grafo Morfosintáctico

Como mencionamos, motivados por la hipótesis cognitiva de Mota et. al. proponemos generar un nuevo grafo que capture nueva información. Para generar este grafo, usamos la herramienta de Part-of-speech tagging de NLTK [17]. Lo primero que hicimos fue separar el reporte en oraciones para luego evaluar cada una con la herramienta de POS-tagging generando una secuencia de *tags*. A partir de este resultado, generamos la transformación del reporte original a una secuencia de *tags* con la que armamos el grafo como lo hacíamos en los casos anteriores, es decir cada *tags* corresponde a un nodo y un par de tags consecutivos de la secuencia generará una arista entre dos nodos.

Por ejemplo, tomando el relato de un paciente esquizofrénico (3.2), del proceso de POS-Tag se obtiene la siguiente tabla de *tags* 3.6:

Palabra	**Tag**
With	IN
my	PRP$
evangelic	JJ
daughter	NN
She	PRP
is	VBZ
crying	VBG
No	DT
I	PRP
only	RB
saw	VBD
Jesus	NNP

Palabra	**Tag**
She	NNP
sometimes	RB
it	PRP
appears	VBZ
to	TO
me	PRP
laughing	VBG
crying	VBG
and	CC
sometimes	RB
appears	VBZ

Tab. 3.6: Tabla de POS-tagging inducida de reporte de la figura 3.2.

Por lo que para el reporte anterior, obtenemos el texto de la figura 3.12 y su grafo inducido donde se puede notar la reducción en cantidad de nodos y cómo cambia la morfología del grafo en comparación a los anteriores.

3.4.2. Clasificación

A partir de los tres grafos y las medidas tomadas sobre éstos, podemos constituir una caracterización cuantitativa más basta de un mismo relato por lo que nos disponemos a usar estos valores para clasificar los tres grupos y comparar nuestro desempeño con el del trabajo original. Consideramos dos opciones: clasificación utilizando un máximo de 2 parámetros y otras clasificación utilizando 3 parámetros.

En el primer caso, utilizando 2 *features*, observamos que no hay ganancia agregando el nuevo grafo con respecto a los resultados anteriores. En 3.13 podemos ver que para

IN PRP$ JJ NN PRP VBZ VBG DT PRP RB VBD NNP NNP RB PRP VBZ TO PRP VBG VBG CC RB VBZ

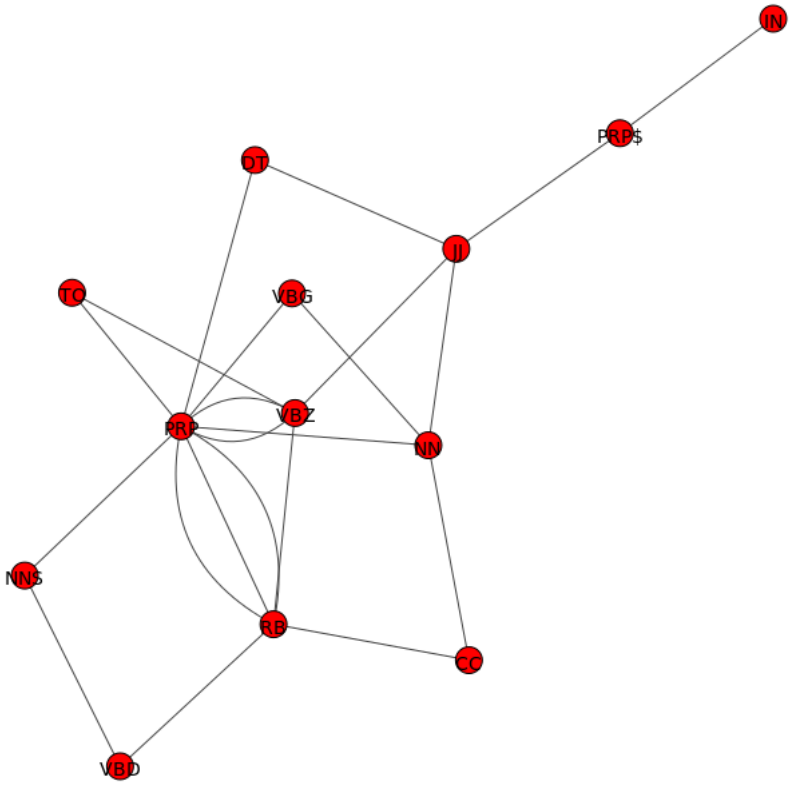

Fig. 3.12: Relato transformado a texto POS-tagging para armar grafo y su grafo *morfosintáctico* inducido a partir de reporte de sujeto esquizofrénico.

la comparación usando hasta dos features no encontramos diferencia alguna en la performance tras haber agregado los features del grafo *Morfosintáctico*, dado que sigue siendo mejor medida para combinar hasta dos, usando solo features del grafo lematizado. En este caso, las features usadas son las mismas que expresamos anteriormente: el promedio de grados en los nodos (ATD) del grafo lematizado y la cantidad de loops de tamaño 1 (L1) del mismo grafo. Seguimos teniendo una performance de 0.75 significativa.

En cambio, al liberar el clasificador para la utilización de tres parámetros, vemos un incremento en el rendimiento. En 3.13 podemos ver que para la comparación usando hasta tres features, sí aporta una mejora en la performance de la clasificación. Haciendo el análisis del Random μ y su desvío, vemos que la performance en la clasificación es significativa. En este caso se da uso de un atributo del grafo *morfosintáctico* dado que usamos los features promedio de grados en los nodos (ATD) del grafo lematizado, la cantidad de loops de tamaño 1 (L1) del mismo grafo y el promedio de grados en los nodos (ATD) del grafo *morfosintáctico*. En la tabla 3.14 se puede apreciar la matriz de confusión que comparándola con la obtenida anteriormente (tabla 3.5) podemos ver que la clasificación para el grupo esquizofrénico mejora.

	SxMxC	Random μ	Random *std*
Trabajo Original	0.6250	-	-
lematizado_ATD, lematizado_L1	0.7500	0.3168	0.1167
lematizado_ATD, lematizado_L1, sintactico_ATD	0.7917	0.3224	0.1168

Fig. 3.13: Tabla de comparación de performance de clasificadores para la clasificación de 3 grupos (esquizofrénico, maníaco y control) para el trabajo original, la automatización del grafo *naive*, *lematizado* y *morfosintáctico*.

maníaco	control	esquizofrénico	$<-$ clasificados como
5	2	1	maníaco
1	6	1	control
0	0	8	esquizofrénico

Fig. 3.14: Matriz de confusión para la clasificación de 3 grupos (esquizofrénico, maníaco y control) usando NB, lematizado_ATD, lematizado_L1 y sintactico_ATD.

3.5. Conclusiones

A partir de la evidencia recolectada en las secciones Grafo *Naive* y Grafo *Lematizado* podemos decir que el proceso de automatización conserva las singularidades del trabajo previo hecho por Mota et. al. [7], tras haber mostrado con éxito que las dos *tablas de tests* preservan su *forma* y singularidad con respecto a las del trabajo original. Esta observación está basada en la generación de las matrices al azar y en como las dos matrices de tests (producto de las *tablas de tests*) son claramente *outliers* en la distribución de las matrices al azar.

Una prueba aún más fuerte de que la automatización es un buen camino se presenta en la sección de clasificación 3.3. En esta sección vemos como los procesos automatizados usando tecnologías del procesamiento del lenguaje natural tienen un buen rendimiento en la detección de particularidades en el estudio del discurso. La clasificación presentada funciona tan bien como la versión manual del trabajo original y sobre todo, tiene un excelente consenso con la clasificación clínica hecha por especialistas médicos. Este resultado fue obtenido tanto si se la usa para clasificar entre grupos de dos, como si se la usa para clasificar a los tres grupos por separado. En este último caso, usando el mismo clasificador que el trabajo original con tres features se llega a una performance de 0.7917 usando la información del grafo *lematizado* y el grafo *morfosintáctico*. Para el caso de las clasificaciones de a pares estamos en una desempeño de: Maníacos vs Control 0.8125, Esquizofrénicos vs Control 0.8750, Esquizofrénicos vs Maníacos 1.

La extensión a otro tipo de extracción de información, como es el caso de los *features* que genera el grafo *morfosintáctico*, parece ser fructífera e invita a indagar más en el uso de técnicas automáticas de procesamiento del lenguaje natural que recauden información interesante. Sin embargo, creemos que resulta relevante la comprensión de las hipótesis cognitivas ya que direccionan la búsqueda de la información pertinente. Ésto se evidencia en el caso del grafo *morfosintáctico*, el cual surge de la motivación de satisfacer el resultado cognitivo del trabajo original de Mota et al.: *La particularidad en los discursos está en*

la abstracción del cómo se dice un discurso y no de qué habla un discurso.

Quedando muchas direcciones por las cuales seguir buscando particularidades del discurso creemos que las herramientas del procesamiento del lenguaje natural pueden capturar mucha información relevante que caracterice el estado mental del sujeto que genera los textos.

4. CASO DE ESTUDIO: ECSTASY

El Ecstasy (MDMA) es una droga psicoactiva, alucinógena, cuyos efectos principales son la euforia, aumento de la sociabilidad, la energía física y emocional. Los efectos de esta droga incluyen patrones similares a los de esquizofrenia y psicosis como por ejemplo alucinaciones visuales, problemas de memoria o comportamiento violento descontrolado. Al igual que en el caso del estudio anterior, nos proponemos analizar relatos de sujetos ante la ingesta de ecstasy en búsqueda de patrones estereotipados del pensamiento y su modificación con la ingesta de MDMA.

En el New York State Psychiatric Institute, Department of Psychiatry, Columbia University de New York se realizaron –hace alrededor de 10 años– entrevistas a 13 sujetos voluntarios. Estas entrevistas consistieron en una simple consigna: hablar sobre un ser cercano. Cada sujeto fue sometido a cuatro sesiones. En cada una de las sesiones, el sujeto debió consumir una de cuatro posibles pastillas: Placebo, MDMA dosis baja (0,75mg), MDMA dosis alta (1,5mg) o Metanfetamina (20mg) (Meth). Para cada sujeto, el orden en el que se consumen las distintas pastillas es al azar y en ningún momento sabe que pastilla le tocó en cada sesión, con el objetivo de no condicionarlo. Tras la ingesta se realiza la entrevista de aproximadamente 10 minutos con el paciente, dirigida por un psiquiatra. A partir de la transcripción de estas entrevistas es posible analizar cambios en los relatos dependiendo de la pastilla ingerida.

4.1. Análisis preliminar

Como primer iteración del análisis en búsqueda de regularidades en los textos –inspirados en la descripción del cambio de ánimo, euforia, y otros efectos mencionados– nos preguntamos cuáles eran los temas sobre los que hablaban los relatos. Para ellos, decidimos calcular la similitud de un texto a distintos conceptos mediante el uso de técnicas para la estimación de similitud semántica entre palabras. En particular, para medir la distancia (o similitud) semántica que hay entre dos palabras usamos Latent Semantic Analysis, entrenado con el corpus TASA (ver la sección de *Trabajo Relacionado*). Estimar la similitud de un relato a un cierto concepto fue aproximado calculando la media de la similitud de las palabras del relato a la palabra que representa el concepto seleccionado. Formalmente,

$$S(r, c) = \frac{\sum_{i=1}^{N(r)} \cos(W_i - c)}{N(r)}$$

donde $N(r)$ es la cantidad de palabras del relato r, W_i es la representación LSA de la i-ésima palabra del relato y c el valor LSA de la palabra que representa el concepto para el que se está midiendo la similitud.

Una vez definida esta medida, es necesario elegir los conceptos contra qué comparar. Siguiendo el articulo de Wikipedia sobre Ecstasy[1] se eligieron palabras relacionadas con estados emocionales, por ejemplo *friend*, *love*, entre otras. En la figura 4.1 se pueden ver resultados de estas mediciones.

[1] `http://en.wikipedia.org/wiki/Ecstasy`

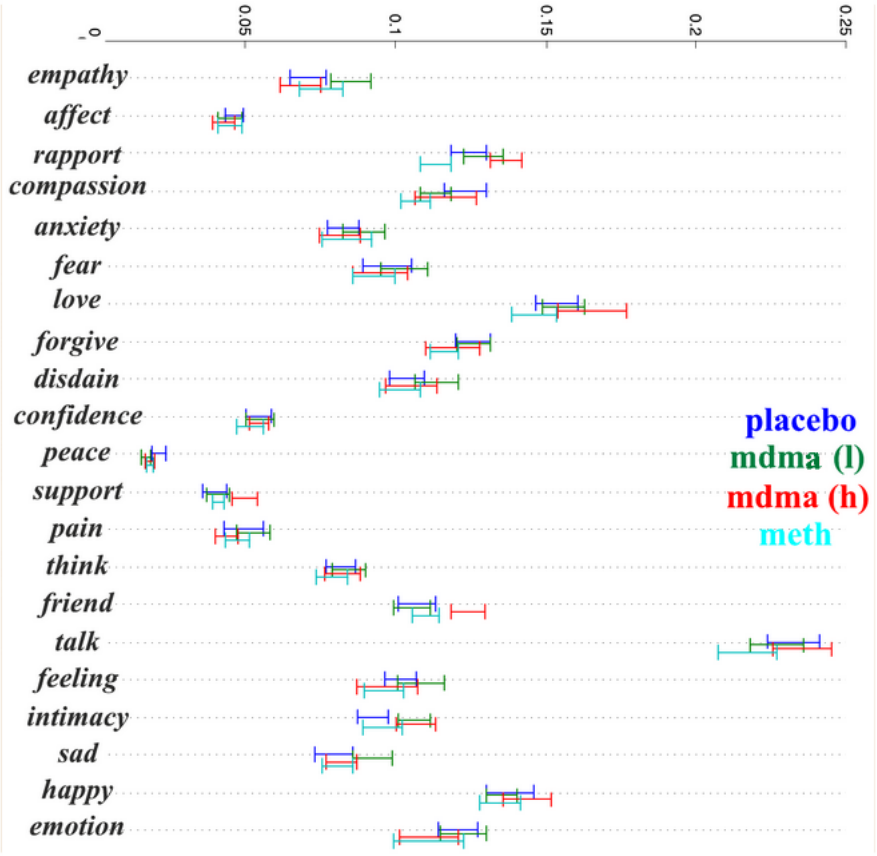

Fig. 4.1: Distancias promedio a palabras elegidas a mano contra discursos de sujetos agrupados por condición (placebo, MDMA(l), MDMA(h) y Metanfetamina). Cada vector representa la media y el estándar error de cada grupo.

Se puede apreciar que varias de las palabras seleccionadas, permiten separar a algunos grupos. El caso más sobresaliente es *friend*, donde la distancia al grupo MDMA (dosis alta) es estrictamente más alta que el resto de los grupos.

Si bien podríamos utilizar esta estrategia para discriminar entre grupos, ésta depende del procedimiento manual de selección de palabras. Al igual que en el caso de detección de psicosis y esquizofrenia del capítulo anterior, nos interesa el desarrollo de técnicas sin supervisión para la detección de estos marcadores. Por lo tanto, a continuación nos concentramos en buscar técnicas generales que permitan detectar patrones del léxico para encontrar grupos discriminantes.

4.2. Buscando grupos discriminantes

En el análisis anterior habíamos calculado la distancia de los relatos a palabras elegidas arbitrariamente. En esta sección buscamos un mecanismo general para detectar un conjunto de palabras que podrían discriminar a las 4 condiciones (placebo, MDMA low, MDMA high y Meth). Para encontrarlas, inicialmente decidimos buscar en todas las palabras de TASA que comprende más de 78000 palabras distintas. Sin embargo, para disminuir la

cantidad de palabras a analizar decidimos quedarnos únicamente aquellas palabras que tuvieran una definición formal en el diccionario. Debido al origen de los textos de este corpus, existen palabras como onomatopeyas y otras palabras sin significado. Para descartar las palabras, decidimos buscar un diccionario tradicional y quedarnos con todas las palabras de TASA que estuvieran definidas en este diccionario; en este caso, elegimos Wordnet [26]. Aplicando este filtrado, quedaron aproximadamente 35000 palabras posibles.

Al igual que como hicimos en el análisis anterior, tomamos las distancias de los relatos a todas las palabras de TASA (intersección Wordnet). Esto nos da como resultado un número entre 0 y 1 de cada una de las 35000 palabras. En trabajos pasados sobre LSA TASA [27] sugieren el uso de un umbral – digitalizando las distancias – como técnica para limpiar el ruido. La idea convertir las distancias mayores o iguales a 0.1 a 1 y las menores a 0. El motivo de ésto es que la gran mayoría de los valores de las distancias en LSA TASA son muy chicos, por lo que a través del umbral se potencian aquellos valores que sobresalen sobre la enorme mayoría de número pequeños. Si bien esta técnica propone una perdida de granularidad en los datos, muestra mejores resultados a la hora de comparar similitud de conceptos.

Teniendo en cuenta este proceso, por cada relato tomamos cada una de sus palabras y calculamos la similitud en LSA contra todas las del corpus, para luego digitalizar usando el umbral mencionado. Generamos así un vector de dimensión 35000 por cada palabra del relato. De este modo, un relato de n palabras, generó una matriz de 35000 filas por n columnas, con 0 o 1 en cada elemento. A partir de esta matriz, tomamos el promedio por filas obteniendo así una matriz de 35000 por 1, donde la fila i representa la distancia entre la palabra i-ésima de TASA WordNet y el relato.

Como siguiente paso, armamos una matriz (detallada en la figura 4.2) que contiene por columna cada sujeto y sus cuatro condiciones (ordenadas placebo, mdma low, mdma high y Meth), es decir cuatro veces 35000 distancias a las palabras mencionadas anteriormente. El resultado obtenido fue una matriz de 140000 distancias a palabras por 13 sujetos.

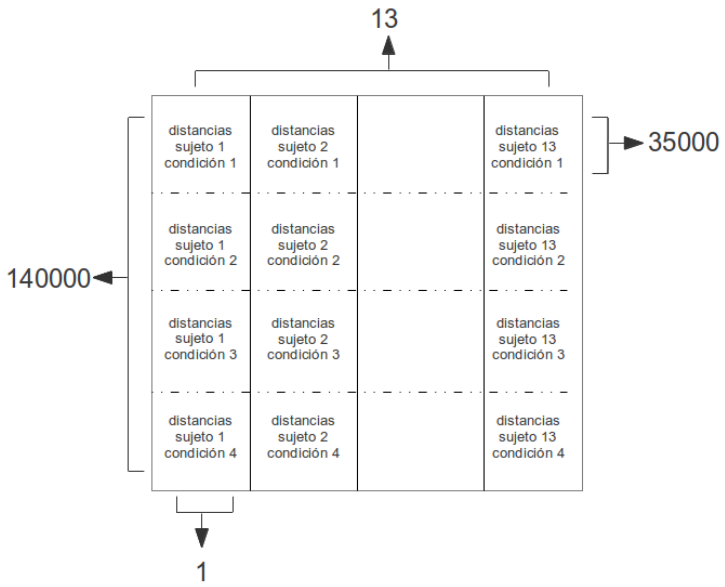

Fig. 4.2: Matriz de distancias ordenada por sujetos y condición de sesión.

Analizar la matriz mencionada y encontrar regularidades resulta muy complejo dado

su tamaño por lo que empleamos Singular Value Descomposition (SVD) como técnica de reducción de dimensionalidad. Usamos esta técnica porque si existiera dependencia entre las filas de la matriz, este método factoriza preservando esta particularidad.

Sea D la matriz expuesta en la figura 4.2 de dimensiones de 140000×13, entonces

$$SVD(D) = U.S.V$$

donde U es de 140000×140000, S es de 140000×13 y V de 13×13. Dado que es una factorización, U puede ser pensada como componente de una nueva base y $S.V$ como los pesos. A partir de este nuevo esquema podemos analizar las componentes de U, es decir, sus columnas.

En la figura 4.3 podemos ver una ilustración de las primeras 6 componentes de U. Recordar que en SVD las componentes están ordenadas de mayor a menor según cuánto

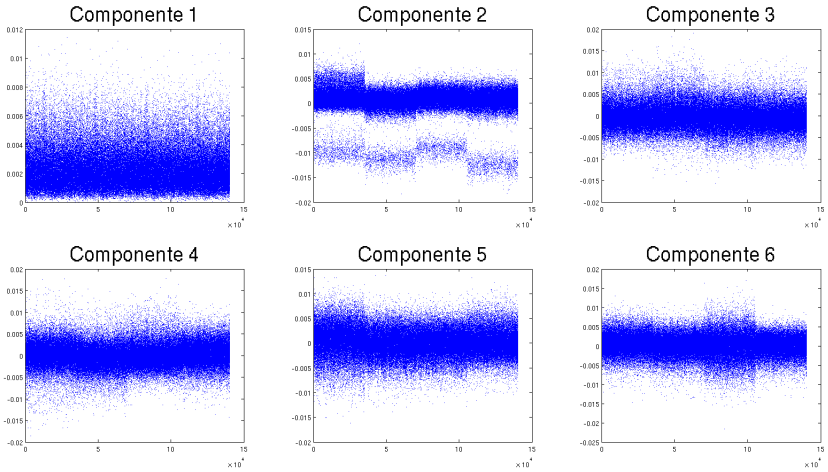

Fig. 4.3: Esquema de las 6 primeras componentes de U (matriz de factorización de SVD sobre D 4.2). A simple vista se puede observar una particularidad en la segunda componente (segunda de arriba).

explica los datos ya factorizados. A simple vista se ve que la segunda componente tiene una particularidad en su nube de puntos, viéndose una separación de los puntos en dos grupos (ver detalle en la figura 4.4) Para poder separar entre ambos grupos analizamos el histograma de esta componente (ver figura 4.4, panel derecho) y encontramos un mínimo en -0.0055, indicándonos el valor para el corte que separa los grupos de puntos.

Tomando solo los puntos menores a -0.0055 encontramos aproximadamente 1300 puntos para cada grupo de 35000 palabras referidas a cada condición. Podríamos haber encontrado distintas cantidad de puntos para cada condición pero, sorprendentemente, para las cuatro condiciones fueron del orden de 1300. Más sorprendente aún, resulta que si tomamos la intersección de los 1300 puntos para cada condición tenemos una intersección de 1200. Tomando la unión de los puntos de las cuatro condiciones obtenemos 1326. Es decir que, a través de la componente obtenida mediante SVD hemos identificado un grupo de palabras que, en principio, podrían servir para discriminar entre las condiciones. A estas palabras las llamamos *palabras emergentes*. En la sección *Trabajo Futuro* hacemos un breve análisis de la particularidad de este conjunto de palabras.

A continuación seguimos el análisis concentrándonos en estas palabras. En primer lugar, hicimos la proyección de la matriz D con las *palabras emergentes*, obteniendo D'

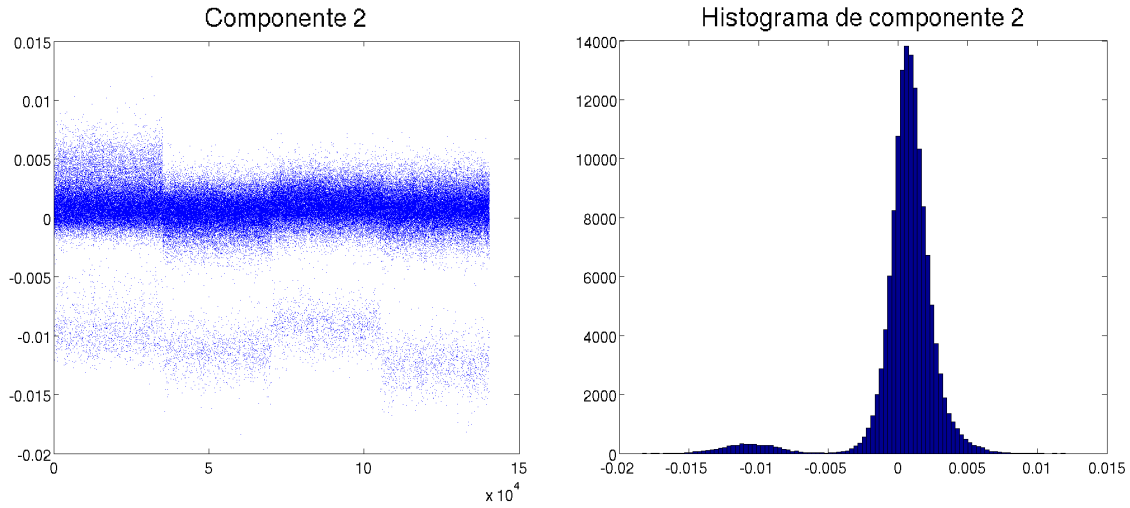

Fig. 4.4: Segunda componente de SVD (izq). Se observan la separación en 2 de la nube de puntos. En la nube de puntos inferior se ve un indicio de separación en las condiciones. (der) Histograma de valores de los puntos del panel izquierdo, con un mínimo en -0.0055.

que se puede ver en la figura 4.5

Factorizamos nuevamente con SVD, esta vez, a la matriz D' (factorización que llamamos SVD1326). Nuevamente, se ilustran las primeras 6 componentes en la figura 4.6, donde podemos ver que algunas de las mismas permiten separar entre las distintas condiciones. Por ejemplo, para la segunda componente se ve una gran diferencia entre la cuarta condición (Meth) y las demás (reflejado en la altura). En cambio, para la tercer componente se evidencia una mayor diferencia entre todas las condiciones, exceptuando el caso de la comparación entre la segunda condición (MDMA low) y cuarta (Meth) que se ven aproximadamente a la misma altura.

Mirando más detalladamente la tercer componente (figura 4.7), observamos que ésta permite diferenciar entre las condiciones, en 4 grupos de 1326 palabras. Sin embargo, esta separación no es clara entre las condiciones 2 y 4 (panel izquierdo de la figura 4.8). En cambio, estas condiciones se ven bien diferenciadas en la componente 5. Combinando ambas componentes, por lo tanto, podemos establecer un criterio que permite diferenciar los grupos (figura 4.8, panel derecho).

Dado que todos los grupos son distinguibles usando las componentes 3 o 5 podemos ver qué resulta de separar los cuatro grupos a la vez usando estas dos componentes. Al analizar la figura 4.9 vemos a cada condición como una nube de puntos de un color en particular. Cada punto representa una palabra para una condición (1326 por condición) proyectada en la componente 3 (eje x) y en la componente 5 (eje y). Se puede ver cómo las 4 condiciones están notablemente separadas. Para el caso de las condiciones placebo y MDMA (high) vemos que las nubes de puntos son prácticamente disjuntas. En el resto de los casos, podemos identificar sectores privilegiados para cada condición, con muy poco solapamiento entre distintos grupos.

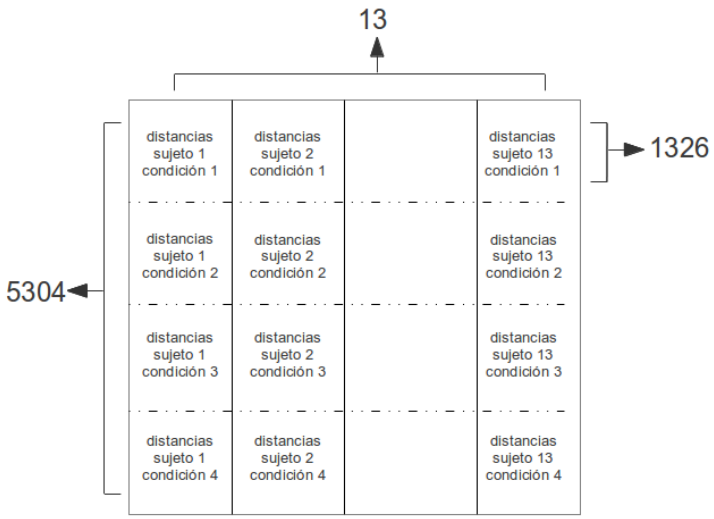

Fig. 4.5: Matriz de distancias ordenada por sujetos y condición de sesión filtrada por 1326 palabras emergentes.

4.3. Identificación

A partir de la propuesta de factorizar la matriz D encontramos una serie de palabras para una componente que separaban en 2 a los puntos. Tomándolas armamos D' reduciendo mucho la dimensionalidad de D. Al factorizar D' con SVD obtuvimos diferentes componentes que separaron a los puntos según la condición a la que éstos pertenecían. Como consecuencia de este mecanismo automático, en este momento nos preguntamos cómo podríamos usar este mecanismo de separación de condiciones para clasificar nuevas muestras.

Para analizar el método y sus características que emergen de este análisis decidimos proponer el siguiente esquema: apartar a un sujeto (con sus 4 condiciones) de los 13, confeccionar el mismo análisis anterior para los 12 restantes y verificar con el sujeto apartado si éste es consistente a los otros 12, tanto en su separación como en la integración a las condiciones de los otros sujetos. Luego, repetir lo mismo con todos los demás. Lo que primero intentamos analizar es cuan separables son las condiciones de una nueva muestra, más allá de intentar identificar a qué clase corresponde cada condición.

4.3.1. Separar condiciones

El proceso para resolver esta cuestión consiste primero en armar una matriz como D' pero dejando un sujeto afuera, es decir con una columna menos (llamada D_{12}). Factorizando con SVD a esta matriz y repitiendo el procedimiento anterior generamos una descomposición:

$$SVD(D_{12}) = U_{12}.S_{12}.V_{12}$$

con U_{12} de 5304×5304, S_{12} de 5304×12 y V_{12} de 12×12.

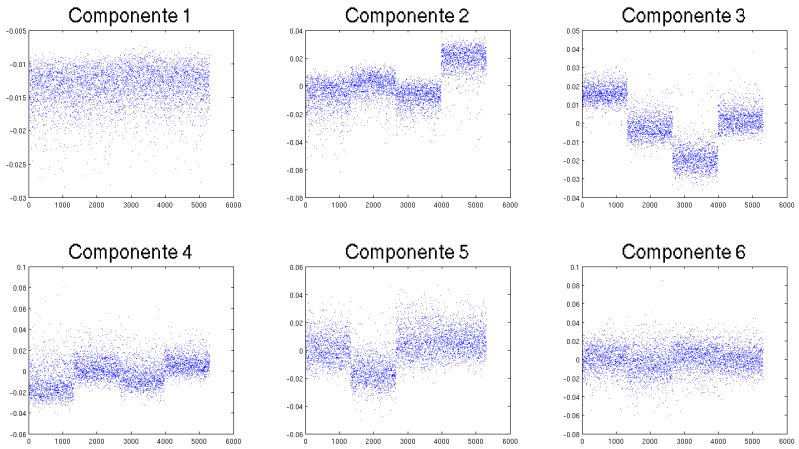

Fig. 4.6: Esquema de las 6 primeras componentes de U (matriz de factorización de SVD sobre D' 4.5). A simple vista se puede observar que algunas componentes presentan separación para las distintas condiciones y otras parecen no presentar diferencias.

Podemos pensar al sujeto que sacamos como una nueva muestra y buscar su proyección en la nueva base. Lo llamamos Y, de 5304×1. Por ejemplo, si quisiéramos calcular la proyección de Y para la componente 3, llamémosla Y_{p3}. Hacemos:

Sea $comp_3$ la tercer columna de U transpuesta, osea $comp_3$ es de 5304×1.

$$Y_{p3} = Y \ . \quad PINV(\ S(3,3)\ .\ comp_3) \quad .\ (\ S(3,3)\ .\ comp_3)$$

donde $PINV$ es la pseudo inversa. Una vez hecho ésto, contamos con la proyección de la nueva muestra Y en la base producida por hacer SVD en D_{12}

Teniendo la proyección de la nueva muestra, necesitamos encontrar una forma de constatar que existe una separación de las condiciones. En la figura 4.10 se ilustra un ejemplo del gráfico que se produce sacando el sujeto 1. En la parte de arriba en azul se ven las componentes de la factorización de todos los sujetos salvo el 1 de las componentes 2, 3 y 4. En verde, el sujeto apartado, mismas componentes. Abajo se ven los boxplots de las distribuciones por condición del sujeto 1 según cada componente. Se puede ver a simple vista que para el sujeto 1 (los puntos en verde), para las componentes 2 y 3 (primera y segunda de arriba) hay una diferencia en la proyección según la condición que se esté viendo, para el caso de la componente 4, no se observa. Sin embargo, la diferencia entre condiciones podría detectarse analizando la distribuciones de los datos, por ejemplo, en las desviaciones.

Los boxplots de la figura 4.10 muestran como son las distribuciones de los puntos verdes (las proyecciones del sujeto apartado) según cada componente. A simple vista se pueden ver diferencias significativas para algunos de los grupos. Por ejemplo comparando las condiciones Placebo y MDMA high (es decir primera y tercera) en la componente 3 vemos que los boxplots son disjuntos y cuantificando esta diferencia, obtenemos significancia al hacer tests (Wilcoxon, $p = 0$).

Haciendo el mismo análisis con los otros sujetos observamos que hay una tendencia a que las componentes que separan las condiciones para los 12 sujetos de entrenamiento, también separan las condiciones de los sujetos nuevos.

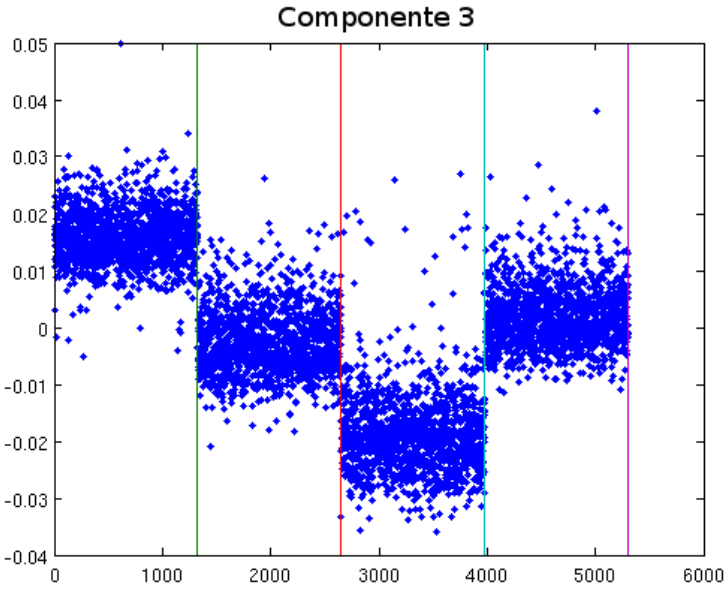

Fig. 4.7: Componente 3 de SVD1326. Las lineas de colores separan las 1326 medidas para cada condición (de izquierda a derecha: Placebo, MDMA low, MDMA high, Meth).

4.3.2. Clasificación

Teniendo una matriz de *entrenamiento* logramos, en la subsección anterior, tomar una nueva muestra, proyectarla y constatar de que podíamos seguir viendo la separación de condiciones por lo que la factorización propuesta y el hecho de sólo evaluar las *palabras emergentes* parecía haber capturado la información necesaria para separar a las muestras.

Como siguiente paso, dado una muestra de un nuevo sujeto con 4 condiciones, nos gustaría poder identificar a qué condición corresponde cada relato. En la factorización anterior, vimos que tenemos diferentes componentes que logran separar distintas condiciones, algunas con más énfasis que otras. Es por eso, que simplificando el problema, nos abocamos a tratar de clasificar sólo dos condiciones radicalmente opuestas: Placebo y MDMA high.

En la primera iteración lo que hicimos fue tomar la matriz D' y observar qué sucede pensándola de a condiciones. Tomamos las primeras 1326 correspondientes a la primera condición y contamos cuántos valores mayores a 0.1 tenían, agrupándolo por sujeto. Repitiendo ésto por todas las condiciones obtuvimos una matriz de sujetos por condición que nos dice cuántos valores cercanos de palabras tenían. El problema de tomar esta medida como informativa es que tiene mucha varianza, dado que los sujetos son independientes entre sí. Por esta razón, pensamos cómo normalizar a cada sujeto, esto quiere decir, que la diferencia entre las condiciones podría existir pero sería relativa a cada sujeto. Decidimos normalizar los valores mencionados por la suma de ellos.

En las tablas 4.1 podemos ver a la izquierda los valores sin normalizar y a la derecha los valores normalizados observándose que los de la tabla izquierda tiene un diferencia importante ya que tenemos sujetos con 1306 palabras cercanas (de 1326 posibles) y otros con solo 47, para los reportes en condición placebo. Viendo la tabla normalizada, los valores son más coherentes entre sujetos por lo que la utilizamos como fuente y usamos

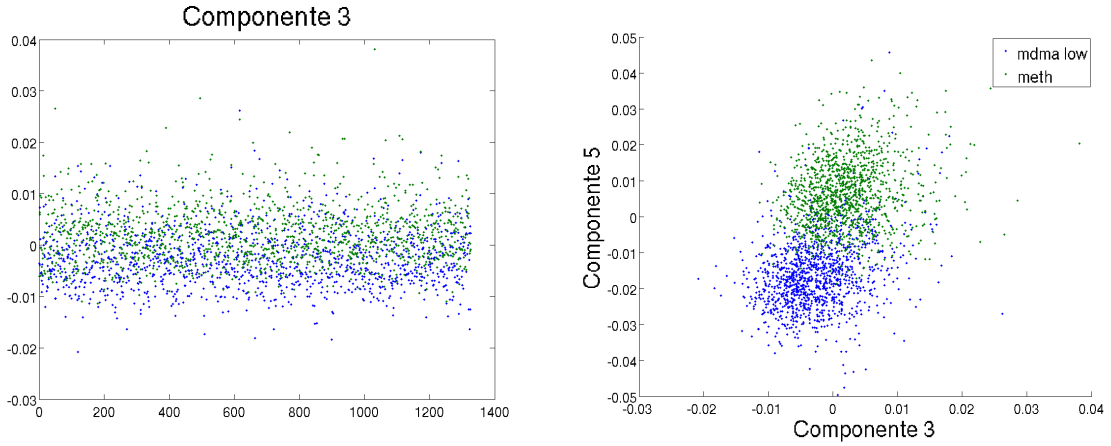

Fig. 4.8: MDMA low vs Meth. A la izquierda componente 3 de SVD1326. A la derecha componente 3 vs componente 5 de SVD1326.

Naive Bayes con 10 *folds* para ver cuan bien clasificaba esta información. Obtuvimos 84.6154 % de performance. Pero en los datos de entrenamiento existen pares de valores que pertenecen al mismo sujeto, lo que va en contra de las hipótesis del clasificador lo cual nos impide afirmar que la clasificación es significativamente buena dado que no queda en claro la independencia muestral.

Placebo	MDMA high
186	112
193	142
679	249
85	48
1306	1044
1084	922
277	180
822	84
282	1005
403	628
47	21
500	459
245	117

Placebo	MDMA high
0.6242	0.3758
0.5761	0.4239
0.7317	0.2683
0.6391	0.3609
0.5557	0.4443
0.5404	0.4596
0.6061	0.3939
0.9073	0.0927
0.2191	0.7809
0.3909	0.6091
0.6912	0.3088
0.5214	0.4786
0.6768	0.3232

Tab. 4.1: A la izquierda se muestra la tabla con la cantidad de palabras cercanas según la condición. A la derecha la misma tabla normalizada con la suma de los valores.

4.4. Conclusiones

Como se mencionó, se comprende el mecanismo biológico que produce la ingesta de MDMA, sin embargo los cambios emocionales no son tan claros. Perceptiblemente se asocia a la intoxicación por esta droga con un aumento en la euforia, aumento de la sociabilidad, entre otras cosas.

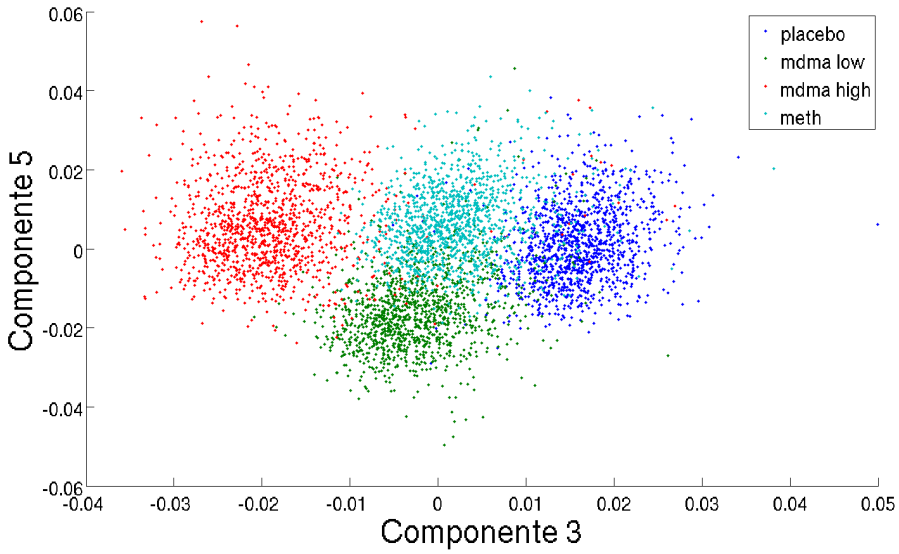

Fig. 4.9: Placebo, MDMA low, MDMA high y Meth en componentes 3 vs componente 5 de SVD1326.

Obtuvimos reportes de 13 sujetos en 4 condiciones distintas y los analizamos en búsqueda de singularidades estereotipadas. Como primera aproximación, medimos la distancia de cada reporte contra algunas palabras elegidas ad hoc usando Latent Semantic Analysis (entrenado con el corpus TASA). Mediante esta técnica, encontramos diferencias significativas entre las 4 condiciones para algunas de las palabras seleccionadas.

Luego, buscamos un método sin supervisión que nos permita prescindir de la elección manual de palabras. Para ellos, calculamos la distancia de cada uno de los relatos contra todo el diccionario proveniente del corpus TASA. Utilizando Singular Value Descomposition para reducir la dimensionalidad, encontramos un grupo pequeño de palabras estaba notablemente separada del resto Usando estas palabras *privilegiadas* logramos diferenciar significativamente las 4 condiciones.

Para determinar si el método era robusto, planteamos un esquema de dejar un sujeto afuera con sus 4 condiciones, realizando el análisis anterior con los 12 sujetos y testeando si la proyección del sujeto separado en la base que generaban los otros 12 era consistente con los entrenados. Entonces, dado un sujeto nuevo con sus 4 condiciones, podemos determinar la existencia de las condiciones a partir del entrenamiento con los restantes.

Finalmente, nos preguntamos si es posible clasificar los relatos de un sujeto en las cuatro condiciones. Para eso, propusimos analizar las distancias de los sujetos por condición, simplificando el problema solo a la comparación entre Placebo y MDMA high. Como resultado preliminar se verificó que con el clasificador Naive Bayes es posible clasificar bien estas dos condiciones. Dejamos como trabajo futuro tratar de involucrar la factorización de SVD para reducir la dimensionalidad y evaluar distintos clasificadores acordes a los datos.

El método propuesto tiene una gran virtud, ya que no es específico a este problema en particular, pudiéndose a extrapolar a cualquier otro dominio. En este dominio particular, también funda un nuevo análisis cuantitativo basado en la diferenciación de condiciones permitiéndonos un análisis parcial del estado mental a partir de la producción del discurso.

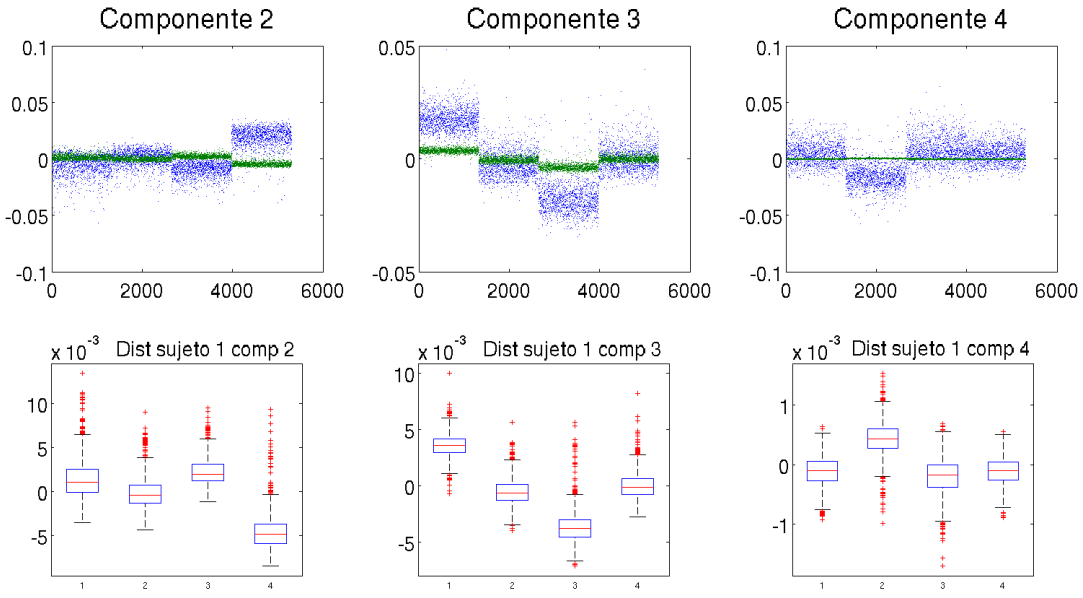

Fig. 4.10: Arriba en azul se ven las componentes de la factorización de todos los sujetos salvo el 1 de las componentes 2, 3 y 4 y en verde, el sujeto apartado, mismas componentes. Abajo se ven los boxplots de las distribuciones por condición del sujeto 1 según cada componente.

5. CONCLUSIONES GENERALES

La medicina tradicional basa sus decisiones en diversos factores, en particular, el diagnóstico clínico. La psiquiatría, como una de las áreas de la medicina, no escapa a esta regla. Sus métodos de diagnóstico estandarizados son complejos de implementar, debido a la intrínseca complejidad que define al objeto de estudio: el individuo y su mente.

La psiquiatría computacional puede contribuir con valiosa información intentando unir la brecha explicativa entre los procesos moleculares, celulares y mentales, complementando los métodos tradicionales de diagnóstico. El paradigma propuesto por este área describe los fenotipos computacionales, los modelos computacionales y su intento por aparearlos inequívocamente con los fenotipos cognitivos y biológicos, es decir entender cuáles son las características del modelo computacional producto de las patologías.

En este trabajo usamos la producción del discurso como ventana a la mente. De esta forma intentamos encontrar características en las producciones ligados con las patologías de los sujetos que las producen. Nos centramos en dos casos de estudio: la caracterización de discursos de pacientes psiquiátricos y de discursos de sujetos bajo el efectos de Ecstasy.

Para el procesamiento de discursos en pacientes psiquiátricos tomamos como punto de partida el trabajo propuesto por Mota et. al. el cual propone la extracción de fenotipos computacionales a partir de del análisis de grafos inducidos a partir del discurso. Nuestro aporte se basó en la automatización de este método y el mejoramiento de la clasificación obtenida a partir de estos resultados.

El método automático propuesto captura las singularidades de trabajo manual y logra una mejor clasificación para los mismos reportes del trabajo original. Como corolario de la automatización de los métodos, se está diseñando una aplicación, *PsycoGraph*, que permita acercar estas medidas al campo clínico como complemento de diagnóstico psiquiátrico tradicional.

En el estudio del discurso en sujetos bajo los efectos de Ecstasy, buscamos particularidades en reportes bajo cuatro condiciones distintas: Placebo, MDMA (baja dosis), MDMA (alta dosis) y Metanfetamina. Verificamos que a través de técnicas de similitud entre palabras es posible detectar diferencias entre los discursos que permiten separar entre las condiciones. Más aún, propusimos en método general para la reducción de dimensionalidad del problema, permitiendo encontrar grupos de palabras destacadas capaces de discriminar entre los grupos.

Esta tesis presentó evidencia de que los métodos automáticos de diagnóstico son posibles con el uso de herramientas de interpretación del lenguaje natural. El primer caso de estudio nos permitió ahondar en métodos que lograran abstraer información relevante a la forma en *cómo* se expresan los sujetos, descartando la información relevante a los tópicos del discurso. Esta información resultó relevante a la hora de clasificar a los pacientes psiquiátricos. El segundo caso de estudio, nos permitió indagar en los tópicos, es decir sobre *qué* hablan los sujetos bajos los efectos en la ingesta de Ecasty. Dos enfoques distintos, lograron capturar diferentes accidentes del discurso producidos por alteraciones en la mente.

6. TRABAJO FUTURO

El campo de la psiquiatría computacional tiene muchas preguntas por resolver. En principio, en este trabajo, nos abocamos al estudio del discurso, pero en su transcripción escrita. Todos los reportes fueron tomados de entrevistas donde los sujetos improvisan y componen en el momento lo que van a decir. Existe mucha evidencia [28] [29] de que en la *prosodia* se encuentra información relevante a estos estudios. Por lo que puede llegar a haber mucha información en el audio de los discursos.

Para las entrevistas del capítulo *Caso de estudio: Discursos en pacientes psiquiátricos* conseguimos algunos archivos de audio. La calidad de estos, junto al inmenso trabajo de preproducción que habría que hacerles (cortar las palabras, alinear la transcripción, etc) generó que no puedan ser parte de este trabajo. Sin embargo, creemos que en muchos aspectos prosódicos podríamos encontrar información valiosa para este tipo de estudios.

La automatización de estos procesos permite comenzar a pensar en políticas de diagnóstico masivo a través de métodos de comunicación como el teléfono y, por excelencia, internet. Ya existen casos de estudio llevados a la práctica donde un sujeto puede llamar por teléfono a un número y se le pide que contesten preguntas. Una de ellas es si fue diagnosticado con alguna enfermedad psiquiátrica, luego se lo graba para tomar mediciones acústicas y así generar una gran base de datos colaborativa con muestras de sujetos diagnosticados con enfermedades y sujetos control.

A continuación ampliamos dos temas que quedaron como trabajo futuro a lo largo de la tesis.

6.1. Discursos psiquiátricos: Explorando clasificadores

En el capítulo *Caso de estudio: Discursos en pacientes psiquiátricos* propusimos un método automático que analizaba discursos de pacientes psiquiátricos en búsqueda de particularidades que permitan identificar los distintos grupos (esquizofrénicos, maníacos y control). Para ésto generamos distintos tipos de grafos inducidos a partir del relato y les tomamos medidas topológicas que luego fueron consideradas *features* para la clasificación. En el trabajo original de Mota et. al., usaron el clasificador Naive Bayes para identificar a los 3 grupos y obtuvieron un rendimiento de 0.6250. Nosotros, automatizando este proceso alcanzamos, con el mismo clasificador un rendimiento de 0.75 (usando hasta 2 features). Sin embargo nos pareció interesante probar con otros clasificadores.

Para experimentar con otros clasificadores usamos la herramienta Weka [8]. Para no salir del esquema del trabajo y poder hacer una comparación decidimos usar hasta 2 features en la clasificación.

	SxMxC	Random μ	Random *std*
lematizado_L1 sintactico_LCC	0.9167	0.4331	0.1230

Tab. 6.1: Performance de clasificador LogitBoost para la clasificación de 3 grupos (esquizofrénico, maníaco y control).

Comparando las diferentes performance obtuvimos la mejor clasificación para estos datos (usando cross-validation 10 *folds*) con el clasificador *LogitBoost* [30]. Se puede ver su desempeño en la tabla 6.1 donde a se ve que obtuvimos una performance de 0.9167. Repitiendo el análisis de capítulo antes mencionado, simulando 100000 muestras construidas al azar y obteniendo Random μ y Random *std* concluimos que la performance de la clasificación usando este nuevo clasificador es significativa.

Habiendo obtenido una mejora significativa en la performance experimentando con otros clasificadores nos queda como trabajo futuro explorar más este universo. También, indagar en distintas heurísticas que nos permitan seleccionar features automáticamente y explorar más en técnicas de reducción de dimensionalidad.

6.2. Ecstasy: Análisis de tópicos en palabras emergentes

En el capítulo *Caso de estudio: Ecstasy* propusimos un método automático que nos permitió identificar relatos de sujetos bajo 4 condiciones distintas (la ingesta de placebo, MDMA low, MDMA high y Metanfetamina). En este método, analizando la muestra inicial de los relatos, nos encontramos con un conjunto de palabras (1326) que resultaron particularmente distintas a las demás. Estas distinción surgió del análisis de una factorización hecha sobre los relatos de los 13 sujetos.

Al observar que estas palabras, nos sirvieron como un filtro para *limpiar ruido* y nos permitirnos encontrar la forma de identificar las condiciones, nos preguntamos qué particularidad tendrían. Podría ser que estas palabras hayan servido para este caso particular y no tengan nada en común. Motivados por esta pregunta, decidimos hacer un breve análisis. En principio nos pareció que leer 1326 palabras e intentar abstraer tópicos manualmente no era posible. Por lo que comenzamos a analizar automáticamente las distancias entre conceptos. Par esto, construimos con la matriz de LSA TASA intersección WordNet, una matriz de distancias cosenos entre todas las palabras. Esta es una matriz es de 35112×35112 simétrica. Dibujándola, con un color cuando las palabras estuvieran cercanas (mayor a 0.1) y con otro cuando estuvieran lejanas, no se logro ver nada más que puntos dispersos al azar. Decidimos entonces, ubicar primero las 1326 *palabras emergentes* y graficar la matriz (para poder apreciar la figura mostramos solo las 4000 primeras palabras). Podemos en la figura 6.1 que las 1326 palabras forman un cuadrado más oscuro y uniforme que las demás regiones. Haciendo zoom sobre este cuadrado vemos en la figura 6.2 que el cuadrado con las *palabras emergentes* es más denso. Esto indica que las palabras del conjunto son más cercanas entre si. También observamos que la diagonal es de un color muy fuerte, eso es así debido que en la diagonal tenemos las distancias entre cada misma palabra.

Viendo esta misma figura observamos algo sorprendente, vemos una linea horizontal muy intensa. Esta linea indica que la palabra que corresponde a ese número resulta ser extremadamente cercana a todas las otras 1326. Buscando en cual es la palabra que corresponde a ese número, concluimos que la palabra más cercana a todas las demás *palabras emergentes* es: **affective**. Resulta interesante que la palabra más distinguible del conjunto sea *affective*, las descripciones de los síntomas emocionales producto de la ingesta del Ecstasy describe un aumento en la afectividad de los consumidores.

El estudio de los kernels de palabras cercanas para las medidas semánticas resulta interesante, dado que existen diversas formas de clusterizar las palabras. Nos sorprende que las distancias al kernel que obtuvimos resulte efectivo a la hora de separar los sujetos bajo la ingesta de Ecstasy. También resulta sorprendente que exista una dirección, es decir

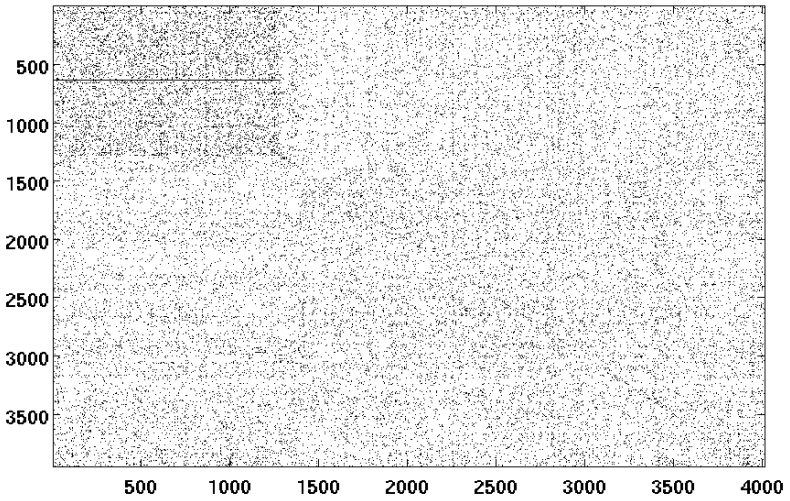

Fig. 6.1: Distancias LSA de palabras de TASA intersección Wordnet. Las primeras 1326 palabras son las *palabras emergentes*. Se muestran las primeras 4000 para que se aprecie. Se puede ver un cuadrado arriba a la izquierda de un color más fuerte. Estas son exactamente las 1326 *palabras emergentes*.

una palabra, tan privilegiada en ese kernel que sea tan coherente con la descripción de los síntomas emocionales.

Queda como trabajo futuro entender en profundidad la técnica automática y su generalización a otros dominios, basándose en el capitulo *Caso de estudio: Ecstasy* y cómo este método logra capturar las características del efecto en la ingesta de ecstasy.

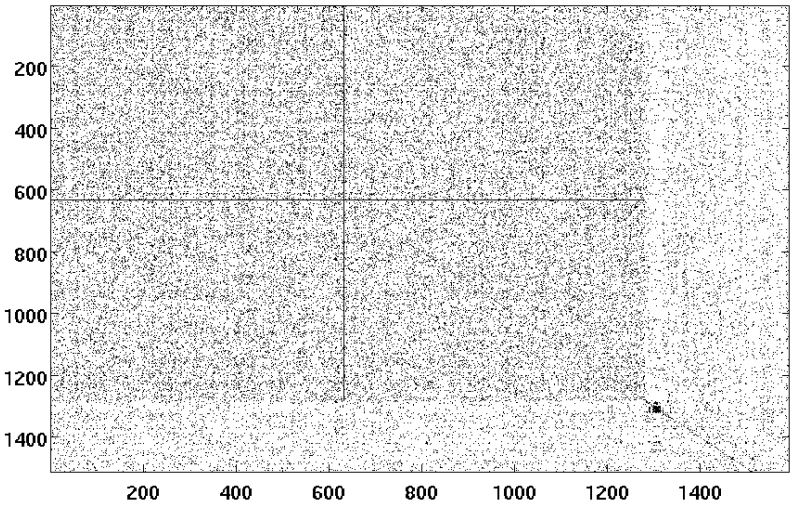

Fig. 6.2: Zoom de la figura 6.1 en las primeras 1600 palabras.

Bibliografía

[1] P.R. Montague, R.J. Dolan, K.J. Friston, and P. Dayan. Computational psychiatry. *Trends in cognitive sciences*, 2011.

[2] A.M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[3] T.V. Maia and M.J. Frank. From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, 14(2):154–162, 2011.

[4] V.B. Gradin, P. Kumar, G. Waiter, T. Ahearn, C. Stickle, M. Milders, I. Reid, J. Hall, and J.D. Steele. Expected value and prediction error abnormalities in depression and schizophrenia. *Brain*, 134(6):1751–1764, 2011.

[5] Russell A. Poldrack, Jeanette A. Mumford, Tom Schonberg, Donald Kalar, Bishal Barman, and Tal Yarkoni. Discovering relations between mind, brain, and mental disorders using topic mapping. *PLoS Comput Biol*, 8(10):e1002707, 10 2012.

[6] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. *TODO*.

[7] N.B. Mota, N.A.P. Vasconcelos, N. Lemos, A.C. Pieretti, O. Kinouchi, G.A. Cecchi, M. Copelli, and S. Ribeiro. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4):e34928, 2012.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[9] M.E. Liechti, A. Gamma, F.X. Vollenweider, et al. Gender differences in the subjective effects of mdma. *Psychopharmacology*, 154(2):161–168, 2001.

[10] G.A. Ricaurte and U.D. McCann. Experimental studies on 3, 4-methylenedioxymethamphetamine (mdma,"ecstasy") and its potential to damage brain serotonin neurons. *Neurotoxicity research*, 3(1):85–99, 2001.

[11] J.C. Kraner, D.J. McCoy, M.A. Evans, L.E. Evans, and B.J. Sweeney. Fatalities caused by the mdma-related drug paramethoxyamphetamine (pma). *Journal of analytical toxicology*, 25(7):645–648, 2001.

[12] K.S. Leung and L.B. Cottler. Ecstasy and other club drugs: a review of recent epidemiologic studies. *Current opinion in psychiatry*, 21(3):234–241, 2008.

[13] L.B. Cottler, S.B. Womack, W.M. Compton, and A. Ben-Abdallah. Ecstasy abuse and dependence among adolescents and young adults: applicability and reliability of dsm-iv criteria. *Human Psychopharmacology: Clinical and Experimental*, 16(8):599–606, 2001.

[14] AL Stone, CL Storr, and JC Anthony. Evidence for a hallucinogen dependence syndrome developing soon after onset of hallucinogen use during adolescence. *International journal of methods in psychiatric research*, 15(3):116–130, 2006.

[15] W.N. Francis, H. Kučera, and A.W. Mackie. *Frequency analysis of English usage: lexicon and grammar*. Houghton Mifflin Harcourt (HMH), 1982.

[16] M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[17] E. Loper and S. Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.

[18] J. Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

[19] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2001.

[20] R.E. Tarjan. Applications of path compression on balanced trees. *Journal of the ACM*, 26(4):690–715, 1979.

[21] T.K. Landauer and S.T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, 104(2):211, 1997.

[22] Paul MBVitanyi Rudi L. Cilibrasi, Rudi LCilibrasi — Paul M.B. Vitanyi. The google similarity distance. *TODO*, 19(3), 2007.

[23] W.H. Kruskal and W.A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[24] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[25] M. Singh and G.M. Provan. A comparison of induction algorithms for selective and non-selective bayesian classifiers. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 497–505. Citeseer, 1995.

[26] G.A. Miller et al. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[27] I. Raskovsky, D. Fernandez Slezak, CG Diuk, and GA Cecchi. The emergence of the modern concept of introspection: a quantitative linguistic analysis. In *Young Investigators Workshop on NAACL*, pages 68–75, 2010.

[28] K.A. Kobak, S.L. Dottl, J.H. Greist, J.W. Jefferson, D. Burroughs, J.M. Mantle, D.J. Katzelnick, R. Norton, H.J. Henk, R.C. Serlin, et al. A computer-administered telephone interview to identify mental disorders. *JAMA: the journal of the American Medical Association*, 278(11):905–910, 1997.

[29] D. Sturim, P. Torres-Carrasquillo, T.F. Quatieri, N. Malyska, and A. McCree. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[30] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.