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

Modelos computacionales de discurso libre sobre tratamientos con drogas psiquiátricas y su relación con datos farmacoquímicos

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

Matias Laporte

Director: Carrillo, Facundo
Codirector: Tagliazucchi, Enzo
Buenos Aires, 2019

RESUMEN

En los últimos años, las ciencias de la computación contribuyeron a las ciencias médicas de maneras disruptivas, mejorando los diferentes campos de aplicación. En particular, en la psiquiatría/psicología y demás ciencias cognitivas, los modelos y algoritmos de procesamiento del lenguaje natural (NLP) mejoraron el estado del arte tanto en las descripciones y modelado fenomenológico, como en la práctica clínica.

Trabajos pasados demostraron la factibilidad de la aplicación de técnicas de NLP para el diagnóstico de patologías psiquiátricas [9, 10, 12, 22], la caracterización del discurso por intoxicaciones farmacológicas [3], y la asociación entre los reportes de efectos subjetivos y el perfil de afinidad de fenetilaminas y triptaminas sustituidas [35], entre otros.

El presente trabajo propone el análisis de características farmacoquímicas (estructura química, afinidad por receptores) de distintos tipos de drogas psiquiátricas (antipsicóticos, antidepresivos) y su relación con los efectos subjetivos reportados por pacientes que se encuentran bajo tratamiento con ellas.

Para la caracterización de la semántica de los reportes, recolectados de diversos sitios web, se utiliza la técnica de Análisis de la Semántica Latente (LSA) aplicada a dos corpus distintos, uno de ellos externo. Los datos de afinidad, por su parte, surgen de bases de datos públicas, y los de estructura química se calculan mediante algoritmos de *fingerprinting* a partir de sus descriptores moleculares.

Los resultados del trabajo evidencian una correlación positiva significativa entre la semántica obtenida de los reportes asociados a una droga, su estructura química y su afinidad a los distintos receptores. Asimismo es posible distinguir a partir de los valores de similitud semántica entre los subtipos de los distintos pares de drogas (antidepresivos de tipo SSRI, SNRI, TCA, etc.).

A su vez, el trabajo como punto de partida para análisis más profundos en el área. El objetivo final de esta línea de investigación, que excede esta tesis, es optimizar la acción de farmacos sobre distintos receptores cerebrales para minimizar las componentes del discurso natural asociadas a efectos indeseables.

Palabras claves: lenguaje natural, fármacos psiquiátricos, análisis semántico, afinidad por receptores, trastornos mentales.

ABSTRACT

In the last years, Computer Science's contribution to the Medical Sciences has been disruptive, improving different fields of application. Particularly, in psychiatry/psychology and other cognitive sciences, models and natural language processing (NLP) algorithms have refined the state of the art in the description and phenomenological modeling, as much as the clinical practice.

Past works have shown the factibility of using NLP techniques for automated psychiatric diagnosis [9,10,12,22], speech characterization of psychoactive drug effects [3], and the association between reported subjective effects and binding affinity profiles of substituted phenetylamines and tryptamines, [35], among others.

This thesis proposes the analysis of pharmacochemical characteristics (chemical structure, binding affinity) of several types of psychiatric drugs (antipsychotics, antidepressants) and its relation with the reported subjective effects of patients undergoing treatment with them.

For the semantic characterization of the reports, collected through diverse websites, latent semantic analysis (LSA) is used and applied to two distinct corpus, one of them external. Binding affinity profiles are obtained from public databases, while chemical structure data is calculated using fingerprinting algorithms.

The work's results evidences a significative positive correlation between the semantic of drug reports, its chemical structure and binding affinity to receptors. Likewise it is possible to distinguish from the semantic similarity the subtypes between the different pairs of drugs (SSRI, SNRI, TCA-type antidepressants, etc.).

Findings show that it may prove beneficial to continue research in the area. The objective of this line of research, which exceeds this thesis, is to optimize the drugs action over cerebral receptors to minimize speech components associated to undesirable effects.

Keywords: natural language, psychiatric drugs, semantic analysis, binding affinity profile, psychiatric disorders.

AGRADECIMIENTOS

A mis directores, Facundo y Enzo, y a Federico, por agarrar este hierro caliente y ayudar a forjarlo, transmitiendo infinidad de enseñanzas y consejos en el proceso.

A mi madre, por darme la libertad de elegir cualquier camino.

A mi padre, por haberme hecho elegir uno.

A Sebastián, por enseñarme a transitarlo.

A la educación pública, que en todos sus niveles me dio infinitas posibilidades; a sus docentes, que con la estructura que tenemos logran sacarle agua a las piedras.

A la Antártida, que me dejó sin palabras.

A la Dra. Carla Capozucca, por sus atinadas críticas y sus consejos.

A los amigos de la facultad, con los que compartimos tantas horas de clases y de insomnio y que son también parte de esto, Julián, Joaquín, Javier, Gastón, Juan, Vanesa, Jennifer, Francisco; a los docentes que se tornaron amigos, como LuisA.

A los amigos de siempre, Mara, Federico y Germán, por permitirme dejar de pensar en binario de vez en cuando.

A Leo y Pupi, por tanto camino recorrido juntos, que nos llevó tanto tiempo, y la competencia en broma, que nos motivó a terminar.

A Vizcky, que compartió muchas de sus noches durante la primer etapa; a Mimi, que hizo lo propio en la última parte.

A mi familia extendida, Liliana, Jorge, Franco y Nadia, por haberme incluido como uno más y por la frescura que aportaron a mi vida.

A los Co, por las tardes de charlas y cariño.

A mi familia, Paula y Camila; Sebastián, Julia, Alfonso y Lucía; Eugenia, Diego, Astor y Fermín; Santiago, Daniela, Bautista y Fausto; Corina, Elisa y Lorena; por relajarme y divertirme en cada encuentro, y por hacerme notar que siempre están ahí.

A la Dra. Romina Mayra Lasagni Vitar, por su compañía incondicional, por su apoyo teñido con su experiencia en esta etapa definitoria, por ser la musa que inspira mis sueños, y por devolverme las palabras.

La cualidad de la fuerza de voluntad es, precisamente, el crecimiento. El logro supone su cancelación. Para ser, la fuerza de voluntad debe aumentar con cada logro, haciendo que el logro no sea más que un paso dado hacia la siguiente aspiración. A mayor poder obtenido, mayor apetito.

–Ursula K. Le Guin, *La rueda celeste*

A mi familia, y a la ciencia ficción

Índice general

1. INTRODUCCIÓN

Los falsos comienzos y futilidades de los años anteriores resultaron ser basamentos, cimientos puestos a ciegas aunque bien puestos.
–Ursula K. Le Guin, *Los desposeídos*

Desde los inicios de la Computación se ha perseguido la construcción de dispositivos (en un principio, mecánicos) que pudieran facilitar operaciones realizadas por la mente de un individuo, permitiendo ahorrar de ese modo tiempo y como costos (y, por qué no, sufrimiento a dicho individuo).

Se pueden considerar como primeros ejemplos el mecanismo de Anticítera, o la máquina diferencial de Babbage, dispositivos que permitían realizar cálculos de tipo astronómico y en el último caso, además, la creación de tablas matemáticas.

A medida que evolucionó la tecnología (y nuestro entendimiento y relación con ella), las exigencias fueron en aumento: una computadora debía no solo poder realizar cálculos, sino poder ofrecer razonamientos provenientes de un nivel de abstracción superior, conocimiento, e información.

Vannevar Bush, en su artículo *Como podríamos pensar* [8], veía en la tecnología el potencial de un *índice de memoria* (*memex*, por su nombre en inglés *memory index*), una manera de poder organizar mecanizadamente el conocimiento humano (tal como realiza la memoria humana) y acceder a él con flexibilidad y rapidez. Teniendo en cuenta que el de Bush fue un concepto teórico, no obstante, se lo considera vaticinador de muchos elementos actuales: el hipertexto, las enciclopedias online, Internet, y más.

Otros, en cambio, veían a la computadora como un mero accesorio, sin esperar que la comunicación entre la misma y un humano pasara de algo superficial. Con el fin de demostrar eso mismo, Joseph Weizenbaum diseñó ELIZA[1], un programa que parodiaba las respuestas de un psicoterapeuta no directivo [2] y, básicamente, repetía lo que el paciente decía (mediante una serie de reglas, realizaba búsqueda de patrones[3] y deducía qué debía repetir). A pesar de su cometido, Weizenbaum logró lo contrario: la gente se involucraba emocionalmente con el programa, y olvidaba estar hablando con una computadora. ELIZA consiguió ser, además, uno de los primeros *chatbots*[4], y uno de los primeros programas de **procesamiento del lenguaje natural** (**NLP**[5]), campo de las Ciencias de la Computación que será introducido en la sección 1.3 Procesamiento de Lenguaje Natural).

Siendo el cerebro el único órgano que puede aprender sobre sí mismo, y teniendo en cuenta que las computadoras fueron evolucionando y obteniendo funcionalidades antes reservadas únicamente para los humanos, el próximo paso sería poder utilizar a las computadoras para aprender sobre el cerebro y la mente.

[1] Versión moderna del programa puede ser consultada aquí
[2] Terapia centrada en la persona, buscando empatizar con el paciente y que hace foco en lo que el mismo dice
[3] Pattern Matching
[4] Programas de computadora (*bots*) capaces de mantener una conversación (*chat*)
[5] Por sus siglas en inglés, *natural language processing*

Dado que sería ambicioso pretender comprender la totalidad de ambos[6], deberemos conformarnos con estudiar una única parte: en particular, podríamos intentar entender mediante modelos computacionales cuáles son los efectos subjetivos reportados por pacientes psiquiátricos de la acción de diferentes drogas antipsicóticas y antidepresivas de prescripción.

Resulta innegable que Internet ha probado ser un ecosistema invaluable para un amplio abanico de propósitos. En lo que respecta a la ciencia, por ejemplo, permite la realización de análisis de datos a gran escala sobre situaciones que podrían ser difíciles de replicar en un ambiente controlado, ya sea por razones de recursos y, por qué no, hasta legales.

Teniendo en cuenta lo difícil que es poder realizar un ensayo masivo consistente en la administración de psicofármacos a individuos con distintos desórdenes psiquiátricos (de los cuales se hablará en la sección 1.1 Psiquiatría), y lo fácil que es publicar en Internet lo que uno siente y piensa, se hará uso de esta ventaja para poder obtener los reportes subjetivos de los pacientes, y se utilizarán datos asimismo públicos (como la estructura química de las drogas y su afinidad, conceptos definidos en la sección 1.2 Farmacología) para poder realizar una caracterización farmacoquímica de las drogas y relacionarlas con los comentarios que las mismas suscitan.

Es la hipótesis de este trabajo que existe una correlación entre las distintas propiedades de los fármacos (estructura química, y afinidad por distintos receptores) y la semántica que emerge de los relatos de los pacientes bajo tratamiento con los mismos.

1.1. Psiquiatría

La psiquiatría es una ciencia, rama de la medicina, que se dedica al estudio y tratamiento de las enfermedades mentales. De acuerdo al manual diagnóstico y estadístico de los trastornos mentales (DSM-5 [1]):

> Un trastorno mental es un síndrome caracterizado por una perturbación clínicamente significativa del conocimiento, la regulación emocional o la conducta de un individuo, que refleja una disfunción en los procesos psicológicos, biológicos o evolutivos que subyacen al funcionamiento mental.

Para tratar los trastornos mentales, la psiquiatría propone dos tipos de abordaje. Por un lado, los tratamientos psicoterapéuticos y, por el otro lado, los biológicos, mediante el suministro de medicamentos que actúen sobre la bioquímica del cerebro.

El **DSM-5** es una herramienta para la clasificación de los distintos tipos de desórdenes mentales, que sirve para ayudar al profesional a la hora del diagnóstico de un paciente. Entre algunos que se discutirán más adelante en este trabajo, podemos encontrar:

- Trastornos de ansiedad.
- Trastorno bipolar y relacionados.
- Trastornos depresivos.
- Espectro de la esquizofrenia y otros trastornos psicóticos.

[6] Emerson W. Pugh: *"Si el cerebro del ser humano fuera tan simple que pudiéramos entenderlo, seríamos tan simples que no lo entenderíamos."*

Trastornos de ansiedad

> Los trastornos de ansiedad incluyen trastornos que comparten características de miedo excesivo y ansiedad y otras perturbaciones relacionadas con el comportamiento. *Miedo* es la respuesta emocional a una amenaza inminente, sea percibida o real, mientras que la *ansiedad* es la anticipación a una amenaza futura.[...] Los ataques de pánico aparecen prominentemente dentro de los trastornos de ansiedad como un tipo particular de respuesta al miedo.
>
> **DSM-5**

Trastorno bipolar y relacionados

> En la quinta edición del DSM, el trastorno bipolar y relacionados se encuentra entre los capítulos del espectro de la esquizofrenia y los trastornos depresivos por ser un *puente* entre ambas clases de diagnóstico en términos de sintomatología, historia familiar y genética. Los diagnósticos incluyen **Trastorno bipolar I**, **Trastorno bipolar II**, **Trastorno ciclotímico**, **Trastorno bipolar inducido por sustancias/medicación**, **Trastorno bipolar debido a otra condición médica**, y otros [...]. **DSM-5**

Si bien la categorización de la enfermedad cambió en las distintas ediciones del manual, el trastorno bipolar de tipo I puede verse (con algunas salvedades) como el clásico trastorno maníaco-depresivo, o psicosis afectiva, taxonomía utilizada en el S.XIX. El de tipo II requiere al menos un episodio de depresión mayor y al menos un episodio de hipomanía.

Trastornos depresivos

> El trastorno depresivo mayor representa la condición clásica de este grupo de trastornos. Se caracteriza por episodios discretos de al menos dos semanas de duración [...] que involucran cambios bien definidos en el afecto, la cognición y funciones neurovegetativas, con remisiones entre episodios. **DSM-5**

Espectro de la esquizofrenia y otros trastornos psicóticos

> El espectro de la esquizofrenia y otros trastornos psicóticos [...] se define por anormalidades en uno o más de los siguientes campos: delirios, alucinaciones, pensamiento desorganizado (visible en el discurso), comportamiento motriz anómalo o severamente desorganizado (incluida la catatonia), y síntomas negativos.
>
> - Los **delirios** son comportamientos fijos que no presentan cambios a la luz de evidencia que los contradiga.
> - Las **alucinaciones** son experiencias perceptuales que ocurren sin estímulos externos.
> - El **pensamiento desorganizado** se infiere a través del discurso del individuo. Éste puede cambiar de un tópico a otro, o dar respuestas poco o nada relacionadas con la pregunta hecha. Raramente, el discurso puede estar tan desorganizado de manera de tornarse incomprensible.

- El **comportamiento motriz anómalo** puede manifestarse de varias maneras, desde una torpeza similar a la infantil o como agitación impredecible. [...] Lleva a dificultades en actividades diarias.

DSM-5

1.2. Farmacología

La farmacología se define como la *ciencia biológica que estudia las acciones y propiedades de los fármacos en los organismos* [14], donde los fármacos son cualquier tipo de sustancia química (*molécula*) que puede interactuar con organismos vivos al unirse a receptores (*macromoléculas proteicas*) situadas en la membrana, citoplasma o núcleo de sus células, desencadenando una serie de reacciones que alteran sus propiedades fisiológicas. El fármaco se limita a regular procesos propios de la célula (es decir, la misma no desconoce los mecanismos puestos en acción).

Una rama de la farmacología, la **farmacología terapéutica**, es la que *estudia la aplicación de los fármacos en el ser humano con la finalidad de curar o de alterar voluntariamente una función normal* [14].

1.2.1. Receptores

Los **receptores farmacológicos** son moléculas con las que los fármacos (*ligandos*) interactúan de forma más o menos selectiva, produciendo una modificación específica y generalmente transitoria en la función celular.

Las proteínas con más probabilidades de unirse a fármacos son las que tienen la capacidad de regular la comunicación intercelular mediante la influencia de sustancias propias del cuerpo humano (*endógenas*). Entre tales sustancias se encuentran, por ejemplo, los neurotransmisores, los neuromoduladores, las hormonas y otros, que al ser liberados por una célula, pueden influir sobre la actividad de otra o sobre sí misma.

Dos características de un receptor farmacológico son la afinidad y la especificidad. La **especifidad** le permite al receptor distinguir una molécula de otra. Valores elevados de **afinidad** permiten al receptor unirse al fármaco aún cuando se encuentre en concentraciones bajas. La constante de inhibición ($\mathbf{K_i}$) está relacionada de manera inversamente proporcional a la afinidad del fármaco por un receptor determinado.

La **eficacia** de un fármaco es su capacidad para poder generar la respuesta biológica esperada luego de su interacción con el receptor. Si el fármaco logra tal respuesta es considerado **agonista**, y si se une al receptor pero sin activarlo es considerado **antagonista**. Hay, a su vez, subtipos de agonismo y antagonismo. La afinidad no es necesariamente idéntica a la eficacia, ya que un fármaco puede unirse con alta afinidad a un receptor pero sin generar una respuesta biológica.

Los receptores que forman parte de este estudio pueden encontrarse en la tabla 4.1 del apéndice.

Antipsicóticos

Se clasifica como antipsicóticos al *conjunto de fármacos que muestran su mayor eficacia en el tratamiento de algunas psicosis orgánicas y tóxicas, y de las psicosis idiopáticas de*

naturaleza esquizofrénica [14].

Uno de los primeros antipsicóticos descubiertos fue la **clorpromazina**, una fenotiazina; luego, se desarrollaron varias subfamilias de fenotiazinas, tioxantenos y butirofenonas, entre las que sobresale el **haloperidol**. La acción de estos antipsicóticos, llamados **antipsicóticos típicos**, se basa en la acción bloqueante de los receptores dopaminérgicos D_2.

Con el descubrimiento de la **clozapina**, se da el surgimiento de los **antipsicóticos atípicos**, similares en eficacia clínica a los típicos, pero con *menor tendencia a provocar reacciones extrapiramidales y a aumentar la secreción de prolactina* [14]. Los antipsicóticos atípicos exhiben acción bloqueante en los receptores dopaminérgicos D_2 así como en los serotonérgicos $2A$.

Antidepresivos

Algunos de los primeros fármacos a los cuales se les atribuyó acción terapéutica para la depresión fueron la **iproniazida**, inhibidor de la enzima MAO[7], y la **imipramina**, un compuesto tricíclico inspirado en las fenotiazinas y que es uno de los grupos más grandes de antidepresivos.

La novedad introducida por los antidepresivos tricíclicos es la asociación a través de su nombre de una estructura química (*tres ciclos*, fig.1.1) con una acción farmacológica específica.

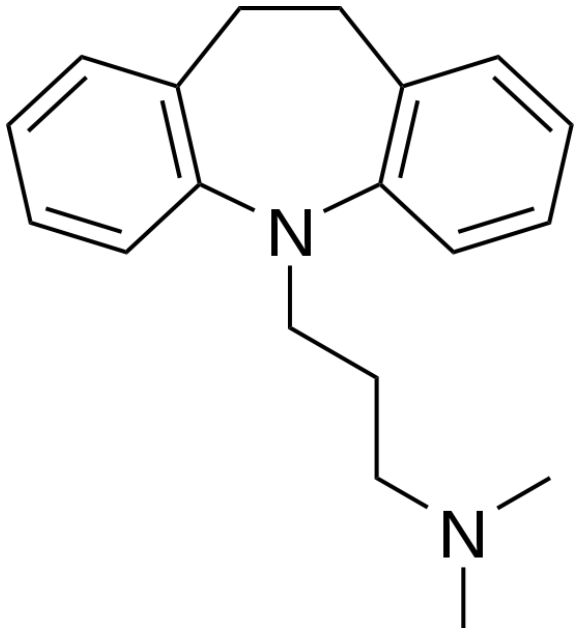


Fig. 1.1: Estructura química de la imipramina. Imagen en el dominio público

Posteriormente fueron desarrollados otros fármacos con acción antidepresiva. Aunque sin estructura tricíclica, su acción se basa en el bloqueo selectivo de la recaptación de algunas de las aminas biógenas (serotonina, noradrenalina, dopamina).

[7] Monoamino oxidasa

Entre estos compuestos se pueden encontrar a los antidepresivos **tetracíclicos** (TeCA), los **inhibidores selectivos de la recaptación de serotonina** (SSRI), los **inhibidores selectivos de la recaptación de serotonina y noradrenalina** (SNRI), los **antagonistas e inhibidores selectivos de la recaptación de serotonina** (SARI), y los **inhibidores de la recaptación de dopamina y noradrenalina** (NDRI).

1.3. Procesamiento de Lenguaje Natural

El procesamiento del lenguaje natural (**NLP**), mencionado anteriormente, es un campo de las Ciencias de la Computación, Inteligencia Artificial y Linguística, encargado de estudiar la interacción entre las computadoras y el lenguaje humano. Involucra diversas áreas (reconocimiento de voz, síntesis de voz, entendimiento del lenguaje), y su objetivo es procesar y analizar grandes volúmenes de datos relacionados con lenguajes naturales.

1.3.1. Análisis semántico

Dentro del campo de NLP, el análisis semántico busca extraer de manera automática el significado de un texto. Aplicado a las reseñas de los fármacos hechas por los pacientes, nos permite descubrir o darle un significado a las mismas, trascendiendo el significado individual de cada palabra que las conforma. El NLP será utilizado luego para analizar la relación entre el dominio de los cambios semánticos, y los dominios de afinidad por receptores y similitudes estructurales entre moléculas.

Para realizar el análisis semántico se utilizaron distintas técnicas ya establecidas de **NLP** relacionadas con *word embeddings*.

Word Embeddings

Word embeddings es el nombre que se le da a un conjunto de metodologías consistente en la transformación de una palabra en un vector. Es decir, se pasa de un espacio de representación donde cada palabra es una dimensión, hacia un espacio de dimensionalidad menor. Esta transformación permite abstraernos del significado como representación de una palabra para utilizar un vector en su lugar. De este modo, se puede definir una noción de distancia semántica entre palabras similares (como Similitud coseno), en base al cálculo de la distancia de los vectores que las representan.

La metodología de *word embedding* utilizada en este trabajo fue **Latent Semantic Analysis**; existen otras como WordNet [21], FastText [7], GloVe [23], y más.

Latent Semantic Analysis

LSA [20] ("análisis de la semántica latente" en español) es un modelo asociativo lineal de alta dimensionalidad utilizado para analizar grandes corpus de texto natural y generar una representación que capture la similitud entre palabras y pasajes de texto.

El método pretende emular, sin ningún tipo de conocimiento lingüístico previo, la adquisición del lenguaje en niños, quienes aprenden a un ritmo mayor que el esperable dada la información a la que fueron expuestos; esto es porque *infieren* información a partir del contexto en el que se encuentran las palabras. **LSA** tiene como sustento principal la hipótesis distribucional [18]: las palabras que aparecen en los mismos contextos comparten significado similar.

En este trabajo, se utilizará LSA de dos maneras. Por un lado, se utilizará el corpus TASA[8], un corpus de textos en inglés consistente de 37.651 documentos que contienen 92.393 términos distintos; por el otro, se aplicará LSA sobre el corpus compuesto por la totalidad de las reseñas obtenidas.

1.3.2. Similitud coseno

Para poder obtener una medida sobre cuán similares son dos vectores en un mismo espacio se puede utilizar la similitud coseno. En nuestro caso, será una medida de cuán parecidas semánticamente son dos reseñas de fármacos.

La similitud coseno se deduce a partir de la fórmula del producto escalar para dos vectores **A** y **B**

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\|\|\mathbf{B}\| \cos(\theta) \tag{1.1}$$

Por lo tanto,

$$similitud = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1.2}$$

La similitud, al ser un coseno, es un valor que se encuentra entre -1 (exactamente opuestos), y 1 (exactamente iguales), mientras que 0 corresponde al caso ortogonal.

[8] Touchstone Applied Science Associates, disponible aquí

2. MATERIALES Y MÉTODOS

El hombre que se propone llevar a un gato por su cola aprende algo que siempre será útil y que nunca se tornará oscuro ni dudoso. –Mark Twain

2.1. Obtención y preprocesamiento del conjunto de datos

Los datos utilizados para el trabajo son de tres tipos:

- La afinidad de las drogas psiquiátricas estudiadas respecto de una serie de receptores[1] presentes en células del sistema nervioso humano
- La estructura química 2D de las drogas[2]
- Los reportes de los pacientes tratados con las drogas consideradas[3].

2.1.1. Afinidad

Para la afinidad de los fármacos con los receptores se utilizó información públicamente disponible en las bases de datos online Ki Database [5] y ChEMBL [16]. De estas fuentes de datos se obtuvo la constante de inhibición (K_i) de cada droga para un amplio espectro de receptores, y en muchos casos, ensayos funcionales que caracterizan la acción farmacológica en cada receptor (agonismo, agonismo parcial, antagonismo, etc.). Se calculó el promedio de los distintos valores recopilados para cada par droga-receptor.

Dado que los valores de K_i se encuentran distribuidos en un rango muy amplio (de 0 a 10.000, donde valores más pequeños se corresponden con afinidades más altas) que dificulta su comparación, se realizó la transformación descrita en Ray (2010) [24], cuya fórmula es:

$$pK_i = -\log_{10}(K_i) \tag{2.1}$$

De este modo, una mayor afinidad del fármaco con los receptores se condice con un pK_i más alto, en unidades que representan el orden de magnitud del valor de K_i.

Finalmente, se elaboró una matriz cuyas filas y columnas corresponden a las distintas drogas, y cada celda consiste de una tupla cuyo primer elemento es la correlación de Pearson de los pK_i de los receptores en común entre ambas drogas, y el segundo elemento indica cuántos receptores en común poseen ambas drogas.

Es decir, sean:

- D_1, D_2: las drogas
- R_{D_1}, R_{D_2}: receptores afectados por D_1, D_2

[1] Los receptores estudiados se listan en el apéndice, tabla 4.1
[2] Las drogas aparecen detalladas en el apéndice, Fármacos y sus receptores
[3] La distribución de reportes se encuentra en el apéndice 4.4

Entonces:

$$M_{D1,D2} =< Pearson(pK_iR_{D_1}, pK_iR_{D_2}), |R_{D_1} \cap R_{D_2}| > \tag{2.2}$$

2.1.2. Estructura molecular

La estructura de las drogas se obtuvo a través de la plataforma de integración de data mining KNIME [4] en combinación con la extension RDKit [25], que codifica la estructura de cada fármaco como una serie de bits. La estructura se determina mediante distintos algoritmos de *fingerprinting*, los cuales tienen en cuenta el rol funcional de los átomos en la molécula, entre otras propiedades químicas.

El *fingerprint* elegido para el trabajo fue **AtomPair** [30], que representa cada molécula como un conjunto de pares de átomos, compuesto por los pares de átomos (sin incluir al hidrógeno) junto con la distancia entre ambos (medida como la cantidad de enlaces que conforman el camino mínimo entre ambos). No hay diferencias significativas entre los distintos tipos de *fingerprint*, que justificaran la elección de uno en particular [26].

Obtenida la cadena de bits correspondiente a cada uno de los fármacos, la comparación entre ellas fue realizada mediante el coeficiente de Tanimoto (o de Jaccard). El mismo se calcula como el cociente entre el tamaño de la intersección de ambos conjuntos y el tamaño de su unión.

$$T(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.3}$$

El resultado, entre 0 y 1, da una medida de similitud entre ambos conjuntos; en nuestro caso, corresponde a la similitud estructural entre pares de drogas.

2.1.3. Reportes

2.1.3.1. Scrapping

Los discursos de los pacientes, donde reportan los efectos producidos por la ingesta de los fármacos, fueron obtenidos de los sitios Drugs y WebMD. Ambos sitios cuentan con una completa base de datos de medicamentos, donde los usuarios pueden reportar con detalle su experiencia al realizar un tratamiento que incluya la ingesta de alguno de ellos. Los relatos incluyen, en mayor o menor medida y entre otros atributos no detallados aquí:

- Nombre comercial del medicamento
- Principio activo
- Información demográfica del paciente
- Patología por la cual recibe el tratamiento
- Duración del tratamiento
- Dosis prescripta
- Valoración (numérica) del medicamento por parte del paciente

En el caso de WebMD.com, algunas de las reseñas de medicamentos incluyen únicamente una evaluación de 1 a 5 (en las dimensiones de efectividad, facilidad de uso y satisfacción) por parte del paciente, sin ningún relato propiamente dicho. Como el objetivo de este trabajo es el análisis semántico sobre el discurso, estas reseñas fueron desestimadas por no contribuir a tal fin.

Mediante *web scraping*[4] y un posterior formateo de los archivos HTML obtenidos, se obtuvieron 24.597 reseñas provenientes de Drugs.com, y 7.986 de WebMD, correspondientes al listado de fármacos analizado (detallado en 3 Orden de las drogas en los ejes y su clasificación).

2.1.3.2. Preprocesamiento del texto

Teniendo el conjunto de reseñas en un formato estructurado (*DataFrame* de Pandas[5] para Python), se procedió a realizar el procesamiento del texto.

Siendo el objetivo del análisis semántico el llevar todas las palabras a un espacio multidimensional para compararlas según su posición en el mismo, resulta conveniente realizar una reducción del universo de palabras para disminuir la variabilidad.

Se comenzó con la conversión a minúsculas de todas las palabras del corpus. Luego, al estar lidiando con reseñas en el idioma inglés, se analizó cuáles eran las contracciones con mayores apariciones en los textos (de este modo, se podían considerar también contracciones con errores de escritura, como *arent* y *doesnt*) y se realizó una expansión manual de las mismas (convirtiendo las dos anteriores, como ejemplo, a *are not* y *does not*).

Cada reseña fue luego convertida en una secuencia de oraciones (mediante la función *sent_tokenize* del paquete NLTK[6]), y cada una de éstas en una secuencia de palabras (utilizando la función *word_tokenize*, del mismo paquete).

Finalmente, se removieron de todas las palabras los caracteres que no fueran letras, limpiando así del corpus indicaciones de cantidades, errores tipográficos, puntuaciones.

Teniendo cada reseña dividida en oraciones, y éstas en palabras, es posible etiquetar cada una de ellas con su función sintáctica, gracias a la función *pos_tag_sents*, también de NLTK. El objetivo de este paso es, sabiendo la función que ocupa una palabra en una oración, poder hallar su lema[7].

El conjunto de etiquetas obtenido como resultado de la función *pos_tag_sents* está basado en el corpus del proyecto Penn Treebank [34], y no coincide con el que necesita el lematizador de WordNet [13]. Por ejemplo, Treebank tiene etiquetas como “JJ”, “JJR”, “JJS” para distintos tipos de adjetivos -común, comparativo, superlativo-, mientras que WordNet engloba a los adjetivos en una única categoría, “ADJ”. Por ende, se realizó un re-etiquetado para que la entrada de la función de lematización sea compatible. Finalmente, se lematizaron las palabras etiquetadas, reduciendo aún más el espacio de palabras, y se descartaron aquellas con menos de 3 caracteres.

[4] Conjunto de técnicas y herramientas que permiten extraer información de sitios web
[5] https://pandas.pydata.org/
[6] Natural Language Toolkit, plataforma estándar para el procesamiento del lenguaje natural en Python
[7] Palabra que se acepta como representante de todas las formas flexionadas de una misma palabra, Wikipedia

Un análisis de frecuencia de palabras mostró que algunas palabras con significado importante para el objetivo del trabajo quedaban fuera del corpus TASA (sec. 1.3.1). La ausencia de estas palabras se debe a la utilización de acrónimos (*ocd*), abreviaturas (*mgs*, *omg*, *pdoc*), errores tipográficos comunes (*perscribed*, *trazadone*), palabras que no fueron lematizadas correctamente -probablemente por un etiquetado no consistente de su función en una oración- (*depressed*, *inhibitors*), o simplemente palabras comunes que no formaban parte del corpus TASA (*downside*, *grogginess*). Para no perder la riqueza semántica que aportaban estas palabras, se realizó un reemplazo manual de las mismas por sus versiones expandidas (*obsessive compulsive disorder*, *miligrams*, *god*, *psychiatrist*), sus correcciones (*prescribed*, *trazodone*, *depress*, *inhibitor*), y sinónimos sí incluidos en TASA (*disadvantage*, *dazed*).[8]

El último paso del preprocesamiento del texto consistió en remover de las reseñas una lista de palabras compilada manualmente [29]. Esta lista siguió, a su vez, el mismo preprocesamiento (i.e., fue lematizada), y tiene una longitud de 6.460 palabras. Son palabras que resultan superfluas al análisis porque, por un lado, pueden no tener significado y, por otro, su significado es intrínseco a lo que se está reportando y no agrega valor (por ejemplo, es muy probable que en una reseña sobre *bupropion* se mencione a la sustancia, y considerarla a la hora de analizar la semántica no tiene sentido).

La lista está conformada por los siguientes tipos de palabras:

- Marcas de las drogas (correctas y con typos)
- Nombres de otras sustancias psicoactivas
- Variaciones de *slang*[9]
- Términos relacionados a vías de administración de las drogas
- Stopwords[10]

Todos los pasos anteriores sirvieron para reducir la cantidad de palabras que se ubicaron en un espacio multidimensional. El preprocesamiento de un texto admite variaciones en cada uno de los pasos (hacer *stemming*[11] en lugar de lematización, no descartar *stopwords*, etc.), pero se siguió la metodología estándar (encontrada en gran parte de la literatura) aplicando, donde correspondieran, ajustes pertinentes para la presente aplicación (remoción de palabras referidas a la jerga drogadependiente, análisis de frecuencia sobre el texto para encontrar casos no considerados en los pasos automáticos, etc.).

2.2. Análisis semántico de las reseñas

2.2.1. Creación del espacio semántico

2.2.1.1. TASA

Como se aclaró en la introducción (sec. 1.3.1), se utilizó un espacio LSA creado a partir del corpus TASA. Primero, se reemplazaron todas las palabras que formaban parte de un

[8] Las palabras son simplemente ejemplos y no conforman la exhaustiva lista utilizada en el trabajo
[9] Lunfardo
[10] Palabras sin significado como artículos, pronombres, preposiciones
[11] Reducir una palabra a su raíz removiendo la terminación. por ejemplo, dejar “gat” en vez de gata/o.

relato por sus vectores TASA (siempre y cuando la palabra formara parte del corpus); luego, para cada uno, se calculó el vector promedio, y se lo normalizó. De este modo, se obtuvo para cada reseña un vector que representa su semántica asociada, en un espacio de 300 dimensiones.

De las 17.154 palabras únicas encontradas a través de todas las reseñas, 9.780 formaban parte del corpus TASA, y 17.154 no. Hubo diversas maneras de agrupar los reportes y obtener en última instancia la similitud semántica que pudiera ser comparada frente a las similitudes estructural y de afinidad. Las distintas agrupaciones serán explicadas en la sección 2.2.2.

2.2.1.2. LSA

Luego del preprocesamiento de los textos ya explicados, se procedió a crear un espació semántico mediante la técnica de LSA (sec.1.3.1), haciendo uso de las herramientas brindadas por la librería **sklearn** de Python.

Se decidió [35] retener las palabras que aparecieran en por lo menos el 5 % de las reseñas, y que no aparecieran en más del 95 % de las mismas. De este modo, se eliminaron palabras que no aportaban información al corpus, ya sea por aparecer muy poco (por ejemplo, errores de escritura, o palabras que no sirven al análisis contextual de la técnica por su escasa frecuencia), o por aparecer demasiado (*stopwords* que no hayan sido detectadas en pasos previos). De este modo, de las 666.692 palabras originales del corpus (13.598 únicas), quedaron 6.772 palabras únicas, apareciendo 508.624 veces originalmente, y cubriendo por lo tanto el 76,3 % de las apariciones.

Se agruparon las reseñas en una colección de *documentos*, donde cada uno representa a una droga (se analizaron 30 drogas [12]). Luego, se creó a partir de ellos la matriz TF-IDF[13] (**TfidfVectorizer** de la librería **sklearn**). Dicha matriz especifica, para cada documento, la cantidad de apariciones de las palabras del corpus en el mismo (por lo que sus dimensiones son 6.722×30).

Sobre la matriz TF-IDF se realizó la decomposición en valores singulares o SVD, por sus siglas en inglés (**TruncatedSVD** de la misma librería), que es una operación de reducción de dimensionalidad. El objetivo de la transformación es preservar la información relevante utilizando un número menor de dimensiones, de modo de optimizar el valor informativo de la representación (y, por lo tanto, el cómputo y espacio). De este modo, de las 6.722 componentes originales de la matriz, se conservan únicamente las primeras 20[14], como manera de reducir el espacio semántico original de las reseñas.

Con esta matriz, se tienen ya las características [15] más importantes que representan a las distintas drogas. Acto seguido, se procedió a realizar las distintas comparaciones presentadas en la sección 2.2.2 Experimentación.

Como se encuentra detallado en dicha sección, agrupando en documentos los textos se transformó este nuevo conjunto de documentos mediante la matriz TF-IDF original

[12] Solo se utilizaron drogas prescriptas para alguna condición que pudiera ser agrupada en una *supercategoría* del DSM-V, tal como se explica en la sección 2.2.2.3

[13] Del inglés, Term frequency, Inverse document frequency

[14] Estas 20 componentes explican el 97.47 % de la varianza de la matriz TF-IDF original, por lo que la pérdida de información es mínima

[15] Features

(obteniendo la representación de estos documentos en el espacio semántico creado a partir del corpus), y se aplicó también la descomposición SVD, manteniendo las 20 componentes principales. Se calculó la correlación de Pearson entre todos los pares de filas de la matriz final (donde los documentos forman las filas y las 20 componentes SVD las columnas), obteniendo por lo tanto la similitud semántica entre todos los pares de drogas.

Cabe destacar que es posible realizar el análisis LSA tomando como corpus de entrada cada uno de los grupos por los cuales se dividirán luego las reseñas para la comparación. Los resultados de este análisis no presentaron diferencias considerables respecto a los obtenidos aplicando LSA sobre el corpus entero.

2.2.2. Experimentación

2.2.2.1. Comparación entre todas las drogas, antipsicóticos y antidepresivos

La primer comparación fue entre todas las drogas, independientemente de su tipo.

2.2.2.2. Comparación intra-antipsicóticos e intra-antidepresivos

El paso siguiente fue realizar la comparación entre las diversas drogas, para cada grupo según su tipo. Para cada par de drogas antipsicóticas, se calculó su similitud semántica; ídem para los antidepresivos.

2.2.2.3. Comparación de drogas para una condición

Se agruparon las reseñas de los pacientes según la condición por la cual reportaron estar siendo tratados. Como este campo en el conjunto de datos no se encontraba estandarizado entre ambos sitios web (cada sitio ofrece su conjunto de opciones para elegir a la hora de realizar un reporte, además de la posibilidad de utilizar un campo libre), existía mucha variabilidad, la cual dificultaba el análisis, dado que se diluía la cantidad de reseñas para cada condición.

Por este motivo, se recurrió al índice de la última edición del DSM-V [1] para obtener super-conjuntos en donde agrupar la mayor cantidad de condiciones posibles. Dicha categorización fue validada y aprobada por una psiquiatra, y redujo la cantidad de condiciones de las 153 originales a 24.

2.2.2.4. Comparación de drogas para un subconjunto de condiciones

Se decidió analizar un subconjunto de patologías que se benefician de los efectos secundarios de este tipo de medicamentos y es por eso que son tratadas con ellos pero que, consideramos, no son condiciones que tengan un sesgo cognitivo que pueda modificar o alterar la semántica de los individuos que las ingieren.

Algunas de estas condiciones son pérdida de peso (que se puede lograr mediante **bupropion**), migrañas (**amitriptilina**, **nortriptilina**, **venlafaxina**), trastornos del ritmo circadiano (**amitriptilina**, **doxepina**, **mirtazapina**, **quetiapina**, **trazodona**), abuso de sustancias (**bupropion**), y otras.

Cálculo de la similitud semántica

Partiendo de las agrupaciones de reseñas definidas anteriormente, el cálculo de la similitud semántica entre las drogas se realizó usando la similitud coseno (sec. 1.3.2) tal como se explica a continuación. Para cada uno de los grupos, se determinaron las 5 drogas con mayor cantidad de reseñas cumpliendo la condición adicional de poseer más de 100 reseñas. Para el conjunto resultante, se consideró cada par de drogas (d_1, d_2) junto a los vectores de sus correspondientes reseñas. Se obtuvieron, por lo tanto, dos matrices M_{d_1} y M_{d_2} (con 300 columnas y una cantidad de filas equivalente a la cantidad de reseñas para cada droga) y se realizó el producto matricial entre cada una y la transpuesta de la otra, obteniendo dos nuevas matrices.

$$M_{d_1 d_2} = M_{d_1} \times M_{d_2} \tag{2.4}$$

$$M_{d_2 d_1} = M_{d_2} \times M_{d_1} \tag{2.5}$$

Hecho eso, se calculó el promedio entre ambas matrices. Este paso se debe a una cuestión de aritmética de punto flotante. Ambas matrices deberían ser en teoría iguales, pero al tener los vectores muchos valores cercanos a cero, a la hora de realizar cuentas con ellos se propaga el error y se llega a dos matrices significativamente distintas[16].

La matriz resultante es de dimensiones $\mathbf{N} \times \mathbf{M}$, siendo $\mathbf{N}$ la cantidad de reseñas de la droga d_1 y $\mathbf{M}$ la de d_2. Dado que el producto matricial de las ecuaciones 2.4 y 2.5 equivale al cálculo de la similitud coseno entre todos los pares de vectores que surgen de elegir uno de cada matriz, cada celda (i, j) de la matriz resultante representa la similitud semántica entre la reseña i de la droga d_1 y la reseña j de la droga d_2.

Se calculó el promedio de esa matriz, para obtener finalmente la medida de la similitud semántica para los reportes de todo par de drogas.

2.2.3. Comparación de las similitudes semántica, estructural y de afinidad

Como resultado de los análisis hasta aquí descriptos, se obtienen las similitudes semántica, estructural y de afinidad entre todos los pares de drogas, dispuestas en tres matrices distintas. Como hay algunas drogas para las cuales no se tiene parte de esta información (en particular, los datos de afinidad), se calculó la intersección entre los tres conjuntos. Un filtro adicional consistió en pedir que cada par de drogas tuviera por lo menos el 60 % de sus receptores con datos de afinidad en común[17].

Luego, para cada una de las matrices de correlación, se removieron los pares de drogas fuera del rango $\mu \pm 2\sigma$[18] para eliminar outliers. De este modo, se crearon tres máscaras booleanas (donde los pares de drogas fuera de ese rango aparecen como **Falso**, y **Verdadero** si se encuentran dentro del rango), luego se calculó la conjunción lógica entre ellas, y el resultado se aplicó a las tres matrices de correlación originales. Así, quedan en las

[16] Se probó también saturar (es decir, poner en cero valores debajo de un umbral arbitrariamente pequeño) las matrices originales, con un resultado comparable obtenido a partir del promedio entre ambas.

[17] Caso contrario, podía ocurrir que dos drogas tuvieran una correlación aparentemente alta por coincidir en un único receptor, pero que cada una de las drogas, por su parte, afectasen a receptores no comunes a ambas, por lo que su correlación no debía ser efectivamente tan alta

[18] Conservando por lo tanto el 95.45 % del conjunto de datos, por la *regla 68–95–99.7*

matrices únicamente los pares de drogas que se encuentran dentro del rango en todas las matrices.

Dado que las matrices son simétricas, se conserva la matriz triangular superior de cada una, y esos valores son los que se comparan entre sí.

3. RESULTADOS Y ANÁLISIS

Ningún número de experimentos puede probar que tengo razón; un solo experimento puede demostrar que no la tengo.
–Apócrifo

En este capítulo se presentarán los resultados obtenidos siguiendo los procedimientos descriptos en el capítulo 2. Para ello, se utilizan los siguientes tipos de visualización:

- Gráfico de dispersión: se utiliza para la comparación de dos tipos de similitudes distintas (entre semántica, estructural y de afinidad) para todos los pares de drogas.
- Mapa de calor: se utiliza para comparar los valores de similitud (semántica, estructural, o de afinidad) entre los pares de drogas.
- Nubes de palabras: muestran visualmente los términos asociados a las componentes principales de la matriz obtenida luego de la aplicación de LSA. El tamaño de las palabras está relacionado a su peso en la componente en cuestión.
- Grafos de comunidades: muestran la conectividad de las drogas en base a su similitud semántica, pudiendo agruparlas en comunidades según su modularidad Louvain [6]. Los gráficos fueron realizados utilizando la herramienta **gephi** [2].

Muchos de los gráficos obtenidos no serán incluidos en la presentación principal de los resultados, por no ser útiles a efectos de nuestro análisis; dichos gráficos se presentarán, por completitud, en la sección correspondiente (sec. 4.5 Gráficos adicionales) del apéndice.

Referencias de las imágenes

Las siguientes imágenes hacen referencia a los marcadores utilizados en los gráficos de dispersión. Cada punto equivale a la correlación (de la métrica que se esté midiendo en el gráfico) entre un par de drogas, donde pueden ser ambas antidepresivas, antipsicóticas, o mezcla de ambas (par fármaco antidepresivo - antipsicótico).

(a) Par antidepresivo *(b)* Par antipsicótico *(c)* Par mezclado

Fig. 3.1: Referencia para los gráficos de la sección de resultados

Por otro lado, los colores de los marcadores correspondientes a los pares antidepresivo, y antipsicótico son los utilizados para diferenciar a las drogas en las etiquetas de todos los gráficos (es decir, los nombres de fármacos antidepresivos aparecen en color azul).

Orden de las drogas en los ejes y su clasificación

Para buscar un orden de las drogas que tuviera sentido a la hora de mostrar los gráficos, se realizó primero un agrupamiento jerárquico (*hierarchical clustering*) sobre la matriz de

similitud estructural (ya que es la matriz para la cual se tienen todos los datos entre pares de drogas).

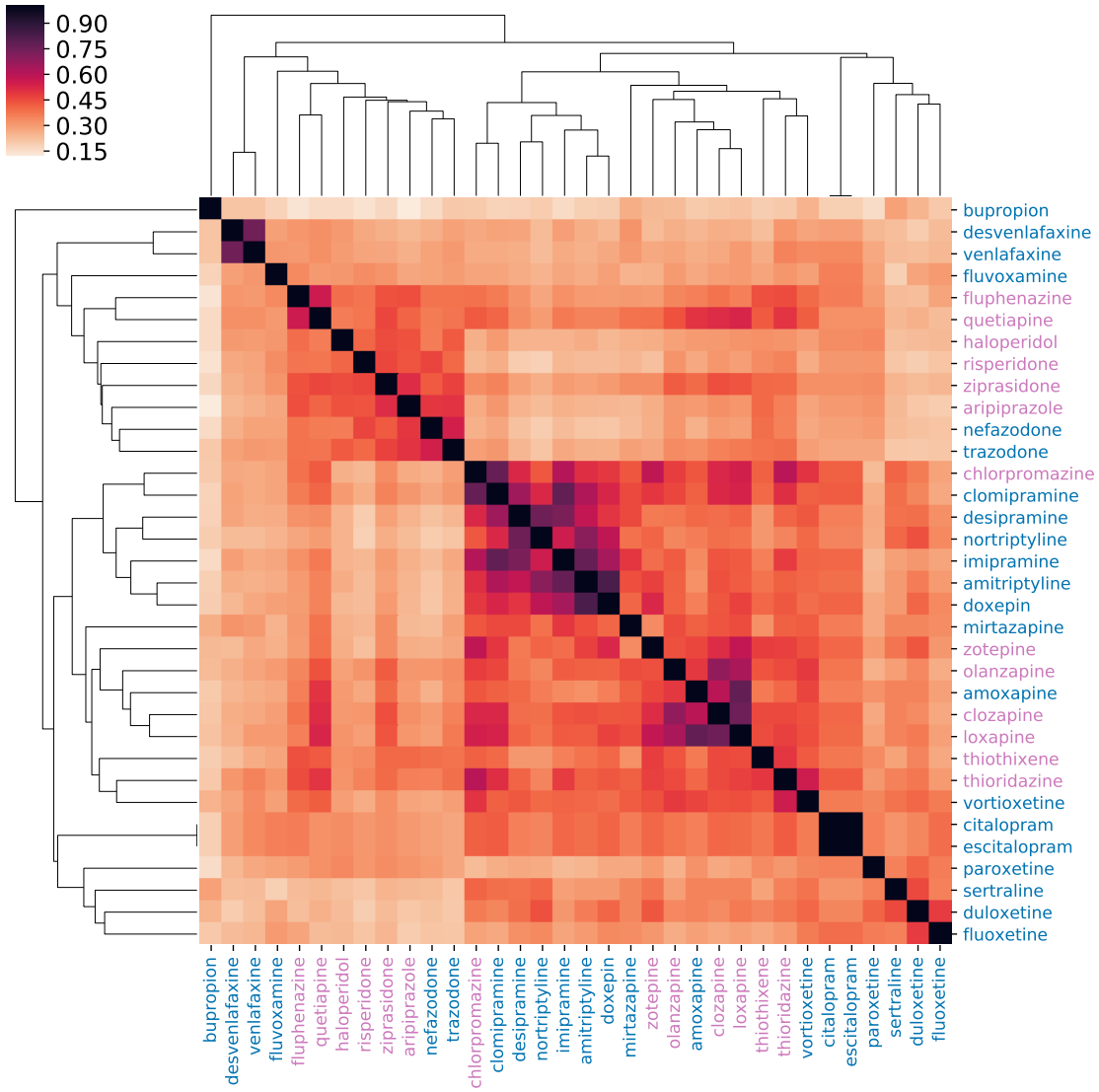


Fig. 3.2: Matriz de similitud estructural entre drogas y agrupamiento jerárquico.

El agrupamiento jarárquico ordenó a las drogas en grupos que se condicen, aproximadamente, con su clasificación (descriptas en 1.2 Farmacología).

Por ejemplo, los antipsicóticos atípicos quedaron en el grupo **quetiapina** ↔ **aripiprazol** (con la intrusión del **haloperidol**); los antidepresivos SSRI [1] en el grupo **vortioxetina** ↔ **sertralina** (con la falta de la **fluvoxamina**), los antidepresivos TCA [2] en el grupo **clomipramina** ↔ **doxepina** (con la falta de la **amoxapina**).

Este orden obtenido se utilizó como base para el final, que fue el siguiente:

[1] Inhibidor Selectivo de la Recaptación de Serotonina
[2] Tricíclicos

1. Quetiapina: antipsic. atípico
2. Risperidona: antipsic. atípico
3. Ziprasidona: antipsic. atípico
4. Aripiprazol: antipsic. atípico
5. Zotepina: antipsic. atípico
6. Olanzapina: antipsic. atípico
7. Clozapina: antipsic. atípico
8. Haloperidol: antipsic. típico
9. Loxapina: antipsic. típico
10. Tiotixeno: antipsic. típico
11. Tioridazina: antipsic. típico
12. Clorpromazina: antipsic. típico
13. Trazodona: antidep. SARI[3]
14. Nefazodona: antidep. SARI
15. Mirtazapina: antidep. TeCA[4]
16. Bupropion: antidep. NDRI[5]
17. Desvenlafaxina: antidep. SNRI[6]
18. Venlafaxina: antidep. SNRI
19. Duloxetina: antidep. SNRI
20. Fluvoxamina: antidep. SSRI
21. Citalopram: antidep. SSRI
22. Escitalopram: antidep. SSRI
23. Sertralina: antidep. SSRI
24. Paroxetina: antidep. SSRI
25. Fluoxetina: antidep. SSRI
26. Vortioxetina: antidep. SMS [7]
27. Clomipramina: antidep. TCA [8]
28. Desipramina: antidep. TCA
29. Nortriptilina: antidep. TCA
30. Imipramina: antidep. TCA
31. Amitriptilina: antidep. TCA
32. Doxepina: antidep. TCA
33. Amoxapina: antidep. TCA

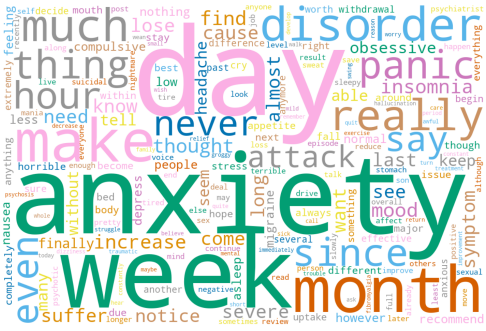
Se aclara que en los ejes de las imágenes las drogas aparecen en orden inverso, y se presentará un subconjunto ordenado de las mismas[9].

[3] Antagonista e Inhibidor de la Recaptación de Serotonina
[4] Tetracíclico
[5] Inhibidor de la Recaptación de Dopamina y Noradrenalina
[6] Inhibidor Selectivo de la Recaptación de Serotonina y Noradrenalina
[7] Modulador y estimulador de la Serotonina
[8] Tricíclico
[9] No todas las drogas fueron utilizadas en los distintos análisis, por razones de cantidad de datos

3.1. Componentes SVD de LSA aplicado al corpus y términos asociados a las mismas

De la matriz SVD que surge como resultado del procedimiento LSA se obtuvieron 20 componentes, cada una de ellas de tamaño 6.772 (la cantidad de palabras distintas del corpus). Las mismas explican el 97.47 % de la varianza encontrada en la matriz TF-IDF original (tal como se explicó en la sección 2.2.1.2 LSA).

Las componentes pueden ser ordenadas de mayor a menor y, de ese modo, distinguir cuáles son los términos con *mayor peso* en cada una de ellas. En los siguientes gráficos se muestran los resultados de este análisis, eligiendo solo las primeras 10 componentes, las cuales explican el 78.38 % de la varianza (cada componente sucesiva se vuelve menos relevante en cuanto a la varianza explicada).



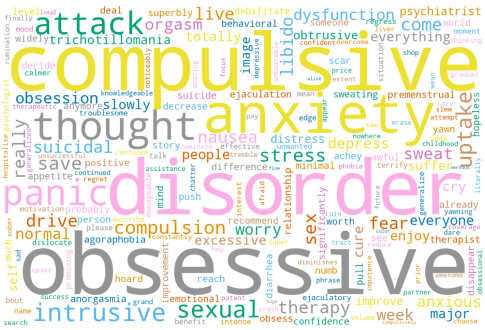
(a) Términos asociados a la primer componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 11.94 % de la varianza.

(b) Términos asociados a la segunda componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 10.80 % de la varianza.

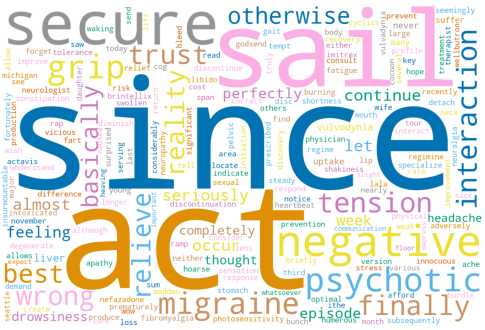
Fig. 3.3

Para la primer componente (fig.3.3a), uno de los términos con mayor peso es el de ansiedad (*anxiety*); consideramos que no casualmente otros términos importantes son de índole temporal (día -*day*-, semana -*week*-, y mes *month*). Es posible argumentar que entre las personas con trastorno (*disorder*) de ansiedad, las preocupaciones excesivas por cuestiones temporales son moneda corriente. Los ataques de pánico (*panic*) son también un síntoma de esta condición.

Los términos asociados a la segunda componente (fig.3.3b) toman ya otro carácter. El término con mayor peso, voz (*voice*), parece referir a la escucha (*hear*) de alucinaciones acústicas, un síntoma que está directamente ligado a trastornos (*disorder*) psicóticos (*psychotic*/*psychosis*) y esquizofrenia (*schizophrenia*). Incluso, se debate [31] sobre la existencia de psicosis con aspectos de neurosis obsesiva (*obsessive*).



(a) Términos asociados a la tercer componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 10.89 % de la varianza.



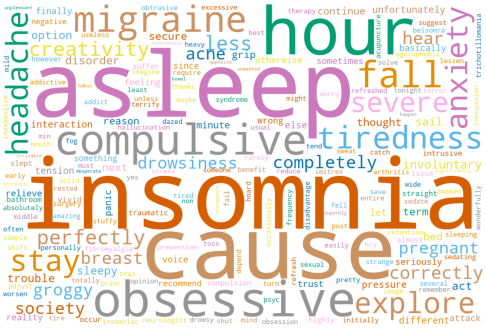
(b) Términos asociados a la cuarta componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 7.94 % de la varianza.

Fig. 3.4

La tercer componente (fig.3.4a) parece estar ligada al trastorno (*disorder*) obsesivo (*obsessive*) compulsivo (*compulsive*). Características como la ansiedad (*anxiety*) generada por la repetición de pensamientos (*thought*) intrusivos (*intrusive*), ataques (*attack*) de pánico (*panic*), están relacionadas a la condición. Los antidepresivos inhibidores selectivos de la recaptación (*uptake*) de la serotonina, además de usarse para la depresión, son también utilizados [17] [33] [32] (aunque en mayores dosis) para tratar el trastorno obsesivo compulsivo.

De la cuarta componente (fig.3.4b) resulta difícil extraer algún significado similar a los anteriores, por tratarse mayormente de palabras que podrían ser utilizadas en diversos contextos (desde -*since*-, actuar -*act*-, seguro -*secure*-, negativo *negative*, interacción -*interaction*-, etc.). Por ejemplo, el término *sail* es utilizado en la expresión *smooth sailing* (teniendo en cuenta la lematización del texto explicada en 2.1.3.2 Preprocesamiento del texto), para decir que algo progresa sin impedimentos ni dificultad[10]; es decir, es aplicable a cualquier tipo de tratamiento farmacológico que surte efecto.

(a) Términos asociados a la quinta componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 7.62 % de la varianza.



(b) Términos asociados a la sexta componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 6.90 % de la varianza.

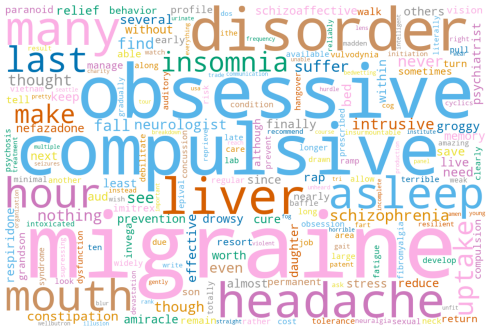
Fig. 3.5

La quinta componente incluye tópicos referidos a migrañas (*migraine*), como dolores de

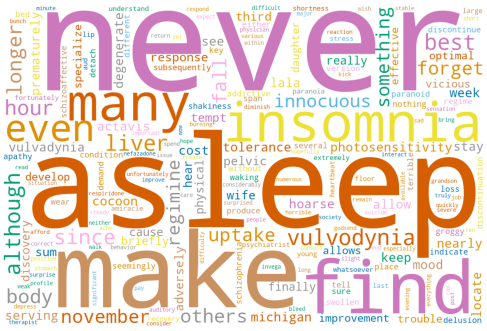
[10] Definición de Merriam Webster

cabeza (*headache*), cansancio (*tiredness*), la posibilidad de alucinaciones auditivas (*hear*, *voice*) presente en las migrañas con aura, y un adjetivo que se suele utilizar para describir a la migraña/dolor de cabeza, como severo (*severe*).

La explicación de estas primeras componentes es solo una interpretación posible, razón por la cual no se continúa con el análisis para el resto de las componentes - sumado a no poder encontrar una relación tan directa entre los términos como hasta ahora. No obstante, cumple el objetivo de ejemplificar cómo es posible encontrar un tipo de orden en el corpus elegido, y que la elección de la técnica LSA parecería haber sido acertada.

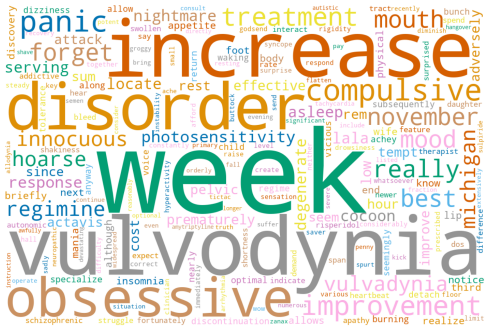


(a) Términos asociados a la séptima componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 6.51 % de la varianza.



(b) Términos asociados a la octava componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 5.94 % de la varianza.

Fig. 3.6



(a) Términos asociados a la novena componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 5.18 % de la varianza.

(b) Términos asociados a la décima componente de la matriz SVD obtenida por el proceso de LSA. Esta componente explica el 4.66 % de la varianza.

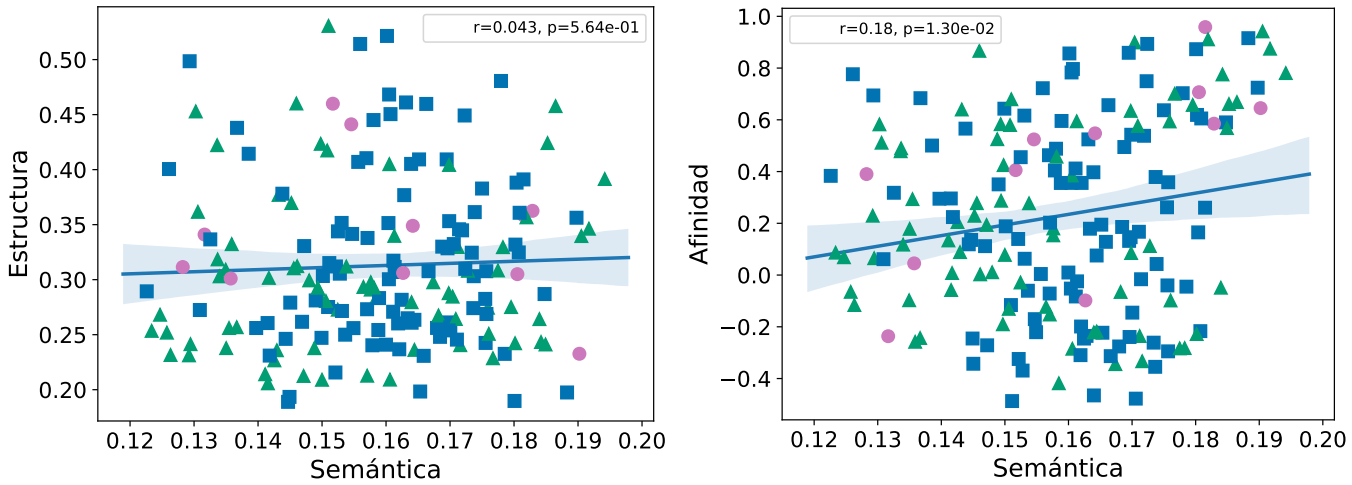
Fig. 3.7

3.2. Comparación entre todos los pares de drogas

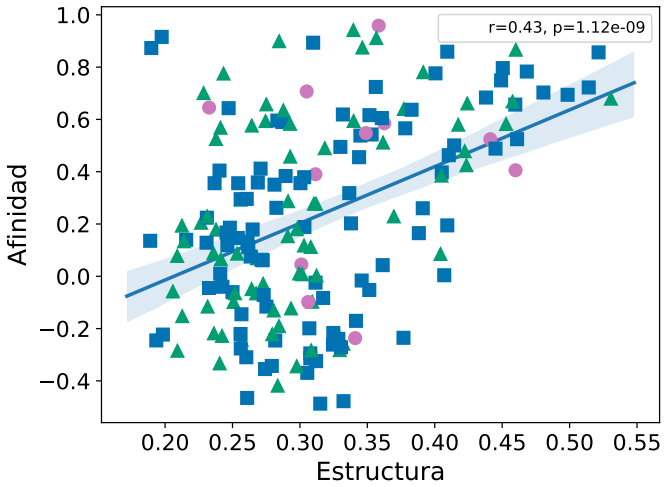
En esta sección se muestran los resultados de la comparación entre todos los pares de drogas, tanto antipsicóticos como antidepresivos.

3.2.1. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)

En esta sección se muestran los resultados de comparar todos los pares de drogas, utilizando la similitud semántica obtenida a partir del corpus TASA (sec.2.2.1.1).



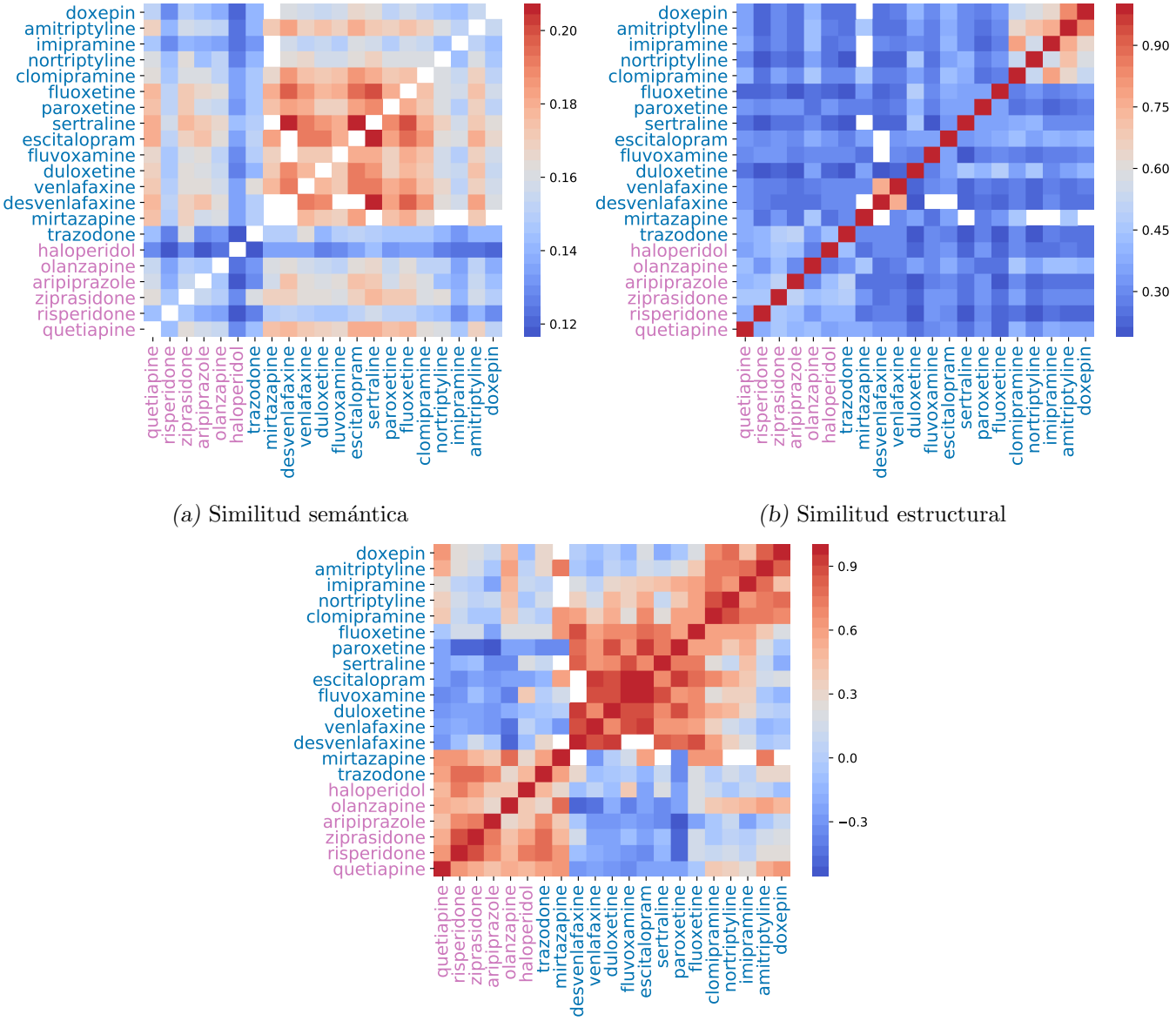
(a) Comparación entre similitudes semántica y estructural. *(b)* Comparación entre similitudes semántica y de afinidad.



(c) Comparación entre similitudes estructural y de afinidad.

Fig. 3.8: Comparación de similitudes para todo par de drogas del corpus. Cada punto representa un par de drogas, y en los ejes se indican sus similitudes estructural, semántica y de afinidad. La similitud semántica surge de la utilización del corpus TASA. La línea, obtenida mediante regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

Analizando la comparación entre las similitudes semántica y estructural (fig.3.8a), no resulta posible elaborar conclusiones dado que el p-valor no resulta significativo. Entre semántica y afinidad (fig.3.8b), sin embargo, parece haber una incipiente correlación positiva. Por último, entre estructura y afinidad (fig.3.8c), la correlación positiva es clara.



(a) Similitud semántica

(b) Similitud estructural

(c) Similitud de la afinidad

Fig. 3.9: Comparación de similitudes entre todos los pares de drogas del corpus. La similitud semántica surge de la utilización del corpus TASA.

Comparando los valores de similitud para cada una de las características (fig.3.9), se puede observar que la diferenciación más marcada ocurre para la comparación de similitud de afinidad (fig.3.9c) entre antipsicóticos vs. antidepresivos. Es decir, los antipsicóticos presentan claramente mayor similitud de afinidad entre sí que al ser comparados con los antidepresivos, y viceversa. Dentro del grupo de antidepresivos, además, es posible

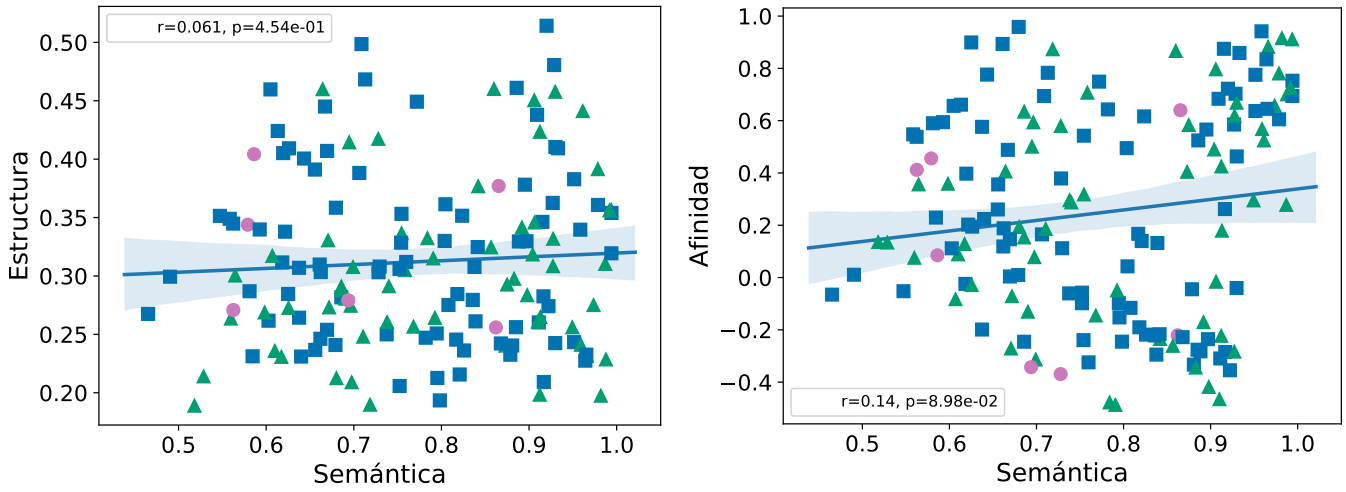
diferenciar a los de clasificación TCA (sec.3) de los SSRI/SNRI. Cabe destacar que, al menos en lo que afinidad respecta, **trazodone** y **mirtazapine** parecen comportarse como antipsicóticos.

Esta consistencia interna a la clase de fármaco se encuentra menos marcada al comparar la similitud estructural (fig.3.9b), donde si bien los valores son menores en términos absolutos, sigue habiendo diferencias para la comparación entre antipsicóticos (de menor consistencia interna) y antidepresivos (más consistentes entre sí, y especialmente bien marcada la diferencia para los TCA).

Por último, al analizar la similitud semántica (fig.3.9a), parece haber una consistencia semántica interna para los antidepresivos; sobre todo en los SSRI, y exceptuando los TCA y la **trazodona**.

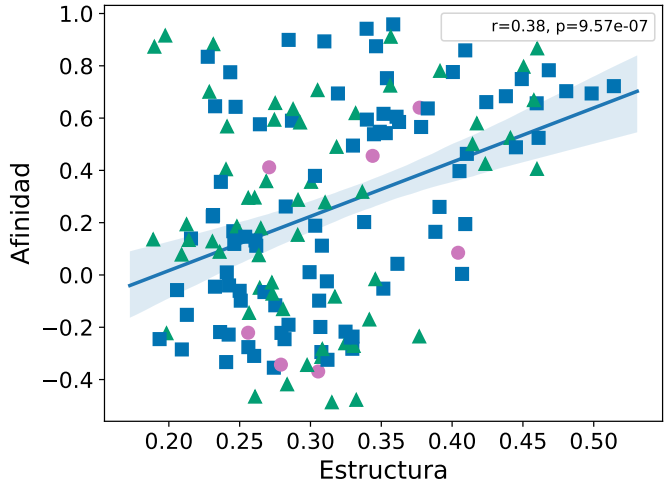
3.2.2. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)

En esta sección se muestran los resultados de comparar todos los pares de drogas, utilizando la similitud semántica obtenida a partir de aplicar la técnica de LSA en el corpus de discurso libre sobre tratamiento con fármacos (sec. 2.2.1.2).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.10: Comparación de similitudes para todo par de drogas del corpus, parte 1. La similitud semántica surge del análisis mediante la técnica LSA. La línea, obtenida mediante regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.



(a) Comparación entre similitudes estructural y de afinidad

Fig. 3.11: Comparación de pares de similitudes para todo par de drogas del corpus, parte 2. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

Los resultados se repiten: no es posible afirmar la existencia de una relación positiva entre semántica y estructura (fig. 3.10a).

Lo mismo ocurre entre las similitudes semántica y de afinidad (fig.3.10b), si bien la relación parece estar un poco mejor definida. La comparación entre estructura y afinidad brinda una esperable correlación positiva, dado que el cambio frente a la comparación anterior es únicamente en la dimensión de la semántica.

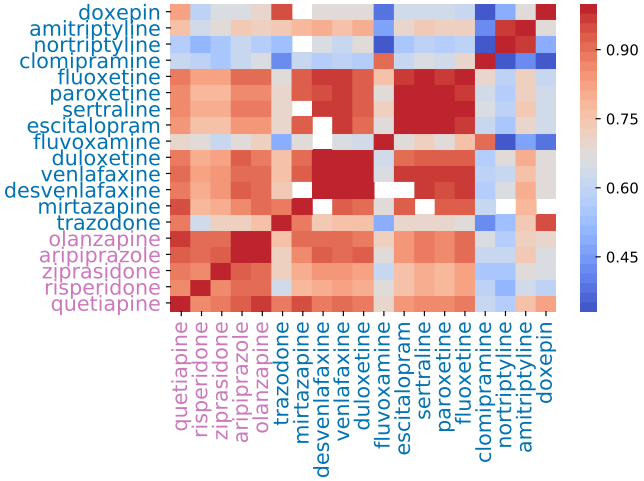


Fig. 3.12: Comparación de similitud semántica para todo par de drogas. La similitud semántica surge del análisis mediante la técnica LSA.

En el análisis utilizando la técnica LSA aplicada sobre el corpus de discurso sobre fármacos aparecen algunos cambios notorios (3.12). En este caso, las drogas antidepresivas presentan una muy alta similitud semántica positiva entre sí, al igual que el grupo de

antipsicóticas (aunque éstas presentan también correlación positiva con los antidepresivos SSRI). Los SNRI son muy consistentes entre sí, al igual que los SSRI (exceptuando la **fluvoxamina**), siendo los TCA los menos parecidos a todos.

La comparación entre similitudes estructural y de afinidad se presentan en el apéndice (c.4.2.1, sec.4.5, fig.4.3).

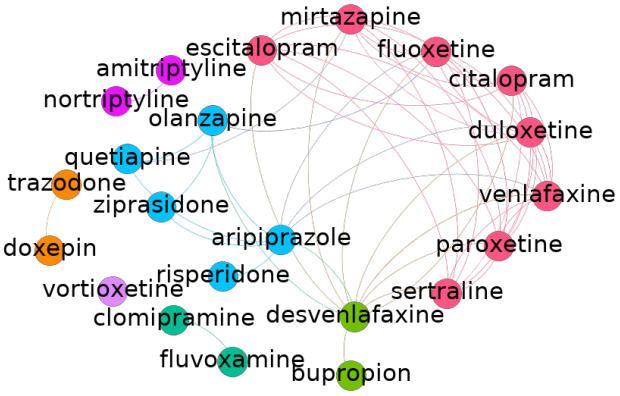


Fig. 3.13: Grafo subyacente a la matriz de similitud semántica para todo par de drogas. Se conservó el 22 % de los ejes con mayor peso.

Hacer un análisis de modularidad sobre el grafo que emerge de la matriz de similitud semántica (mostrada en la fig. 3.12), devuelve como resultado lo presentado en la figura 3.13. Se pueden apreciar 7 comunidades. Por un lado, los antipsicóticos (azul) quedan bien separados del resto. Los antidepresivos muestran un resultado menos claro. Los SSRIs (rojo) están bien agrupados, aunque **paroxetina** y **venlafaxine** pertenecen a los SNRIs (verde claro, que además aparece con el bupropion, que es NDRI). **Clomipramina** y **fluvoxamina** son TCA y SSRI respectivamente (pero se vio también que en la fig.3.12 aparecían relacionadas entre sí). **Amitriptilina** y **nortriptilina** están correctamente agrupadas (con la falta de la **clomipramina**, como ya se dijo, y la **doxepina**, incorrectamente asociada a la **trazodona**, que es un SARI). **Vortioxetina**, que es SMS, aparece solitaria de manera correcta.

3.2.3. Refinación de la comparación

Dado que los resultados expresados en estos gráficos se obtuvieron a partir de la comparación entre todos los pares de drogas sobre los cuales se disponían reseñas, fue difícil poder realizar un análisis minucioso sobre las mismas.

Por lo tanto, se decidió realizar una comparación circunscripta a los dos tipos de drogas: antipsicóticos por un lado, y antidepresivos por el otro, de manera de poder refinar el conjunto de drogas sobre el cual se trabajaba.

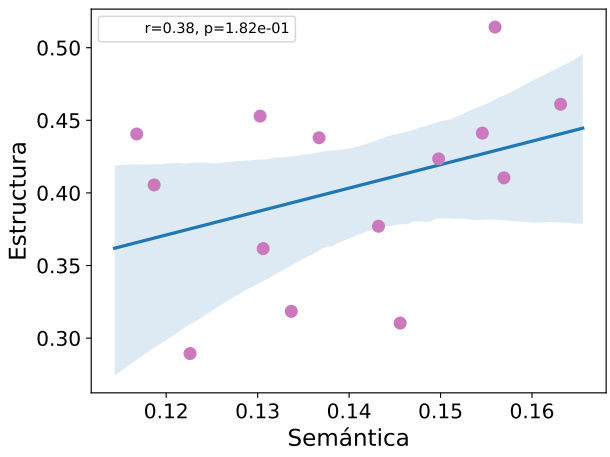
3.3. Intra-antipsicóticos e Intra-antidepresivos

En esta sección se muestran los resultados de la comparación entre todos los pares de antipsicóticos, y entre todos los pares de antidepresivos (los que tienen más de 100 reseñas, y los que tienen más de 1.000).

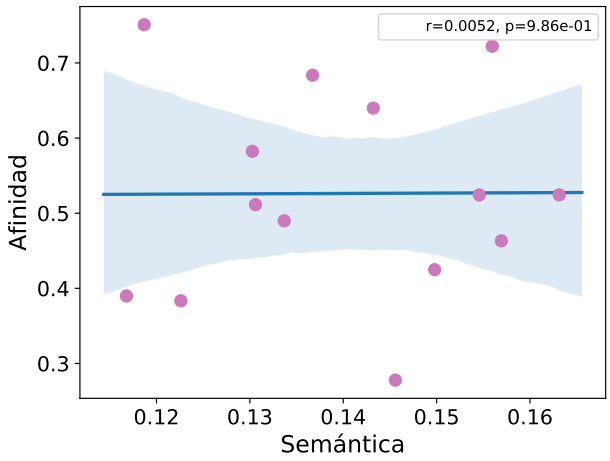
3.3.1. Comparación entre antipsicóticos

3.3.1.1. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)

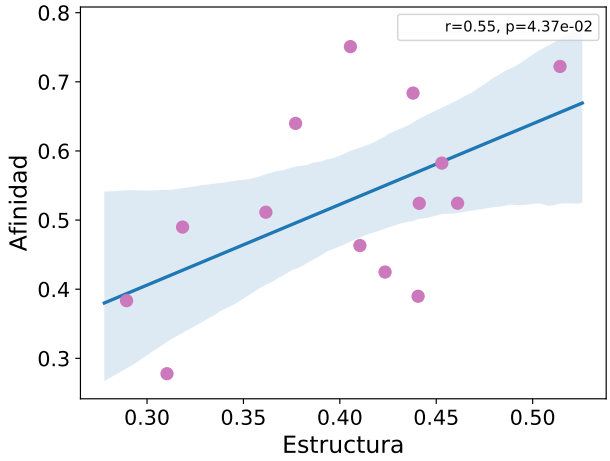
En esta sección se muestran los resultados de comparar todos los pares de antipsicóticos, utilizando la similitud semántica obtenida a partir del corpus TASA (sec.2.2.1.1).



(a) Comparación entre similitudes semántica y estructural



(b) Comparación entre similitudes semántica y de afinidad



(c) Comparación entre similitudes estructural y de afinidad

Fig. 3.14: Comparación similitudes para todo par de drogas antipsicóticas con más de 100 reseñas. La similitud semántica surge de la utilización del corpus TASA. La línea, obtenida mediante una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

Nuevamente, la comparación entre las similitudes estructural y de afinidad (fig.3.14c) muestra una correlación lineal positiva marcada (mejor aún que en la comparación anterior entre todos los tipos de drogas, fig.3.8c).

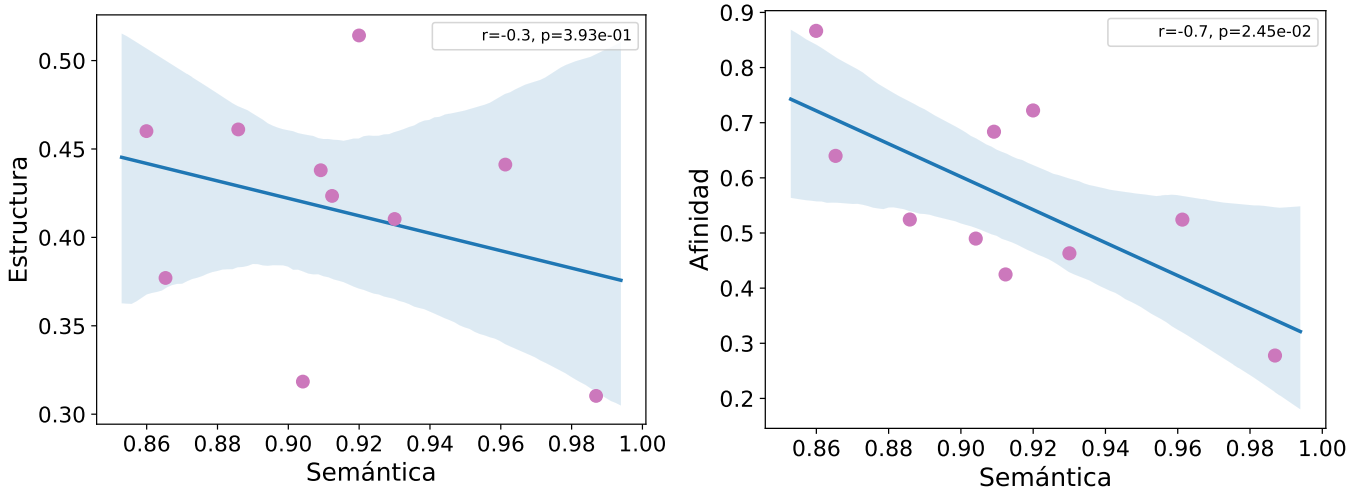
Las otras dos comparaciones, entre las similitudes semántica y estructural (fig.3.14a), y las similitudes semántica y de afinidad, no arrojan resultados significativos, por lo que no se puede decir nada sobre ellas.

Debido a que cambiaron los grupos a comparar, también lo hicieron sus medias y desvíos estándares, razón por la cual puede haber distintos pares de comparaciones (esta afirmación será válida para lo que resta del trabajo).

Las comparaciones de la similitud semántica, estructural, y de afinidad para los pares de drogas antipsicóticas visualizadas mediante los mapas de calor se presentan en el apéndice (c.4.2.1, sec.4.5, fig.4.4).

3.3.1.2. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)

En esta sección se muestran los resultados de comparar todos los pares de antipsicóticos, utilizando la similitud semántica obtenida a partir de aplicar la técnica de LSA en el corpus de discurso sobre tratamientos con fármacos (sec. 2.2.1.2).



(a) Similitud semántica

(b) Similitd de afinidad

Fig. 3.15: Comparación de similitudes para todo par de drogas antipsicóticas con más de 100 reseñas. La similitud semántica surge del análisis mediante la técnica LSA. La línea, obtenida mediante una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

Al observar la comparación entre las similitudes semántica y estructural (fig.3.15a), nuevamente no se puede concluir nada por el alto p-valor.

La comparación entre semántica y afinidad (fig.3.15b), en cambio, muestra un resultado inesperado, pues la correlación es negativa. Una razón posible del resultado adverso puede

ser que al estar trabajando con pocos datos (los antipsicóticos representan el 15 % del corpus) aumente el nivel de ruido.

La comparación entre similitudes estructural y de afinidad, y las comparaciones entre similitud semántica, estructural, y de afinidad para las drogas antipsicóticas analizadas mediante LSA se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.6 y 4.7, respectivamente).

Dado que hay un único antipsicótico atípico, no tiene sentido el análisis de comunidades sobre el grafo.

3.3.2. Comparación entre antidepresivos

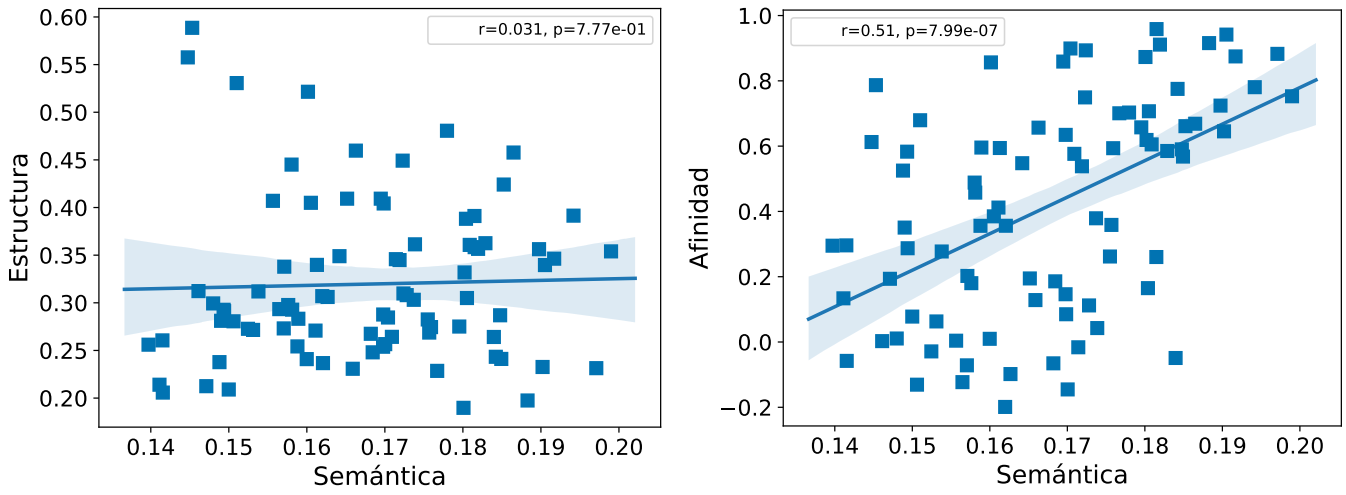
Debido al desbalanceo en el conjunto de datos (ver tabla en el apéndice 4.4) entre las drogas antipsicóticas y las antidepresivas, se decidió separar la comparación entre éstas en dos.

Por un lado, primero se evaluaron las drogas antidepresivas con más de 100 reseñas (resultando en 18 drogas para comparar), y luego, las que tenían más de 1.000 reseñas (resultando en 11 drogas).

Más de 100 reseñas por droga

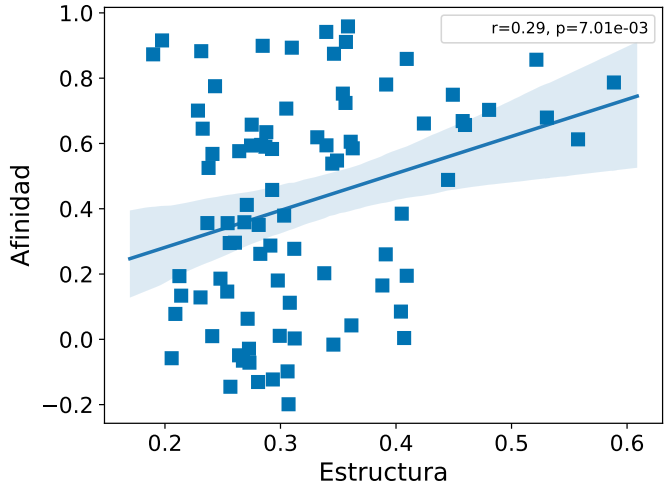
3.3.2.1. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)

En esta sección se muestran los resultados de comparar todos los pares de antidepresivos con más de 100 reseñas, utilizando la similitud semántica obtenida a partir del corpus TASA (sec.2.2.1.1).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.16: Comparación de similitudes para todo par de drogas antidepresivas con más de 100 reseñas cada una, parte 1. La similitud semántica surge de la utilización del corpus TASA. La línea, obtenida mediante una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.



(a) Comparación entre similitudes estructural y de afinidad

Fig. 3.17: Comparación de similitudes para todo par de drogas antidepresivas con más de 100 reseñas cada una, parte 2. La línea, obtenida mediante una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

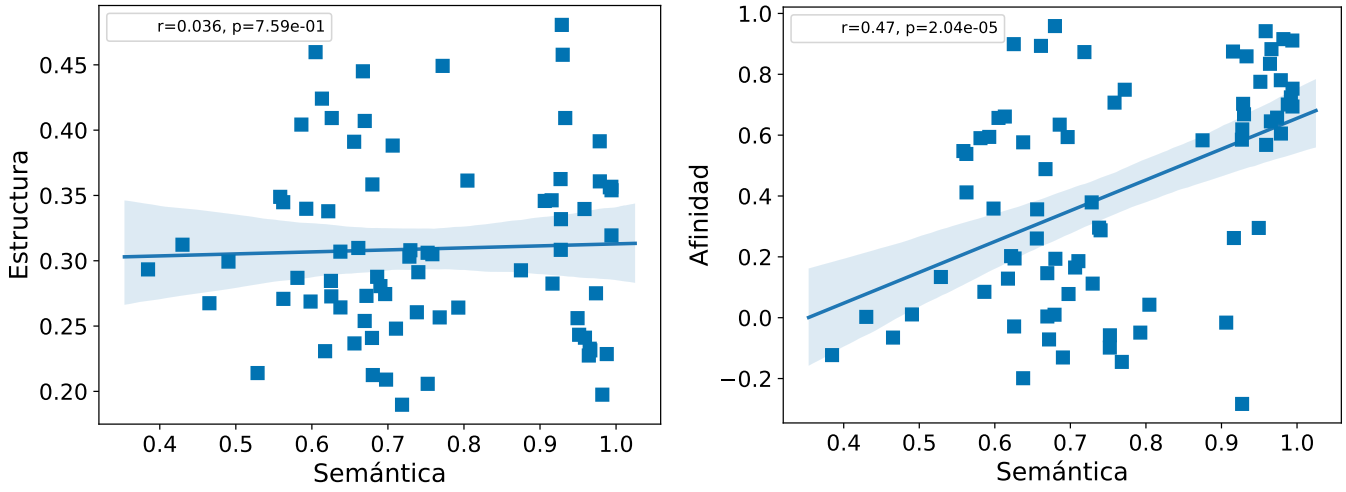
En las figuras 3.16 y 3.17 se repiten algunas de las observaciones pasadas: la comparación entre las similitudes semántica y estructural (fig. 3.16a) no muestra resultados significativos, mientras que en el caso de la comparación entre afinidad y semántica (fig. 3.16b), sí se puede ver una correlación positiva significativa entre ambas medidas.

Estructura y afinidad (fig.3.17a), como en la comparación entre todos los pares de drogas (fig.3.8c), se siguen mostrando relacionadas (aunque en menor medida).

Las comparaciones de similitud semántica, estructural, y de afinidad para las drogas antidepresivas analizadas mediante TASA se encuentran en el apéndice (c.4.2.1, sec.4.5, fig. 4.9).

3.3.2.2. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)

En esta sección se muestran los resultados de comparar todos los pares de antidepresivos con más de 100 reseñas, utilizando la similitud semántica obtenida a partir de aplicar la técnica de LSA en el corpus de discurso libre sobre tratamientos con fármacos (sec.2.2.1.2).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.18: Comparación de similitudes todo par de drogas antidepresivas con más de 100 reseñas. La similitud semántica surge del análisis mediante la técnica LSA. La línea, obtenida mediante regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

A diferencia de lo que sucedió en la comparación entre los pares de drogas antipsicóticas (fig. 3.15), aquí los resultados de las similitudes semánticas obtenidas mediante LSA (fig.3.18) del corpus sí son comparables a los obtenidos con TASA (fig.3.16). Se presentan las mismas relaciones entre las medidas de similitud analizadas: semántica y estructura no presentan una relación significativa, mientras que semántica y afinidad sí y, además, dicha relación es positiva.

La comparación entre similitudes estructural y de afinidad, y las comparaciones entre similitud semántica, estructural, y de afinidad para las drogas antipsicóticas analizadas mediante LSA se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.10 y 4.11, respectivamente).

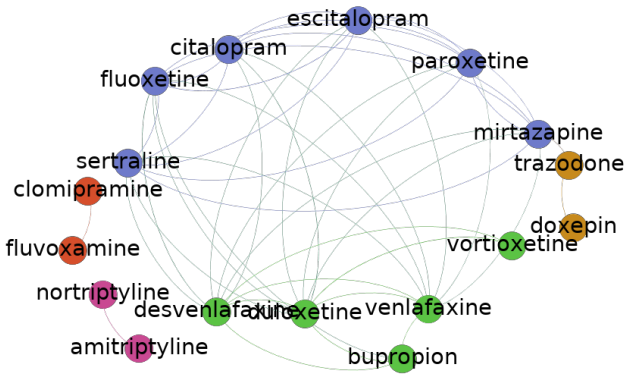


Fig. 3.19: Grafo subyacente a la matriz de similitud semántica para todo par de drogas antidepresivas. Se conservó el 34 % de los ejes con mayor peso.

En este caso, el análisis de modularidad sobre el grafo correspondiente a las similitudes

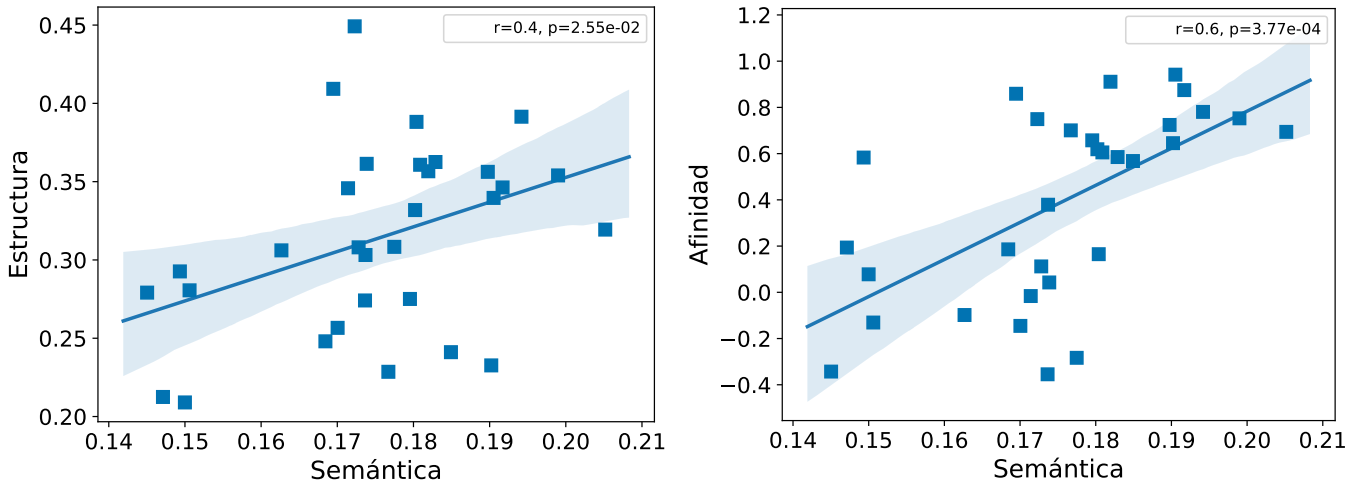
semánticas entre todo par de antidepresivos devuelve cinco comunidades. Podemos ver por un lado a los SSRI (en violeta) más la presencia de la **mirtazapina** (que es un tetracíclico); por otro lado, la comunidad de los SNRI está relativamente bien demarcada (en verde), aunque incluye también al **bupropion** (NDRI). La **doxepina** (TCA) y la **trazodona** (SARI) están una vez más incorrectamente relacionadas, al igual que la **fluvoxamina** (SSRI) y **clomipramina** (TCA). La **nortriptilina** y **amitriptilina**, ambas TCA, se encuentran bien agrupadas.

Más de 1000 reseñas por droga

Al reducir las drogas comparadas a las que cuentan únicamente con más de 1.000 reseñas, se obtienen valores significativos y mejores (en cuanto a la hipótesis de este trabajo) que considerando una mayor cantidad de drogas con menos reseñas.

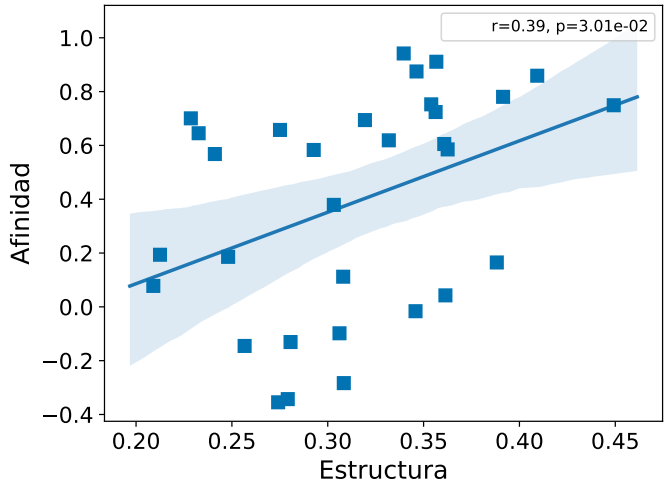
3.3.2.3. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)

En esta sección se muestran los resultados de comparar todos los pares de antidepresivos con más de 1.000 reseñas, utilizando la similitud semántica obtenida a partir del corpus TASA (sec.2.2.1.1).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.20: Comparación de similitudes para todo par de drogas antidepresivas con más de 1000 reseñas, parte 1. La similitud semántica surge de la utilización del corpus TASA. La línea, obtenida mediante regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.



(a) Comparación entre similitudes estructural y de afinidad

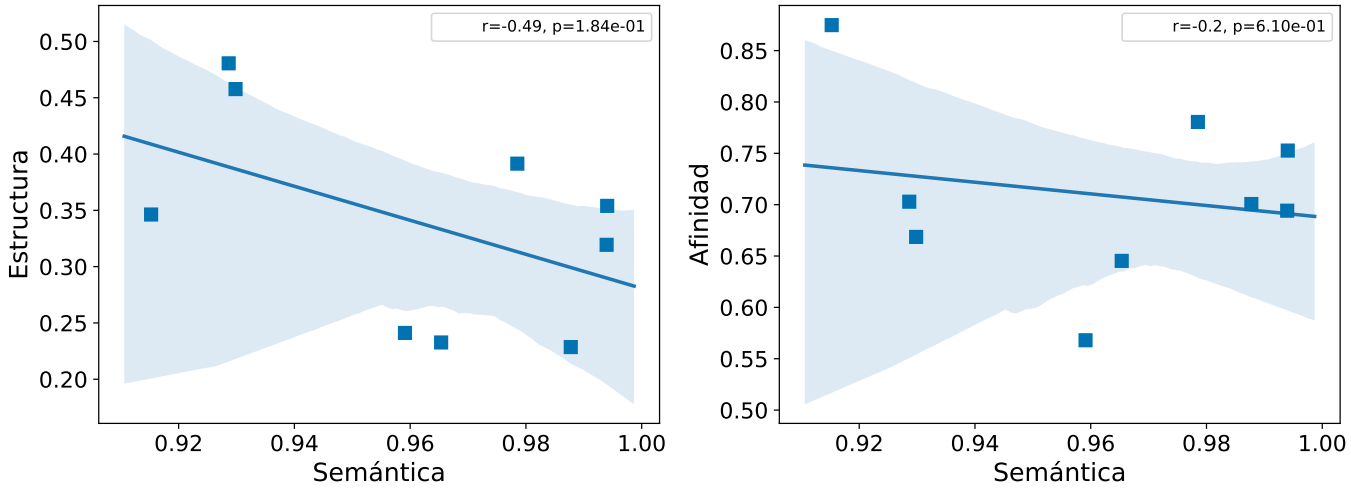
Fig. 3.21: Comparación de similitudes para todo par de drogas antidepresivas con más de 1000 reseñas, parte 2. La similitud semántica surge de la utilización del corpus TASA. La línea, obtenida mediante regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

En este caso, los resultados muestran una correlación positiva considerable y significativa entre todos los pares de características analizadas: semántica contra estructura (fig.3.20a), semántica contra afinidad (fig.3.20b), y estructura contra afinidad (fig.3.21a).

Las comparaciones de similitud semántica, estructural, y de afinidad para las drogas antidepresivas con más de 1.000 reseñas analizadas mediante TASA se encuentran en el apéndice (c.4.2.1, sec.4.5, fig. 4.12).

3.3.2.4. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)

En esta sección se muestran los resultados de comparar todos los pares de antidepresivos con más de 1.000 reseñas, utilizando la similitud semántica obtenida a partir de aplicar la técnica de LSA sobre el corpus de reseñas sobre fármacos (sec.2.2.1.2).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.22: Comparación de similitudes para todo par de drogas antidepresivas con más de 1000 reseñas. La similitud semántica surge del análisis mediante la técnica LSA. La línea, obtenida mediante regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

Para el análisis de este subconjunto de drogas con la técnica LSA no se obtiene un resultado significativo.

La comparación entre similitudes estructural y de afinidad, y las comparaciones de similitud semántica, estructural, y de afinidad para las drogas antidepresivas con más de mil reseñas analizadas mediante LSA se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.13 y 4.14, respectivamente).

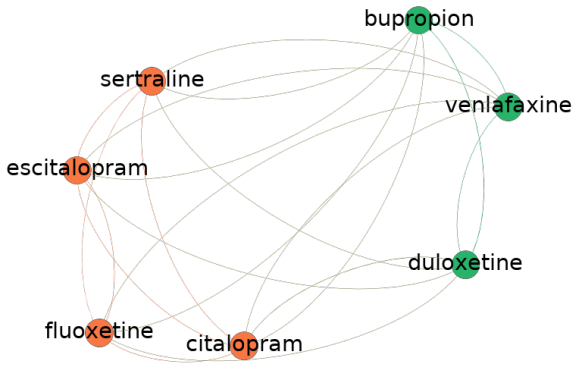


Fig. 3.23: Grafo subyacente a la matriz de similitud semántica para todo par de drogas antidepresivas con más de 1.000 reseñas. Se conservó el 100 % de los ejes.

En este caso, el análisis agrupó únicamente dos comunidades, debido a la poca cantidad de drogas. Por un lado, están los SSRI (en naranja), y por el otro, los SNRI (en verde) junto al **bupropion**.

3.3.3. Refinación de la comparación

Dado que los resultados siguen sin ser significativos en muchos casos, investigamos los resultados obtenidos refinando los grupos de drogas utilizados para las comparaciones.

Puesto que se está tratando con una gran cantidad de condiciones distintas, y con una gran cantidad de drogas, los resultados pueden ser muy variados. Cada persona parte de una química distinta, sumado a estar sufriendo una condición distinta, luego la variabilidad es relativamente grande, y por lo tanto la utilización de distintas drogas (para una misma condición) puede amplificar aún más este efecto.

Por esta razón se decidió agrupar a las reseñas según la condición (reportada) por la que se está tratando al paciente. Consideramos que esta refinación del conjunto de datos a analizar puede brindar resultados menos afectados por la variabilidad de los datos.

3.4. Comparación de drogas prescriptas para una misma condición

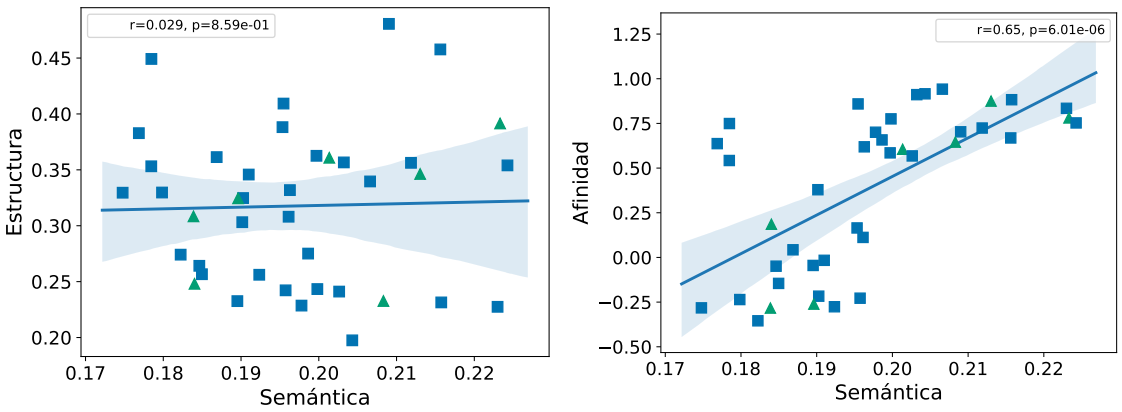
En esta sección se muestran los resultados de la comparación entre todos los pares de drogas prescriptas para una misma condición.

3.4.1. Trastorno de ansiedad

Esta sección comprende los pares de drogas prescriptas para el tratamiento del trastorno de la ansiedad.

3.4.1.1. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)

En esta sección se muestran los resultados de comparar todos los pares de drogas prescriptas para el tratamiento del trastorno de ansiedad, utilizando la similitud semántica obtenida a partir del corpus TASA (sec.2.2.1.1).



(a) Comparación entre similitudes semántica y estructural

(b) Comparación entre similitudes semántica y de afinidad

Fig. 3.24: Comparación de similitudes para todo par de drogas prescriptas para el tratamiento del trastorno de ansiedad. La similitud semántica surge de la utilización del corpus TASA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

La comparación en el contexto de drogas prescriptas para ansiedad se puede ver en la figura anterior (fig.3.24). Todas las drogas son antidepresivas, salvo un antipsicótico, **quetiapina**, utilizado para casos de depresión donde la medicación usual no surte efecto [11].

En lo que respecta a la comparación entre semántica y estructura (fig.3.24a), los resultados no son significativos, algo que se repitió varias veces en la comparación de estas dos características de las drogas.

Por otro lado, sí hay una correlación positiva y significativa al comparar las similitudes semántica y de afinidad (fig. 3.24b).

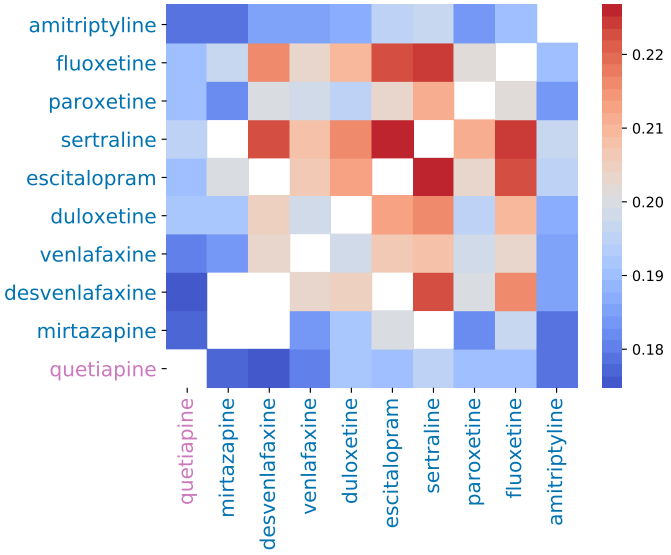


Fig. 3.25: Comparación de similitud semántica para todo par de drogas prescriptas para el tratamiento del trastorno de ansiedad. La similitud semántica surge de la utilización del corpus TASA.

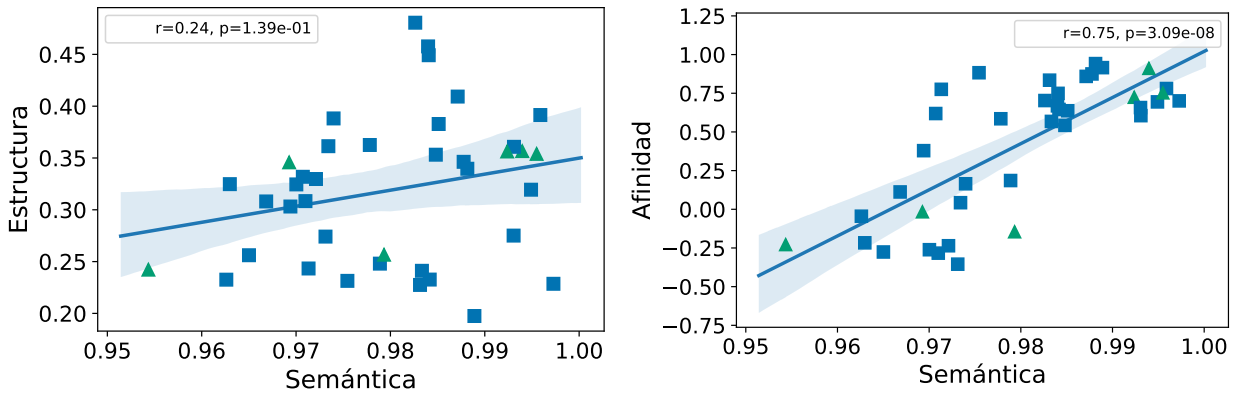
Si se compara la similitud semántica entre todas las drogas prescriptas para la ansiedad (fig.3.25), se ve nuevamente una clara distinción entre los antidepresivos, por un lado, y en particular **quetiapina**, el único antipsicótico presente en esta muestra. Dentro de los antidepresivos, se ven claras demarcaciones entre el grupo de los SSRI[11] (cuya semántica está muy correlacionada entre sí), luego los SNRI[12] en menor medida, y por último, antidepresivos como la *amitriptilina* (TCA) y *mirtazapina* (TeCA), que no resultan ser semánticamente similares a ninguno de los otros.

La comparación entre similitudes estructural y de afinidad, y las comparaciones de similitud estructural, y de afinidad para las drogas prescriptas para ansiedad analizadas mediante TASA se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.15 y 4.16, respectivamente).

[11] Fluoxetina, paroxetina, sertralina y escitalopram
[12] Duloxetina, venlafaxina y desvenlafaxina

3.4.1.2. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)

En esta sección se muestran los resultados de comparar todos los pares de drogas prescriptas para el tratamiento del trastorno de ansiedad, utilizando la similitud semántica obtenida a partir de aplicar la técnica de LSA en el corpus de reseñas sobre fármacos (sec.2.2.1.2).



(a) Comparación entre similitudes semántica y estructural

(b) Comparación entre similitudes semántica y de afinidad

Fig. 3.26: Comparación de similitudes para todo par de drogas prescripto para el tratamiento del trastorno de ansiedad. La similitud semántica surge del análisis mediante la técnica LSA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

El análisis de las drogas prescriptas para ansiedad mediante LSA muestra los mismos resultados que los de TASA: son no significativos para la comparación entre similitudes semántica y estructural (fig.3.26a), y un resultado mucho más marcado entre las similitudes semántica y de afinidad (fig.3.26b).

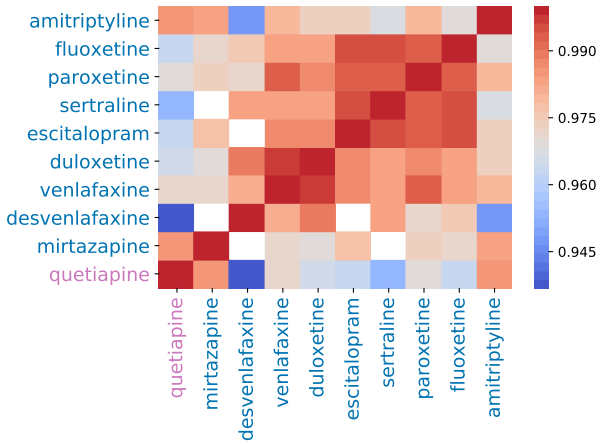


Fig. 3.27: Comparación de similitud semántica para todo par de drogas prescriptas para el tratamiento del trastorno de ansiedad. La similitud semántica surge del análisis mediante la técnica LSA.

Al igual que en la comparación hecha con el corpus TASA (fig.3.25), para el análisis con LSA aplicado al corpus de reseñas (fig.3.27) es posible separar por grupos en base a las similitudes semánticas, y notar que coincide con las clasificaciones de las drogas. Por un lado, el grupo de los antidepresivos SSRI[13]/SNRI[14] está bastante correlacionado entre sí, aunque resulta clara su distinción. Por otro lado, **mirtazapina**, **quetiapina** y **amitriptilina** son más parecidos (en particular los primeros dos) entre sí que al resto (y esto se sostiene comparando la similitud de afinidad, ver el apéndice, fig.4.18).

La comparación entre similitudes estructural y de afinidad, y las comparaciones de similitud estructural, y de afinidad para las drogas prescriptas para ansiedad analizadas mediante TASA se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.17 y 4.18, respectivamente).

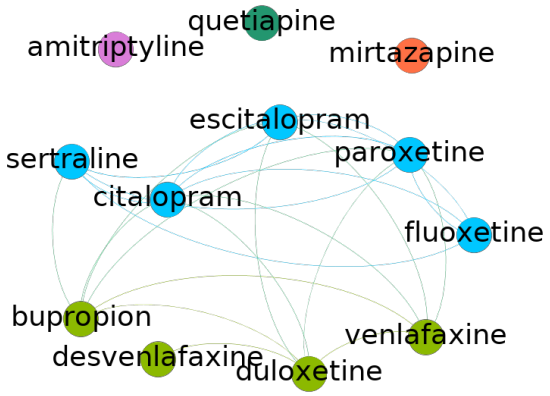


Fig. 3.28: Grafo subyacente a la matriz de similitud semántica para todo par de drogas prescriptas para trastornos de ansiedad. Se conservó el 36 % de los ejes con mayor peso.

En este caso, el análisis de modularidad sobre el grafo correspondiente a las similitudes semánticas entre todo par de drogas prescriptas para trastornos de ansiedad, devuelve cinco comunidades. Tres de las drogas aparecen inconexas de manera apropiada, **quetiapina** (antipsicótica), **amitriptilina** (TCA) y **mirtazapina** (TeCA). De los dos grupos restantes, una comunidad representa a los SSRI (en celeste), y la otra a los SNRI (en verde), con el agregado de **bupropion** (que es un NDRI).

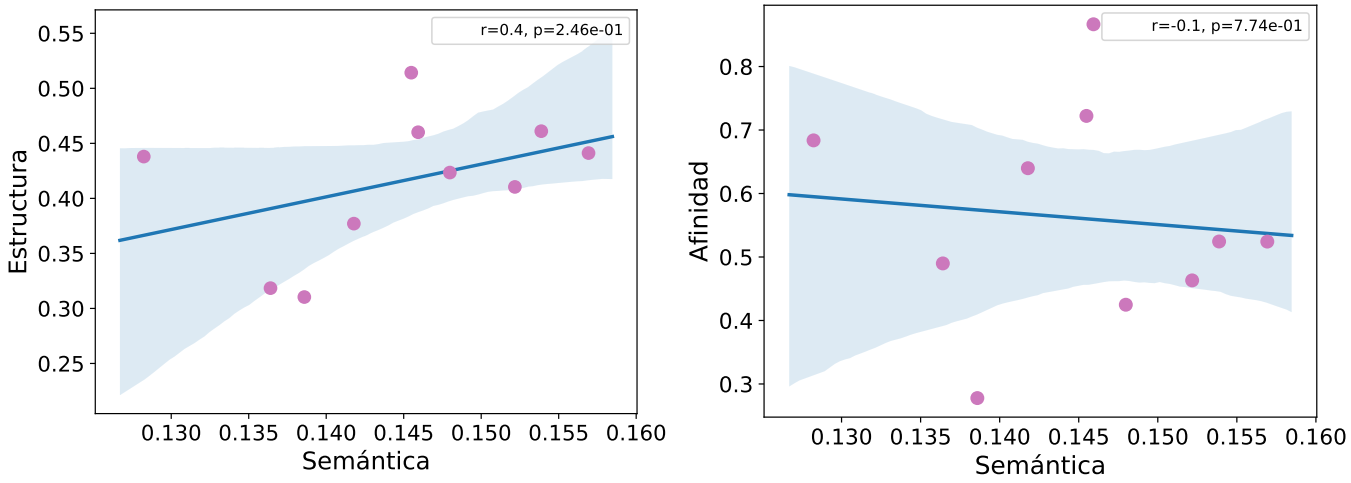
3.4.2. Trastorno bipolar

Esta sección comprende los resultados de la comparación de los pares de drogas prescriptas para el tratamiento del trastorno bipolar.

3.4.2.1. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)

En esta sección se muestran los resultados de comparar todos los pares de drogas prescriptas para el tratamiento del trastorno bipolar, utilizando la similitud semántica obtenida a partir del corpus TASA (sec.2.2.1.1).

[13] Fluoxetina, paroxetina, sertralina y escitalopram
[14] Duloxetina, venlafaxina y desvenlafaxina



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.29: Comparación de pares de similitudes para todo par de drogas prescripto para el tratamiento del trastorno bipolar. La similitud semántica surge de la utilización del corpus TASA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

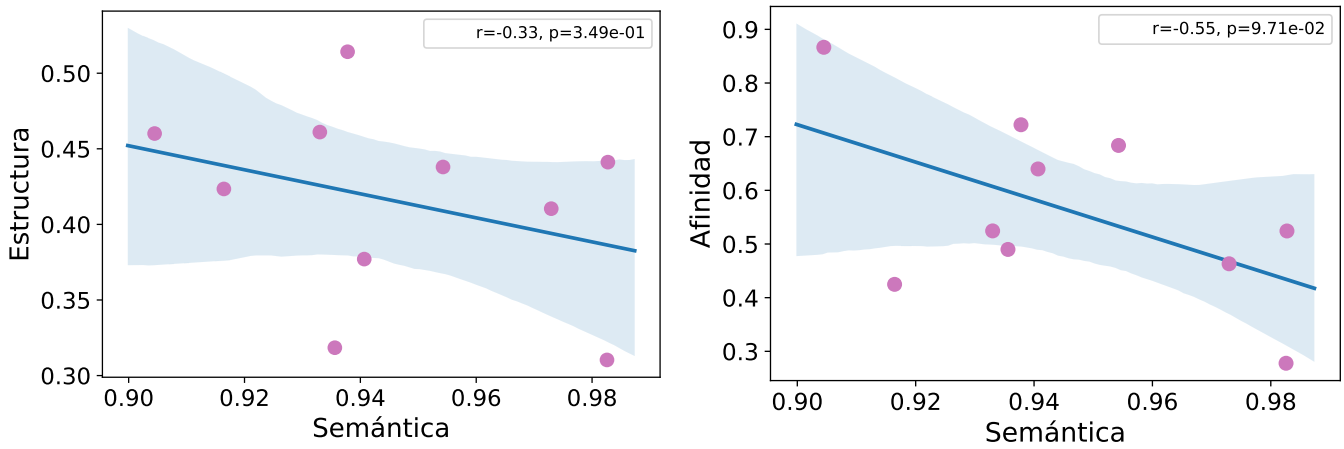
Para el tratamiento del trastorno bipolar las drogas a comparar son relativamente pocas (solo 5 pares). Esto lleva a que los resultados obtenidos al comparar sus similitudes semántica y de estructura (fig.3.29a), y de semántica y afinidad (fig.3.29b) no sean significativos.

Dado que las drogas prescriptas para el tratamiento del trastorno bipolar son todas antipsicóticas, resulta superfluo mostrar las comparaciones de similitudes aquí. Dichas figuras, junto a la comparación entre las similitudes estructural y de afinidad, se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.20 y 4.19, respectivamente).

Por la misma razón no se incluye el grafo con el análisis de modularidad.

3.4.2.2. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)

En esta sección se muestran los resultados de comparar todos los pares de drogas prescriptas para el tratamiento del trastorno bipolar, utilizando la similitud semántica obtenida a partir de aplicar la técnica de LSA en el corpus sobre reseñas de fármacos (sec.2.2.1.2).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.30: Comparación de similitudes para todo par de drogas prescripto para el tratamiento del trastorno bipolar. La similitud semántica surge del análisis mediante la técnica LSA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

El resultado de TASA se repite aquí: tanto la comparación entre las similitudes semántica y estructural (fig.3.30a), y entre las similitudes semántica y de afinidad (fig.3.30b) resultan no significativas.

La comparación entre similitudes estructural y de afinidad, y las comparaciones de similitud estructural, y de afinidad para las drogas prescriptas para trastorno bipolar analizadas mediante LSA se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.21 y 4.22, respectivamente).

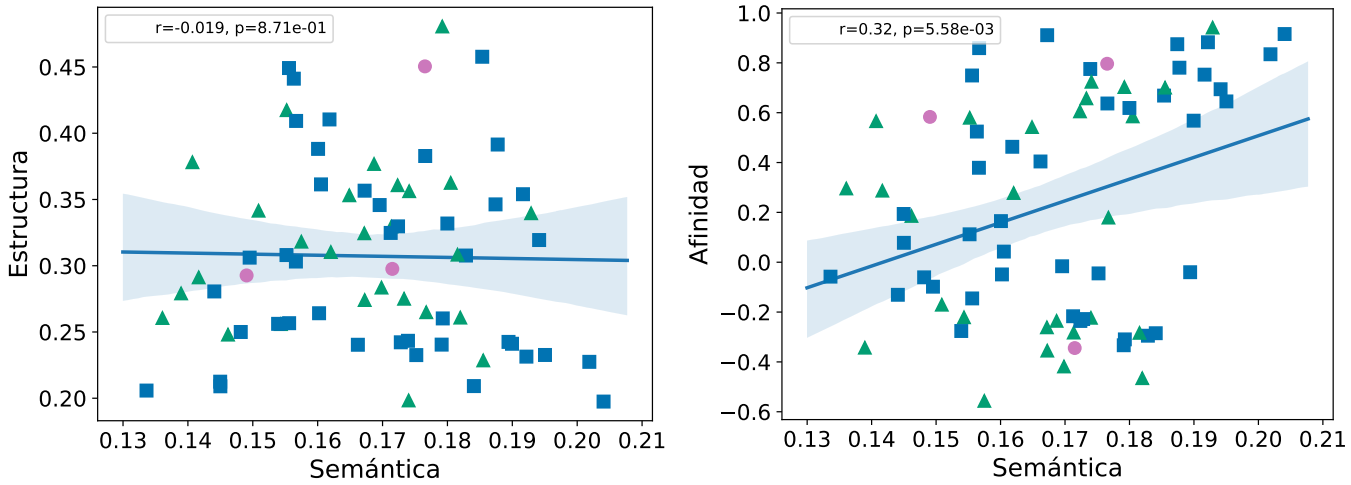
3.4.3. Depresión

Esta sección comprende los resultados de la comparación de los pares de drogas prescriptas para el tratamiento de la depresión.

La depresión es una de las condiciones más preponderantes en nuestro conjunto de datos, acumulando el 35 % de los reportes (tabla 4.4), y siendo tratada con un total de 27 drogas distintas (de las 33 analizadas). Por esta razón, es esperable que los resultados de analizar únicamente esta condición sean parecidos a los del caso inicial, en el que se analizaron todas las drogas juntas (sec.3.2), e incluso a los antidepresivos entre sí (sec.3.3.2).

3.4.3.1. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)

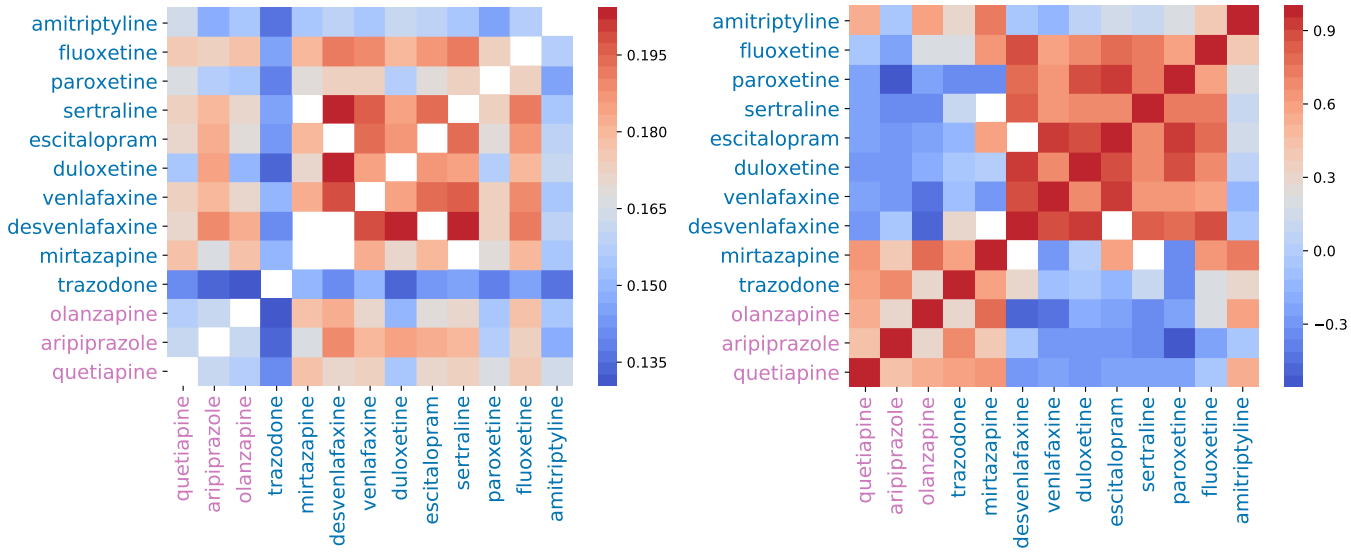
En esta sección se muestran los resultados de comparar todos los pares de drogas prescriptas para el tratamiento de la depresión, utilizando la similitud semántica obtenida a partir del corpus TASA (sec.2.2.1.1).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.31: Comparación de similitudes para todo par de drogas prescripto para el tratamiento de la depresión. La similitud semántica surge de la utilización del corpus TASA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

Tal como mencionamos previamente, para el análisis por LSA de las drogas prescriptas para la depresión se repiten resultados pasados, donde la comparación entre las similitudes semántica y estructural (fig. 3.31a) devuelve un resultado no significativo, mientras que en la comparación entre las similitudes semántica y de afinidad (fig. 3.31b), es significativo y muestra una correlación positiva entre ambas medidas.



(a) Similitud semántica *(b)* Similitud de afinidad

Fig. 3.32: Comparaciones de similitud para todo par de drogas prescriptas para el tratamiento de la depresión. La similitud semántica surge de la utilización del corpus TASA.

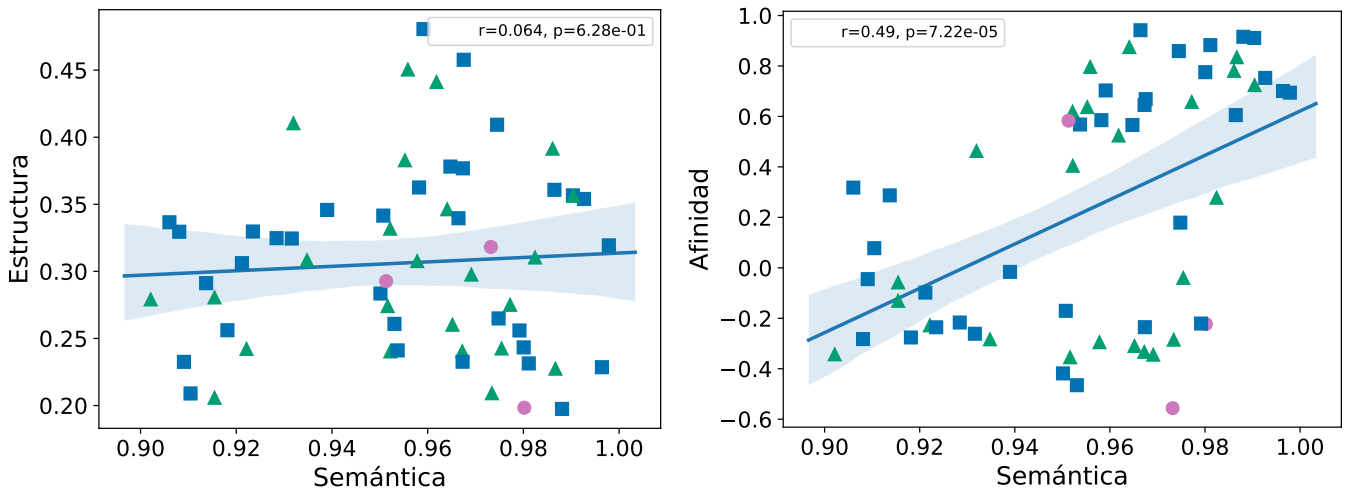
En las comparaciones de similitud semántica (fig.3.32a) se ve la distinción entre las drogas antidepresivas (el grupo de SSRI[15]/SNRI[16]), por lo menos) y las antipsicóticas, presentando tanto **amitriptilina** como **trazodona**, al igual que en ocasiones pasadas, un comportamiento anómalo. Como en la comparación de similitud semántica entre todos los pares de drogas (fig.3.9a), la semántica de las drogas antidepresivas presenta una consistencia interna mayor (tonos rojos) que en el grupo de las antipsicóticas (tonos azules).

Para la comparación de las similitudes de afinidad (fig.3.32b), la diferenciación es mayor (tal como sucedió en 3.9c), con las excepciones (una vez más) de **trazodona**, **mirtazapina**, y **amitriptilina**), más parecidas a los antipsicóticos.

La comparación entre similitudes estructural y de afinidad, y las comparaciones de similitud estructural para las drogas prescriptas para depresión analizadas mediante TASA se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.23 y 4.24, respectivamente).

3.4.3.2. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)

En esta sección se muestran los resultados de comparar todos los pares de drogas prescriptas para el tratamiento de la depresión, utilizando la similitud semántica obtenida a partir de aplicar la técnica de LSA en el corpus sobre reseñas de fármacos (sec.2.2.1.2).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.33: Comparación de similitudes para todo par de drogas prescriptas para el tratamiento de la depresión. La similitud semántica surge del análisis mediante la técnica LSA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

La comparación entre las similitudes semántica y estructural de los pares de drogas prescriptas para depresión (fig. 3.33a), no presenta ninguna correlación significativa.

Ciñéndose a la comparación de similitudes semántica y de afinidad (fig.3.33b), hay una marcada correlación positiva, mejor que en casos anteriores.

[15] Fluoxetina, paroxetina, sertralina y escitalopram
[16] Duloxetina, venlafaxina y desvenlafaxina

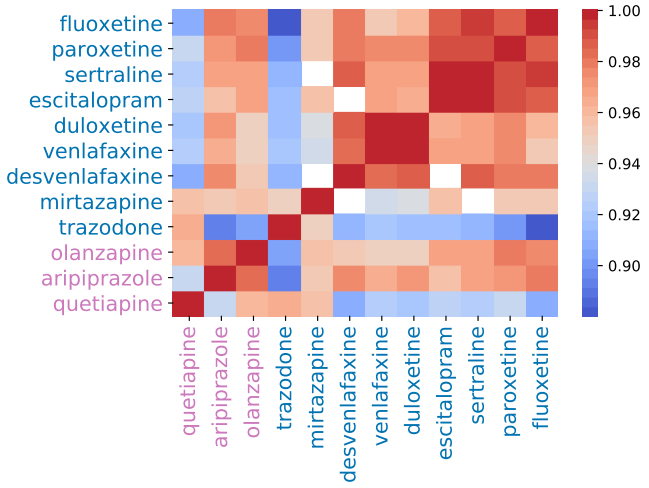


Fig. 3.34: Comparación de similitud semántica para todo par de drogas prescriptas para el tratamiento de la depresión. La similitud semántica surge del análisis mediante la técnica LSA.

En la comparación de similitud semántica (fig.3.34), hay en particular consistencia semántica interna entre los antidepresivos (más marcada en los SSRI[17] y SNRI[18]), y un poco menor entre los antipsicóticos (donde además, **olanzapina** y **aripiprazole** son también parecidos semánticamente a la mayoría de los antidepresivos, aunque son más parecidos entre sí). **Trazodona** no resulta agrupable, teniendo correlación únicamente con **quetiapina** y **mirtazapina** (que presenta correlación alta tanto con antipsicóticos como con los antidepresivos SSRI).

La comparación entre similitudes estructural y de afinidad, y las comparaciones de similitud estructural y de afinidad para las drogas prescriptas para depresión analizadas mediante LSA se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.25 y 4.26, respectivamente).

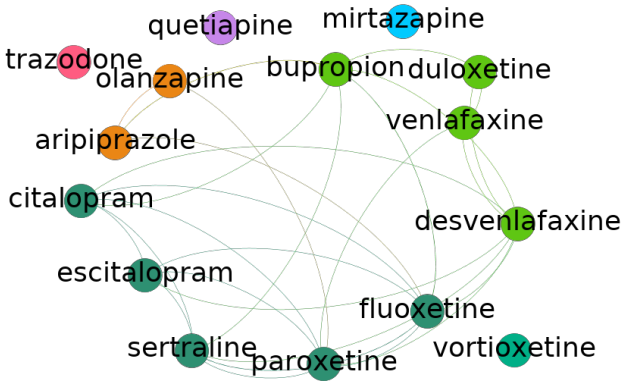


Fig. 3.35: Grafo subyacente a la matriz de similitud semántica para todo par de drogas prescriptas para depresión Se conservó el 27 % de los ejes con mayor peso.

[17] Fluoxetina, paroxetina, sertralina, y escitalopram
[18] Duloxetina, venlafaxina y desvenlafaxina

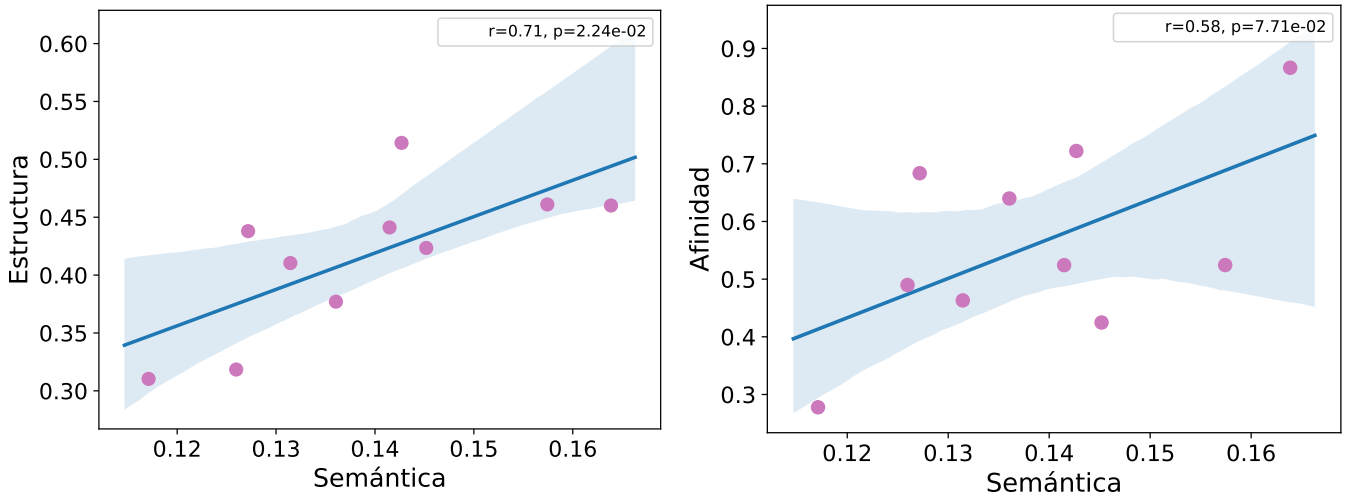
En este caso, el análisis de modularidad sobre el grafo correspondiente a las similitudes semánticas entre todo par de drogas prescriptas para la depresión devuelve siete comunidades. Cuatro de las drogas aparecen como únicas integrantes de su comunidad: los antidepresivos **trazodona** (SARI), **mirtazapina** (TeCA) y **vortioxetina** (SMS), correctamente, y **quetiapina** de manera incorrecta, ya que es un antipsicótico al igual que **aripiprazol** y **olanzapina**. Por otro lado, los SSRI (en verde oscuro) están bien agrupados, y los SNRI (en verde claro) aparecen una vez más acompañados por el **bupropion** (NDRI).

3.4.4. Esquizofrenia

Esta sección comprende los resultados de la comparación de los pares de drogas prescriptas para el tratamiento de la esquizofrenia.

3.4.4.1. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)

En esta sección se muestran los resultados de comparar todos los pares de drogas prescriptas para el tratamiento de la esquizofrenia, utilizando la similitud semántica obtenida a partir del corpus TASA (sec.2.2.1.1).



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.36: Comparación de similitudes para todo par de drogas prescriptas para el tratamiento de la esquizofrenia. La similitud semántica surge de la utilización del corpus TASA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

Para el grupo de las drogas recetadas para el tratamiento de la esquizofrenia, la comparación de las similitudes semántica y estructural (fig. 3.36a) arroja un resultado significativo, siendo la correlación positiva. La comparación entre semántica y afinidad (fig.3.36b), no obstante, se muestra no significativa (aunque con una tendencia hacia una correlación positiva).

Dado que las drogas prescriptas para el tratamiento de la esquizofrenia son todos antipsicóticos, resulta superfluo mostrar las comparaciones de similitudes aquí. Dichas figuras, junto a la comparación entre las similitudes estructural y de afinidad, se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.28 y 4.27, respectivamente).

3.4.4.2. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)

En este caso no fue posible realizar la comparación mediante la técnica LSA aplicada al corpus de reseñas sobre fármacos, debido a que el tratamiento de los datos (removiendo outliers, por ejemplo) dejaba una cantidad de drogas insuficientes para la comparación (3 o menos).

3.4.5. Refinación de la comparación

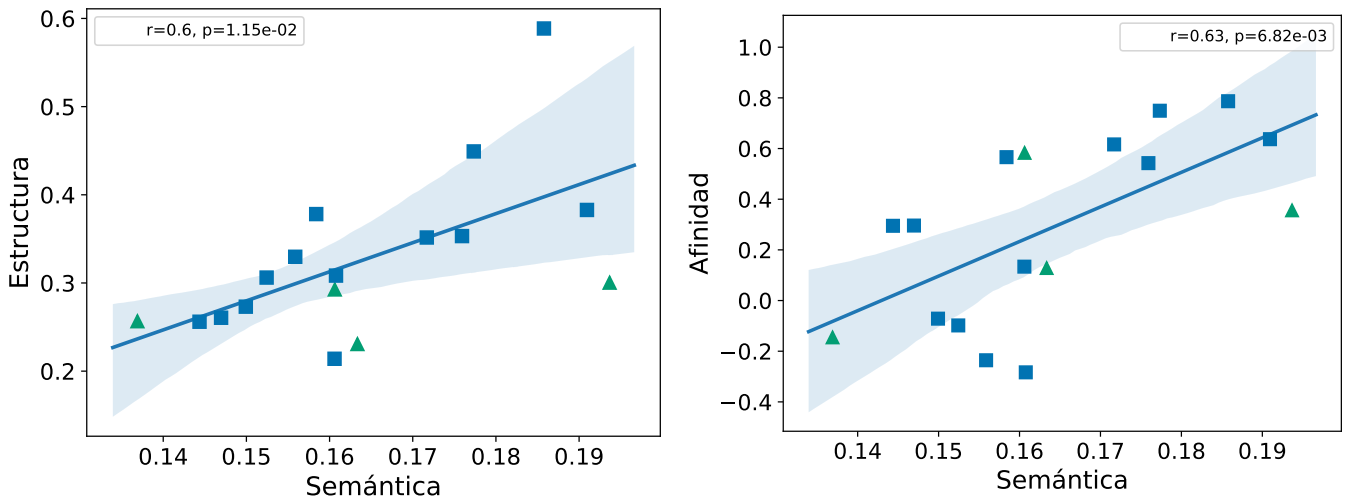
Se optó refinar una vez más el conjunto de datos utilizado para el análisis. En esta ocasión, se partió de la idea de que las condiciones por las que se está realizando el análisis presentan una dimensión cognitiva fuerte, pudiendo ser esto lo que modifica (y agrega variabilidad) a la semántica reportada por los pacientes. Se decidió partir de un conjunto de condiciones que no puedan, en teoría, afectar cognitivamente a los pacientes que escribieron las reseñas.

3.5. Subconjunto de condiciones

En esta sección se muestran los resultados de la comparación entre todos los pares de drogas prescriptas para condiciones que no afectan de manera directa la cognición de los pacientes (condiciones no cognitivas).

Algunas de las condiciones *no cognitivas* que forman parte de esta última comparación son pérdida de peso (tratada con **bupropion**), migrañas (**amitriptilina**, **nortriptilina**, **venlafaxina**), trastornos del ritmo circadiano (**amitriptilina**, **doxepina**, **mirtazapina**, **quetiapina**, **trazodona**), abuso de sustancias (**bupropion**), entre otras.

3.5.1. Comparación de similitudes estructural, de afinidad y semántica (con corpus TASA)



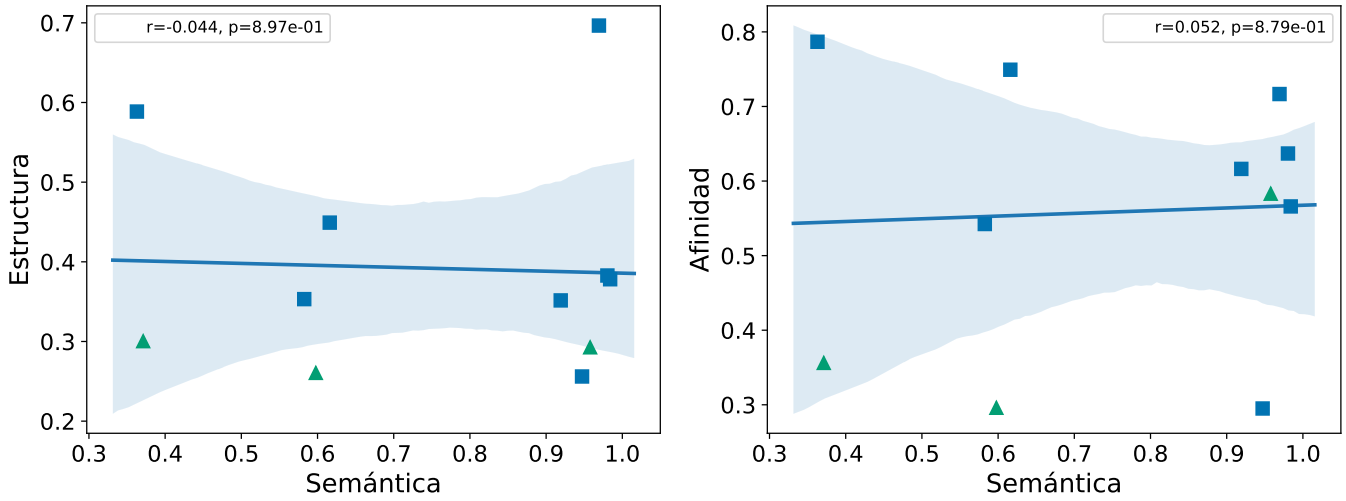
(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.37: Comparación de similitudes para todo par de drogas prescripto para el tratamiento de un subconjunto de condiciones. La similitud semántica surge de la utilización del corpus TASA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

El análisis mediante el corpus TASA del conjunto de drogas prescripto para este subconjunto de condiciones brinda los resultados que se pueden ver en la figura anterior (fig.3.37). Tanto la comparación entre las similitudes semántica y estructural (fig.3.37a) y la de las similitudes semántica y de afinidad (fig.3.37b) muestran una correlación positiva significativa. Cabe destacar que el resultado positivo en ambas comparaciones mediante el corpus TASA sólo había sido obtenido hasta el momento en la comparación entre drogas antidepresivas con más de 1.000 reseñas (fig.3.20).

La comparación entre las similitudes estructural y de afinidad, y las comparaciones de similitud, se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.30 y 4.31, respectivamente).

3.5.1.1. Comparación de similitudes estructural, de afinidad y semántica (con LSA entrenado sobre corpus propio)



(a) Comparación entre similitudes semántica y estructural *(b)* Comparación entre similitudes semántica y de afinidad

Fig. 3.38: Comparación de similitudes para todo par de drogas prescripto para el tratamiento de condiciones *no cognitivas*. La similitud semántica surge del análisis mediante la técnica LSA. La línea, producto de una regresión lineal, representa la recta que mejor se ajusta a los datos. Se incluyen el coeficiente de correlación lineal de Pearson y su p-valor asociado.

Para el caso de LSA aplicado al corpus de reseñas (figs.3.38a y 3.38b), como sucedió cada vez que se compararon pocos pares de droga entre sí, los resultados no son significativos.

La comparación entre las similitudes estructural y de afinidad, y las comparaciones de similitud, se encuentran en el apéndice (c.4.2.1, sec.4.5, figs. 4.33 y 4.34, respectivamente).

4. DISCUSIÓN Y CONCLUSIONES

El trabajo consistió en la recopilación y análisis de datos públicamente accesibles: estructura molecular de drogas antidepresivas y antipsicóticas, afinidad de las mismas a distintos receptores, y reportes subjetivos de pacientes tratados con las mismas. Estos datos sirvieron para obtener la similitud entre cada una de esas dimensiones para todos los pares de drogas. Se analizaron en total cerca de 33 drogas, repartidas (no uniformemente) en casi 33.000 reseñas en total.

Se observó en todos los casos donde se obtuvieron resultados significativos que para las diversas comparaciones de subconjuntos de drogas existe una correlación positiva entre la semántica reportada al utilizar alguna de las drogas analizadas, y la estructura y/o afinidad de las mismas.

Más aún, restringiéndose únicamente a las comparaciones de similitudes semánticas de las drogas mediante los mapas de calor y el análisis de comunidades sobre los grafos, se observó que es posible diferenciar a los grupos de antipsicóticos con los antidepresivos, y dentro de éstos, a los que pertenecían a distintas clases (SSRI, SNRI, TeCA, TCA, etc.), con algunas particularidades según el caso (por ejemplo, el NDRI **bupropion** siempre aparecía ligado a los **SNRI**).

Más allá de las dificultades encontradas y posibles cuestionamientos al trabajo (sec. 4.1), la tesis pretende servir como punto de partida para incrementar la incursión de herramientas computacionales en el análisis de grandes volúmenes de datos que comprendan la relación entre drogas antipsicóticas y antidepresivas y sus efectos poblacionales. Dado que la interacción de estas drogas con el sistema nervioso es ya de por sí muy compleja, trabajos como el presente brindan una manera asequible de aproximarse al problema.

Podemos conjeturar que, en un futuro, se podrán diseñar drogas de manera *ad-hoc* esperando obtener efectos específicos basándose en los resultados de análisis como el presentado en este trabajo. En la medida que la cantidad de pacientes tratados y su conectividad aumente, la extracción de información de datos online presentará una manera útil para evaluar los efectos de fármacos una vez lanzados al mercado. Mediante el desarrollo de técnicas de ingeniería reversa, conjeturamos que será posible diseñar fármacos de acuerdo a los efectos subjetivos deseados.

4.1. Dificultades y posibles mejoras

A pesar de los resultados positivos, no son pocas las cuestiones que podrían mejorarse y que, de hacerlo, podrían dar lugar a estudios más sólidos.

Por un lado, hay que tener en cuenta que los reportes de las drogas fueron obtenidos a través de información disponible públicamente, lo cual podría generar dudas sobre su consistencia, y veracidad. La población de usuarios que se toma el trabajo de comentar en un sitio público qué efectos tuvo o qué sintió al ser tratado con una droga se puede considerar un subconjunto muy específico y con un sesgo bastante particular; sin embargo, es también lo que posibilitó el trabajo, ya que realizar un estudio controlado con cuestionarios estandarizados en una población de 33.000 sujetos que padezcan alguna condición parece inviable como objeto de una Tesis de Licenciatura de Cs. de la Computación.

Esta falta de formalidad conduce a trabajar con un conjunto de datos relativamente sucio, puesto que a la informalidad de los reportes se suma la población que los escribe, en algunos casos gente con patologías que podrían causar alguno de los siguientes problemas:

- Escribir mal (aunque la patología no es condición necesaria para que eso ocurra). Esto requiere, como se mostró en el trabajo (sec.2.1.3.2 Preprocesamiento del texto), de un análisis minucioso sobre cuáles son las palabras mal escritas más comunes, y tratar de corregirlas manualmente, haciendo una interpretación subjetiva adecuada de las mismas.

- Neologismos. Está estudiado [19] que la esquizofrenia, por ejemplo, altera el lenguaje de los individuos que la sufren, llevándolos incluso a producir neologismos. En nuestro estudio los neologismos son una fuente de ruido, puesto que al aparecer en una frase se estaría perdiendo significado en la misma ante la imposibilidad de analizar dicha palabra. No obstante, si varios pacientes producen los mismos neologismos, la técnica de **LSA** podría darles un significado por su co-ocurrencia junto a otras palabras; sin embargo, es poco probable que dos personas inventen exactamente la misma palabra.

- Dosis de los medicamentos. No todos los reportes especificaban de manera estructurada la dosis recetada del medicamento; a veces, la misma era mencionada dentro del reporte del paciente, quien en muchos casos admitía que modificaba la dosis, aumentándola si no sentía efecto alguno o disminuyéndola si los efectos eran demasiado fuertes. Muy pocas veces se especificaba si tal modificación fue por indicación del médico. Dado que la dosis modifica los efectos de una droga (y, bajo nuestra hipótesis, también la semántica), esto es otra fuente considerable de ruido en el análisis.

- Interacción entre medicamentos. Al igual que en el inciso anterior, a veces, el paciente reporta estar tomando otras drogas (del mismo tenor, es decir, otros antipsicóticos u otros antidepresivos), cuya interacción se desconoce. Probablemente, otros pacientes estén bajo la misma situación pero no incluyan en el reporte qué están tomando otros fármacos.

Otra consideración tiene que ver con los tipos de drogas que están utilizando. El análisis llevado a cabo fue utilizado en un trabajo previo en el contexto de drogas de uso recreativo (en particular, psicodélicos serotonérgicos [35]). Hay una distinción entre el reporte que uno hace al tomar una droga de carácter recreativo, donde tiene preponderancia *el viaje*, que el realizado al consumir un fármaco con el objetivo de tratar una condición.

El análisis de experiencias psicodélicas en el trabajo citado permitió extraer componentes de significado muy claro (*percepción* -visual, auditiva, olfativa, etc.-, *carga corporal* -estimulación, estado de ánimo-, *preparación* -para el consumo de la droga-, etc.); al analizar un fármaco, el significado extraído se relaciona también con la condición que se esté tratando, así como con los posibles efectos secundarios de la medicación.

Los datos de afinidad de las distintas drogas con los receptores surge de ensayos publicados en trabajos previos [28]. La replicación de estos resultados podría brindar valores más consistentes y completos de afinidad, y proveer una base más fuerte sobre la cual basar este trabajo o similares. Es importante aclarar dos limitaciones de los datos de afinidad. En primer lugar, y tal como fue mencionado en la introducción, la afinidad no es necesariamente proporcional con la eficacia de la droga en el receptor. Por lo tanto, la

afinidad representa una aproximación al efecto biológico del fármaco. En segundo lugar, el presente análisis no consideró la actividad funcional de la droga en el receptor (agonista, antagonista, etc).

4.2. Trabajo futuro

Cabe destacar que si bien el análisis fue exhaustivo no se considera que de ningún modo esté completo, ya que hay mucho más para explorar en trabajos futuros.

En lo que respecta a la afinidad, por ejemplo, hay muchos detalles que no se tuvieron en cuenta y podrían agregar una mayor riqueza al análisis.

Uno de ellos es el análisis de subconjuntos específicos de receptores, donde se podrían estudiar las diferencias (o similitudes) del impacto en la semántica entre distintas drogas con mayor precisión. En particular, podrían identificarse los subconjuntos de receptores más relevantes para la semántica de sus reseñas.

Por otro lado, y tal como se mencionó en la sección anterior, no se estudió (por falta de datos y tiempo) el efecto específico de los ligandos sobre los receptores: los ligandos pueden activar o desactivar al receptor, y esta activación a su vez puede tanto aumentar como disminuir la actividad de alguna función celular.

Otra clasificación de los ligandos es en agonistas (los que se unen a un receptor y provocan una acción específica en la celula a la cual pertenece; a su vez, se clasifican en parciales y completos, provocando aquellos una respuesta menor que los últimos), y antagonistas (ligandos que también se unen al receptor pero además de no activarlos, impiden que otros agonistas puedan activarlos); esta distinción tampoco fue objeto del estudio actual.

Existen también moléculas catalogadas como **segundos mensajeros**. Las mismas son moléculas intracelulares que al recibir una señal externa (en nuestro caso, los psicofármacos -los primeros mensajeros- uniéndose a los receptores de la célula) se encargan de propagar la señal por el interior de la célula a la que pertenecen, lo que las vuelve determinantes para la consideración de los efectos generados por los psicofármacos. Un análisis como el hecho en este trabajo que contemple el efecto de los segundos mensajeros, sería sin dudarlo más completo.

En lo que respecta a la estructura de las moléculas, la similitud entre ellas fue calculada con AtomPair, como se describió previamente (sec.2.1.2 Estructura molecular). Se podrían repetir los experimentos considerando otras medidas para la similitud entre dos moléculas.

Por el lado de la semántica, y las herramientas y técnicas utilizadas para calcularla, hay alternativas a LSA que podrían haber sido usadas, pero no fueron incluidas por razones de tiempo. Algunas de ellas fueron mencionadas en la introducción (sec.1.3 Procesamiento de Lenguaje Natural), como Word2Vec, GloVe, y fastText, pero otras no (por ejemplo, NMF [15] en lugar de SVD para la reducción de dimensionalidad del espacio semántico).

Es posible cuestionar también la utilización de la regresión lineal en los gráficos de dispersión. Al estar comparando valores que provienen de pares de drogas, donde éstas pueden pertenecer a más de un par, los distintos puntos no serían independientes entre sí. La visualización de estos resultados sin esa transgresión es algo a trabajar en el futuro.

4.2.1. Desde un fingerprint hacia la semántica

En esta sección se introducirá un posible futuro trabajo, utilizando parte de los datos ya recopilados.

Como se mencionó en la sección 2.1.2 Estructura molecular, existen distintos algoritmos de *fingerprinting* para poder codificar la estructura de una molécula en una cadena de bits. La herramienta RDKit permite obtener el *fingerprint* de Morgan, que es un caso particular de huella de conectividad extendida [27].

Por otro lado, es posible mediante la técnica de LSA, ya vista, lograr una caracterización de las drogas en el espacio semántico. Esto es, para cada droga, realizar una reducción de dimensionalidad que permita obtener una representación de la droga en n componentes ($n = 20$ fue lo utilizado en el trabajo), en lugar de la representación conformada por todas las palabras utilizadas en las reseñas de la droga en cuestión.

Entonces, se tendrían para cada droga su *fingerprint* y su representación en 20 componentes. Esta representación, al ser una caracterización semántica de la droga, se puede utilizar para clasificarla: binarizando[1] los valores de la droga a lo largo de las 20 componentes, la droga queda determinada por su presencia o ausencia en cada una. Con estos datos se podría entrenar un modelo de clasificador binario para cada componente con el fin de que, al darle una cadena de bits (una droga) como entrada, prediga su clase (si está *activada* o no en esa componente).

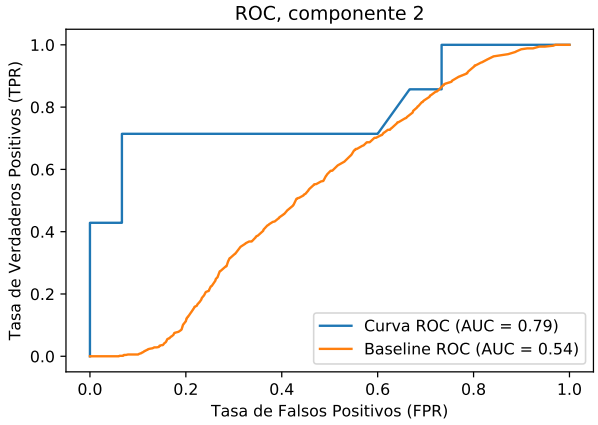
Una forma de evaluar un modelo de clasificación binario es mediante la curva AUC-ROC (Area Under the Curve - Receiver Operating Characteristics). La curva ROC representa la tasa de verdaderos positivos frente a la tasa de falsos positivos, según varía el umbral de clasificación (el valor a partir del cuál se decide que un caso es positivo, o negativo). AUC es el área debajo de esa curva, y puede ser interpretada como la probabilidad de clasificar un ejemplo aleatorio positivo de modo más alto que un ejemplo negativo aleatorio; si el área es 1.0, las predicciones del modelo son 100 % correctas, mientras que si es 0.0, las predicciones son 100 % incorrectas.

A partir del modelo entrenado se pueden obtener, además, las características que el mismo considera **importantes** para realizar la clasificación; en nuestro caso, esto equivaldría a bits importantes del *fingerprint*. Es decir: la activación de una droga en una componente dada, estaría dada por algunos bits del *fingerprint* de la molécula. Teniendo en cuenta que los bits del *fingerprint* representan en última instancia fragmentos de la molécula, y que las componentes principales representan a la semántica de la droga (recordemos la sección 3.1 Componentes SVD de LSA aplicado al corpus y términos asociados a las mismas), se podría llegar a establecer una relación entre fragmentos de una molécula y palabras específicas.

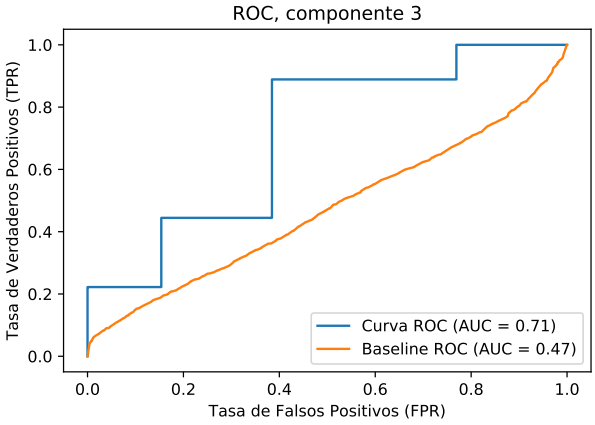
Para demostrar la viabilidad de este experimento, presentamos los siguientes resultados preliminares.

Las figuras 4.1 y 4.2 muestran los resultados de entrenar un modelo de clasificación binario con el *fingerprint* de las moléculas y algunas de las distintas componentes asociadas al análisis semántico por LSA. No se incluye la primer componente pues todas las moléculas

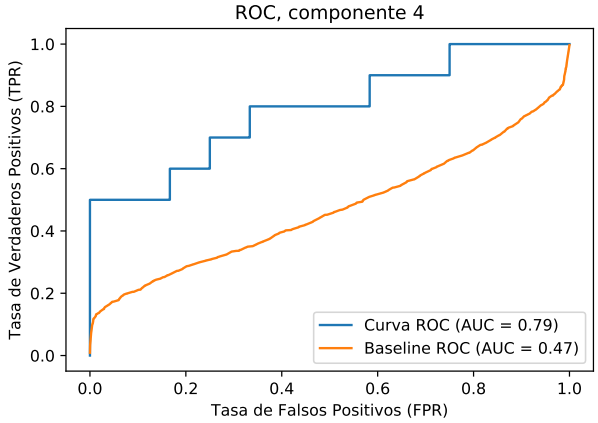
[1] Bajo algún criterio dado, poner un 0 o un 1



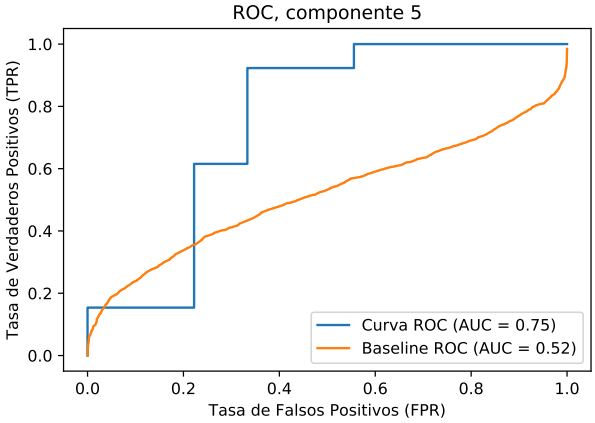
(a) Curva ROC para la componente 2.



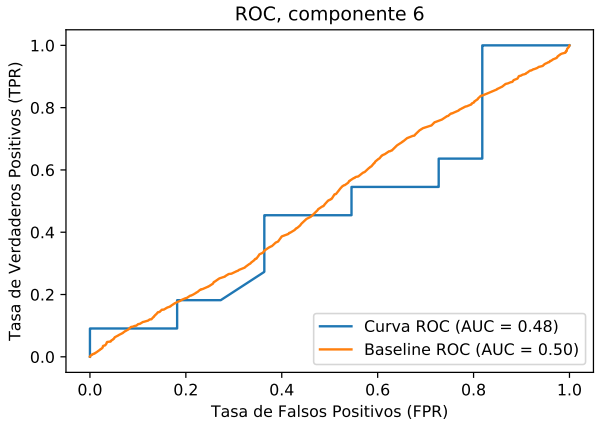
(b) Curva ROC para la componente 3.



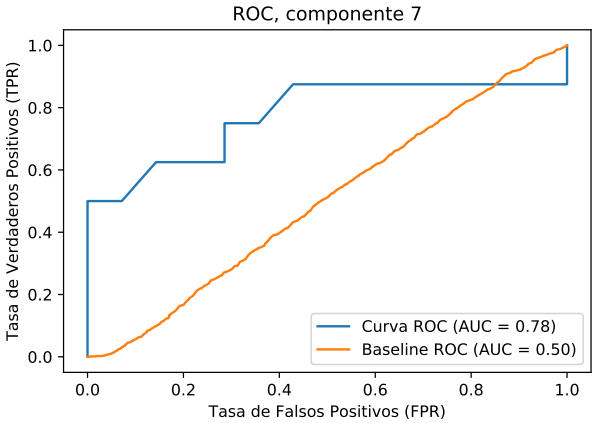
(c) Curva ROC para la componente 4.



(d) Curva ROC para la componente 5.

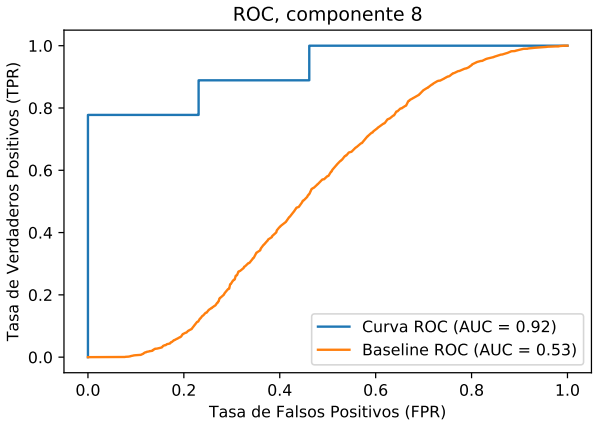


(e) Curva ROC para la componente 6.

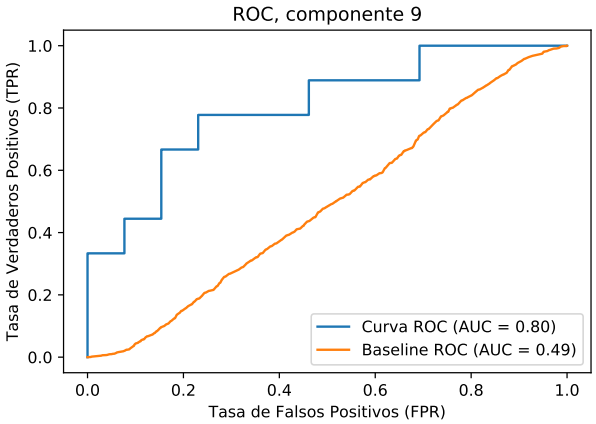


(f) Curva ROC para la componente 7.

Fig. 4.1: Curvas ROC (parte 1) para la evaluación de un clasificador binario de tipo RandomForest con 200 estimadores y 7 folds para cada componente. El clasificador utilizado como baseline fue entrenado 200 veces para cada componente.



(a) Curva ROC para la componente 9.



(b) Curva ROC para la componente 8.

Fig. 4.2: Curvas ROC (parte 2) para la evaluación de un clasificador binario de tipo RandomForest con 200 estimadores y 7 folds para cada componente. El clasificador utilizado como baseline fue entrenado 200 veces para cada componente.

están activadas en ella, imposibilitando el entrenamiento del modelo al no haber dos clases.

Los gráficos representan la evaluación del modelo mediante la curva AUC - ROC ya explicada. La línea naranja representa la curva ROC correspondiente al entrenamiento aleatorio del clasificador binario, que se utiliza como comparación base[2]; es decir, se mezclan las clases de los ejemplos de entrenamiento, lo cual permite saber qué resultado se obtendría a través del azar.

Los resultados muestran que las componentes sirven como método de clasificación de las moléculas, puesto que el valor del AUC es generalmente elevado y, más aún, la curva ROC se encuentra consistentemente por encima de la curva correspondiente a la obtención de los resultados al azar en todos los casos menos en uno (la componente 6, fig. 4.1e).

Esta línea de trabajo será continuada en el futuro inmediato.

[2] Baseline

Bibliografía

[1] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. Autor, Washington, DC, 5th ed. edition, 2013.

[2] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.

[3] Gillinder Bedi, Guillermo A Cecchi, Diego F Slezak, Facundo Carrillo, Mariano Sigman, and Harriet De Wit. A window into the intoxicated mind? speech as an index of psychoactive drug effects. *Neuropsychopharmacology*, 39(10):2340, 2014.

[4] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[5] Jeremy Besnard, Gian Filippo Ruda, Vincent Setola, Keren Abecassis, Ramona M. Rodriguiz, Xi-Ping Huang, Suzanne Norval, Maria F. Sassano, Antony I. Shin, Lauren A. Webster, Frederick R. C. Simeons, Laste Stojanovski, Annik Prat, Nabil G. Seidah, Daniel B. Constam, G. Richard Bickerton, Kevin D. Read, William C. Wetsel, Ian H. Gilbert, Bryan L. Roth, and Andrew L. Hopkins. Automated design of ligands to polypharmacological profiles. *Nature*, 492(7428):215–+, 2012.

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008.

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[8] Vannevar Bush and Jingtao Wang. As we may think. *Atlantic Monthly*, 176:101–108, 1945.

[9] Facundo Carrillo, Natalia Mota, Mauro Copelli, Sidarta Ribeiro, Mariano Sigman, Guillermo Cecchi, and Diego Fernandez Slezak. Automated speech analysis for psychosis evaluation. In *Machine Learning and Interpretation in Neuroimaging*, pages 31–39. Springer, 2014.

[10] Facundo Carrillo, Natalia Mota, Mauro Copelli, Sidarta Ribeiro, Mariano Sigman, Guillermo Cecchi, and Diego Fernandez Slezak. Emotional intensity analysis in bipolar subjects. *arXiv preprint arXiv:1606.02231*, 2016.

[11] Ella J Daly and Madhukar H Trivedi. A review of quetiapine in combination with antidepressant therapy in patients with depression. *Neuropsychiatric Disease and Treatment*, 3(6):855, 2007.

[12] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3267–3276. ACM, 2013.

[13] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[14] J. Flórez. *Farmacología Humana*. Masson, 2003.

[15] Fuzhen Sun and Kim Zhang. Nmf-based method of text classification. In *2010 8th World Congress on Intelligent Control and Automation*, pages 4312–4316, July 2010.

[16] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 09 2011.

[17] Wayne K Goodman, Lawrence H Price, Steven A Rasmussen, Pedro L Delgado, George R Heninger, and Dennis S Charney. Efficacy of fluvoxamine in obsessive-compulsive disorder: a double-blind comparison with placebo. *Archives of General Psychiatry*, 46(1):36–44, 1989.

[18] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

[19] Gina R Kuperberg. Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589, 2010.

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

[21] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[22] Natalia Bezerra Mota, Facundo Carrillo, Diego Fernández Slezak, Mauro Copelli, and Sidarta Ribeiro. Characterization of the relationship between semantic and structural language features in psychiatric diagnosis. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 836–838. IEEE, 2016.

[23] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[24] Thomas S. Ray. Psychedelics and the human receptorome. *PLOS ONE*, 5(2):1–17, 02 2010.

[25] RDKit: Open-source cheminformatics. http://www.rdkit.org. [Online; accedido 2019/06/20].

[26] Sereina Riniker and Gregory A Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics*, 5(1):26, 2013.

[27] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. PMID: 20426451.

[28] Bryan L. Roth, H.Y. Meltzer, and Naseem Khan. Binding of typical and atypical antipsychotic drugs to multiple neurotransmitter receptors. volume 42 of *Advances in Pharmacology*, pages 482 – 485. Academic Press, 1997.

[29] Camila Sanz, Federico Zamberlan, Earth Erowid, Fire Erowid, and Enzo Tagliazucchi. The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports. *Frontiers in Neuroscience*, 12:7, 2018.

[30] Robert P. Sheridan, Michael D. Miller, Dennis J. Underwood, and Simon K. Kearsley. Chemical similarity using geometric atom pair descriptors. *Journal of Chemical Information and Computer Sciences*, 36(1):128–136, 1996.

[31] Stephen F Signer. Is there an obsessive psychosis? 1986.

[32] G Mustafa Soomro. Obsessive compulsive disorder. *BMJ clinical evidence*, 2012, 2012.

[33] G Mustafa Soomro, Douglas G Altman, Sundararajan Rajagopal, and Mark Oakley Browne. Selective serotonin re-uptake inhibitors (ssris) versus placebo for obsessive compulsive disorder (ocd). *Cochrane database of systematic reviews*, (1), 2008.

[34] Ann Taylor, Mitchell Marcus, and Beatrice Santorini. The penn treebank: An overview, 2003.

[35] Federico Zamberlan, Camila Sanz, Rocío Martínez Vivot, Carla Pallavicini, Fire Erowid, Earth Erowid, and Enzo Tagliazucchi. The varieties of the psychedelic experience: A preliminary study of the association between the reported subjective effects and the binding affinity profiles of substituted phenethylamines and tryptamines. *Frontiers in Integrative Neuroscience*, 12:54, 2018.

4. APÉNDICE

4.3. Fármacos y sus receptores

En esta tabla pueden verse los fármacos para los cuales se consiguieron datos sobre su afinidad con los distintos receptores detallados.

Fármaco	**Receptores**
Amitriptilina	5HT1A, 5HT1B, 5HT2A, 5HT2B, 5HT2C, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, BETA2, D1, D2, D3, D5, DAT, H1, H2, H3, H4, M1, M2, M3, M4, M5, NET, SERT, SIGMA
Amoxapina	5HT1A, 5HT2A, 5HT2B, 5HT2C, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, D1, D2, D3, D4, DAT, H1, M1, M2, M3, M4, M5, NET, SERT
Aripiprazol	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, SERT
Clorpromazina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, NET, SERT, SIGMA
Clomipramina	5HT1A, 5HT1B, 5HT1D, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, D1, D2, D3, DAT, H1, H2, M1, M2, M3, M4, M5, NET, SERT, SIGMA
Clozapina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, NET, SERT
Desipramina	5HT1A, 5HT1B, 5HT2A, 5HT2C, 5HT3, 5HT7, ALPHA1A, ALPHA2A, BETA2, D1, D2, DAT, H1, M1, M2, M3, M4, M5, NET, SERT
Desvenlafaxina	D1, D2, D3, DAT, H1, H2, H3, M1, M2, M3, M4, M5, SERT
Doxepina	5HT1A, 5HT2A, 5HT2B, 5HT2C, 5HT6, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, D1, D2, D3, DAT, H1, H2, H4, M1, M2, M3, M4, M5, NET, SERT, SIGMA
Duloxetina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, BETA1, BETA2, D1, D2, DAT, DOR, H1, H2, KOR, M1, M2, MOR, NET, SERT
Escitalopram	5HT1A, 5HT1B, 5HT2A, 5HT2C, ALPHA1A, ALPHA2A, D1, D3, DAT, H1, M1, SERT

Fármaco	**Receptores**
Fluoxetina	5HT1A, 5HT1B, 5HT1D, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT6, 5HT7, ALPHA1A, ALPHA2B, BETA2, D1, D2, DAT, H1, H3, M1, M2, M3, M4, M5, NET, SERT, SIGMA
Flufenazina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, NET, SERT, SIGMA
Fluvoxamina	5HT1A, 5HT2A, 5HT2C, BETA2, D1, D2, DAT, H1, M1, SERT, SIGMA
Haloperidol	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, NET, SERT, SIGMA
Imipramina	5HT1A, 5HT2A, 5HT2C, 5HT3, 5HT6, 5HT7, ALPHA1A, BETA2, D1, D2, D3, DAT, H1, M1, M2, M3, M4, M5, SERT
Loxapina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, H1, H2, H3, H4, M1, M2, M3, M4, M5, SERT
Mirtazapine	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2C, 5HT3, 5HT5, 5HT7, ALPHA1A, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, DAT, H1
Nefazodona	5HT1A, 5HT2A, 5HT2C, ALPHA1A, BETA2, D1, D2, DAT, H1, SERT
Nortriptilina	5HT1A, 5HT1B, 5HT2A, 5HT2B, 5HT2C, 5HT6, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, D1, D3, DAT, H1, H2, M1, M2, M3, M4, M5, NET, SERT
Olanzapina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, SERT
Paroxetina	5HT1A, 5HT1B, 5HT1D, 5HT2A, 5HT2B, 5HT2C, 5HT3, ALPHA1A, BETA2, D1, D2, DAT, H1, M1, M2, M3, M4, M5, NET, SERT, SIGMA
Quetiapina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, SERT, SIGMA

Fármaco	**Receptores**
Risperidona	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, SERT
Sertralina	5HT1A, 5HT2A, 5HT2B, 5HT2C, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, BETA2, D1, D2, DAT, H1, M1, M2, M3, M4, M5, NET, SERT, SIGMA
Tioridazina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, H1, H2, H3, H4, KOR, M1, M2, M3, M4,M5, NET, SERT, SIGMA
Tiotixeno	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, H1, H2, H3, H4, M1, M2, M3, M4, M5, SERT
Trazodona	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, SERT, SIGMA
Venlafaxina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, BETA1, BETA2, D1, D2, D3, DAT, DOR, H1, H2, KOR, M1, M2, M3, M4, M5, MOR, NET, SERT
Ziprasidona	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2B, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, SERT
Zotepina	5HT1A, 5HT1B, 5HT1D, 5HT1E, 5HT2A, 5HT2C, 5HT3, 5HT5, 5HT6, 5HT7, ALPHA1A, ALPHA1B, ALPHA2A, ALPHA2B, ALPHA2C, BETA1, BETA2, D1, D2, D3, D4, D5, DAT, DOR, H1, H2, H3, H4, KOR, M1, M2, M3, M4, M5, MOR, SERT

Tab. 4.1: Drogas y receptores para los cuales se tienen datos de afinidad

4.4. Distribución de reseñas

En las siguientes tablas, la abreviatura "T." equivale a trastorno/trastornos.

Tipo de droga	**# de reseñas**
Antidepresivo	28655
Antipsicótico	4139

Tab. 4.2: Cantidad de reseñas por tipo de droga

Fármaco	**# de reseñas**
Amitriptilina	2165
Amoxapina	1
Aripiprazol	1650
Bupropion	3682
Clorpromazina	76
Citalopram	2837
Clomipramina	190
Clozapina	71
Desipramina	74
Desvenlafaxina	954
Doxepina	485
Duloxetina	1595
Escitalopram	2251
Fluoxetina	1577
Flufenazina	15
Fluvoxamina	241

Fármaco	**# de reseñas**
Imipramine	149
Loxapina	24
Mirtazapina	1244
Nefazodona	69
Nortriptilina	782
Olanzapina	594
Paroxetina	1262
Quetiapina	1392
Risepridona	571
Sertralina	2693
Tioridazina	4
Tiotixeno	16
Trazodona	2667
Venlafaxina	2025
Vortioxetina	887
Ziprasidona	410

Tab. 4.3: Cantidad de reseñas por fármaco

Condición	**# de reseñas**
T. de ansiedad	6717
T. bipolar	2258
Cluster A	36
Cluster B	168
T. de conducta	15
Depresión	11701
T. disociativos	2
T. alimenticios	52
T. de eliminación	31
Fibromyalgia	970
Cambios de vida	123
Menopausia	219

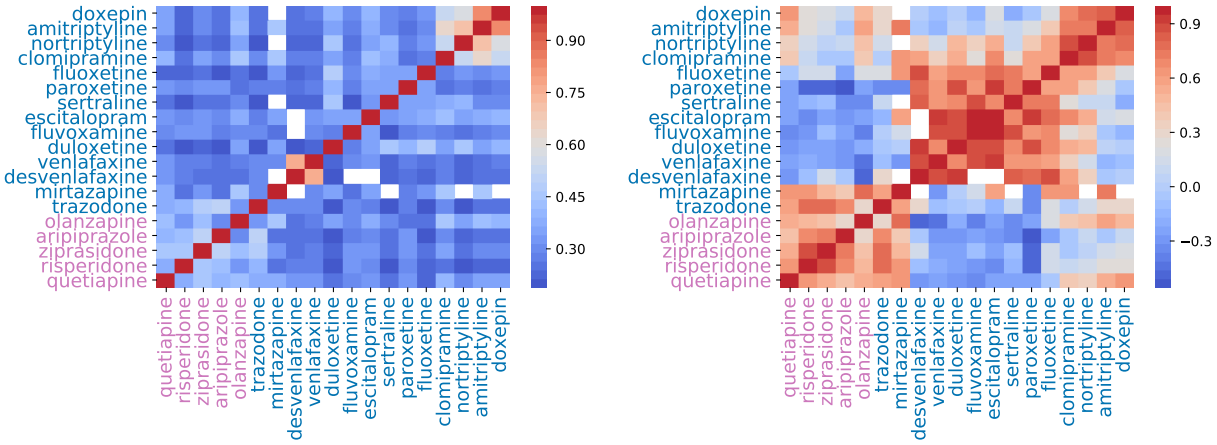
Condición	**# de reseñas**
Migrañas	726
T. neurocognitivos	17
T. del desarrollo neurológico	277
Neuropatías	511
T. obsesivo compulsivo	699
Espectro de la esquizofrenia	886
Disfunciones sexuales	81
T. del ritmo circadiano	2979
T. somáticos	22
T. de abuso de sustancias	314
Trauma	407
Pérdida de peso	1695

Tab. 4.4: Cantidad de reseñas por condición

4.5. Gráficos adicionales

4.5.1. Comparación entre todos los pares de drogas por criterio de comparación

4.5.1.1. Comparación de similitudes (corpus TASA)

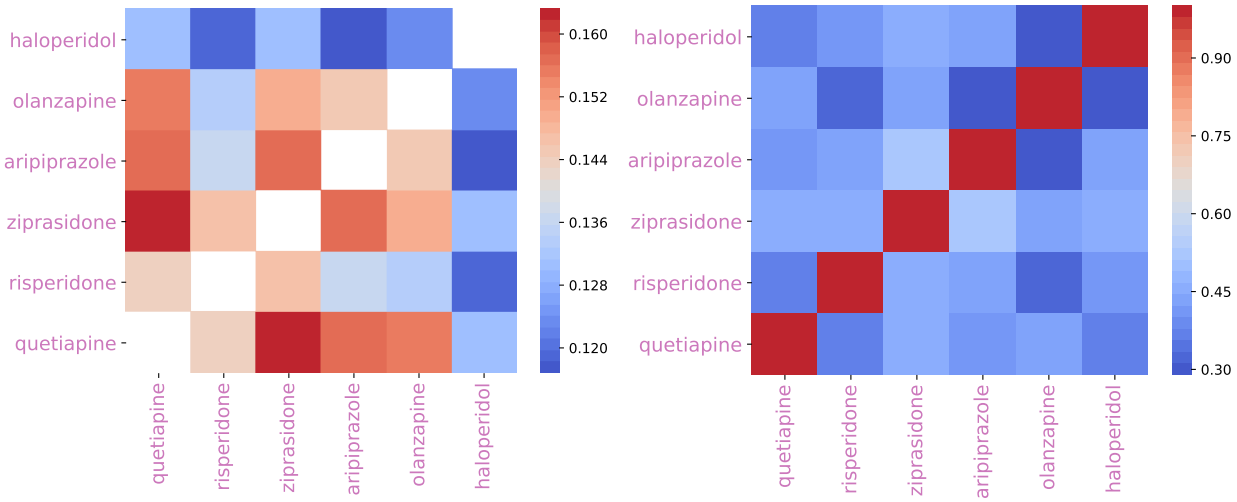


(a) Similitud estructural *(b)* Similitud de afinidad

Fig. 4.3: Comparación de similitudes para todo par de drogas.

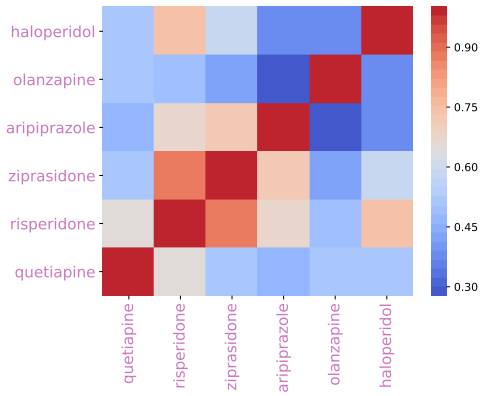
4.5.2. Comparación entre antipsicóticos

4.5.2.1. Comparación de similitudes (corpus TASA)



(a) Similitud semántica *(b)* Similitud estructural

Fig. 4.4: Comparación de similitudes semántica y estructural para todo par de drogas antipsicóticas con más de 100 reseñas. La similitud semántica surge de la utilización del corpus TASA.



(a) Similitud de afinidad

Fig. 4.5: Comparación de similitud de afinidad para todo par de drogas antipsicóticas con más de 100 reseñas.

4.5.2.2. Comparación de similitudes (LSA entrenado sobre corpus propio)

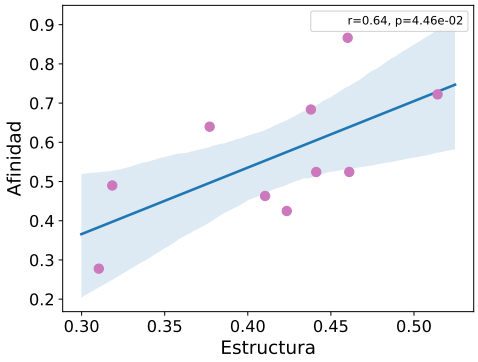
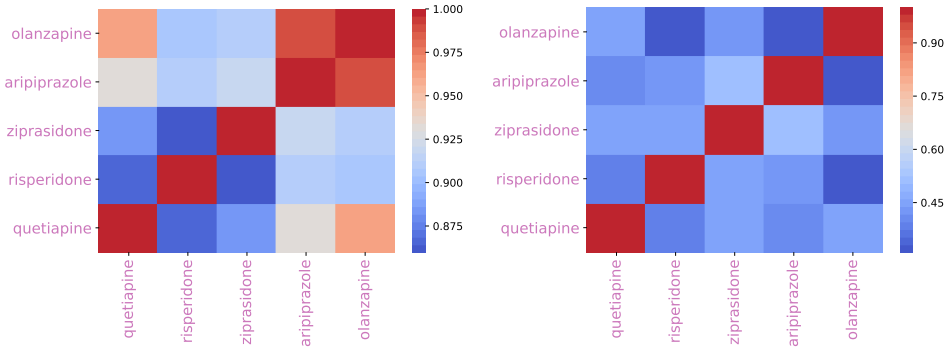
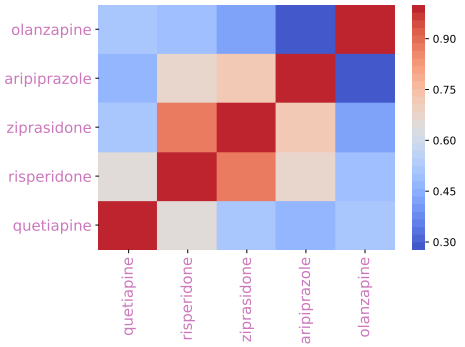


Fig. 4.6: Comparación entre similitudes estructural y de afinidad para todo par de drogas antipsicóticas con más de 100 reseñas.



(a) Similitud semántica *(b)* Similitud estructural

Fig. 4.7: Comparación de similitudes semántica y estructural para todo par de drogas antipsicóticas con más de 100 reseñas. La similitud semántica surge del análisis mediante la técnica LSA.

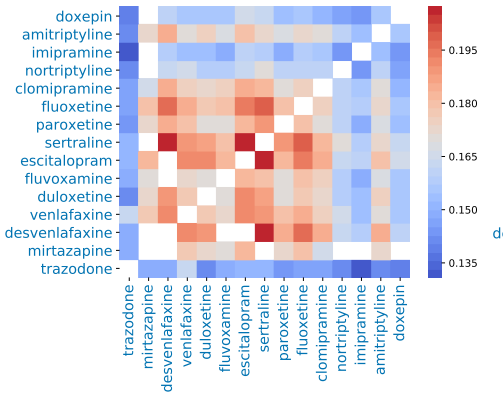


(a) Similitud de afinidad

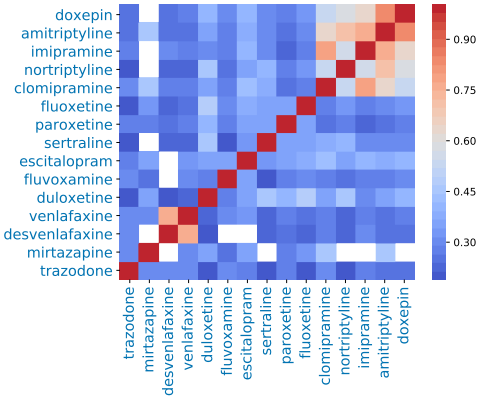
Fig. 4.8: Comparación de similitud de afinidad para todo par de drogas antipsicóticas con más de 100 reseñas.

4.5.3. Comparación entre antidepresivos (más de 100 reseñas por droga)

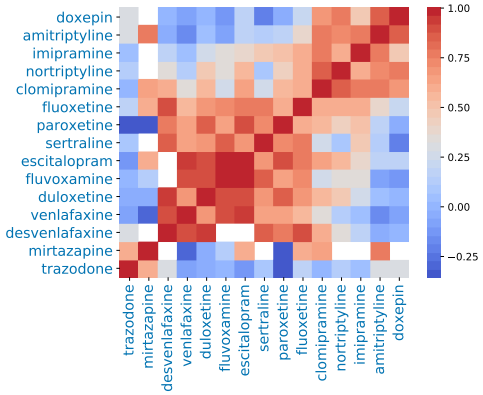
4.5.3.1. Comparación de similitudes (corpus TASA)



(a) Similitud semántica



(b) Similitud estructural



(c) Similitud de afinidad

Fig. 4.9: Comparación de similitudes para todo par de drogas antidepresivas con más de 100 reseñas. La similitud semántica surge de la utilización del corpus TASA.

4.5.3.2. Comparación de similitudes (LSA entrenado sobre corpus propio)

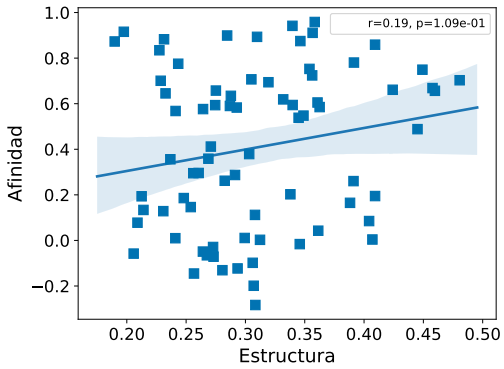
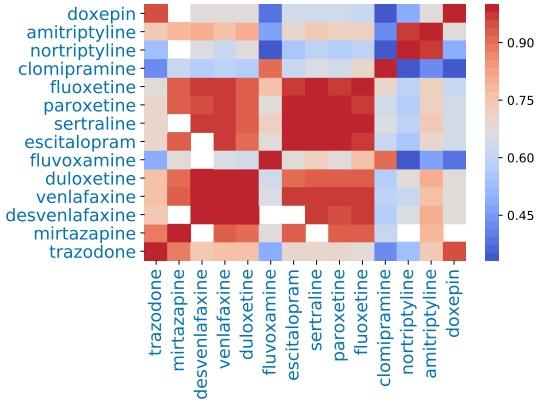
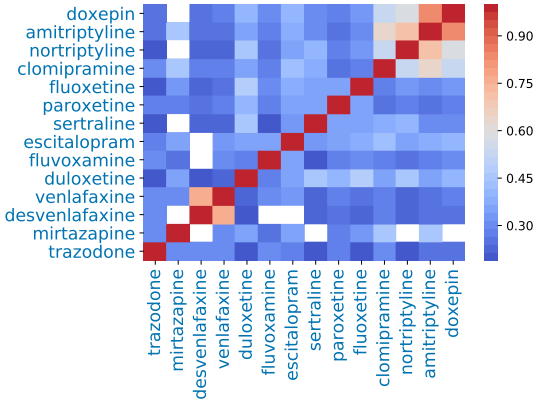


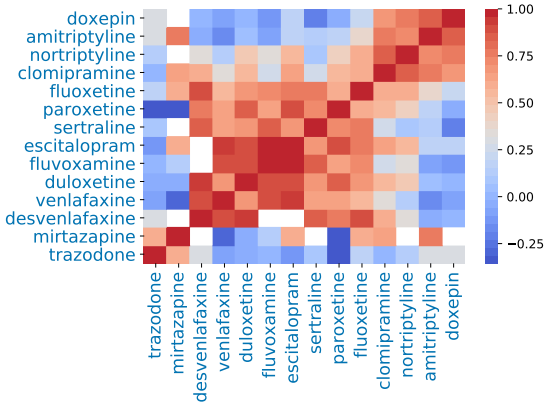
Fig. 4.10: Comparación entre similitudes estructural y de afinidad para todo par de drogas antidepresivas con más de 100 reseñas.



(a) Similitud semántica



(b) Similitud estructural

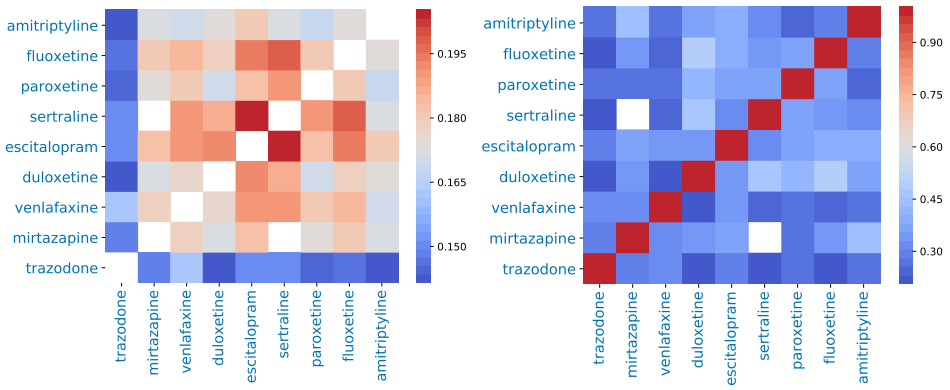


(c) Similitud de afinidad

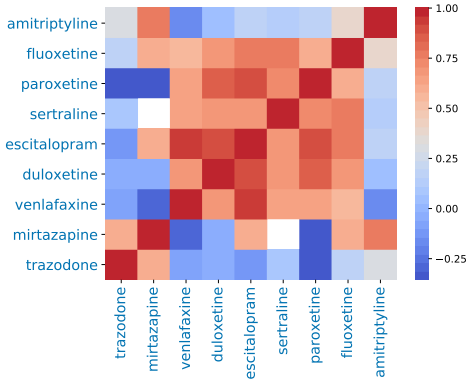
Fig. 4.11: Comparación de similitudes para todo par de drogas antidepresivas con más de 100 reseñas. La similitud semántica surge del análisis mediante la técnica LSA.

4.5.4. Comparación entre antidepresivos (con más de 1000 reseñas por droga)

4.5.4.1. Comparación de similitudes (corpus TASA)



(a) Similitud semántica *(b)* Similitud estructural



(c) Similitud de afinidad

Fig. 4.12: Comparación de similitudes para todo par de drogas antidepresivas con más de 1000 reseñas. La similitud semántica surge de la utilización del corpus TASA.

4.5.4.2. Comparación de similitudes (LSA entrenado sobre corpus propio)

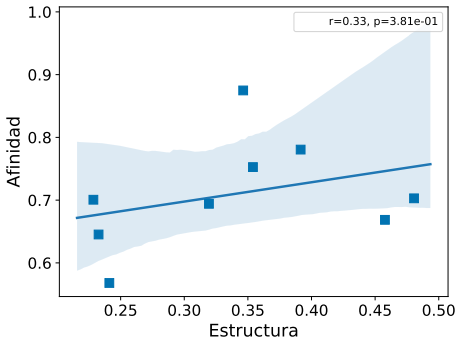
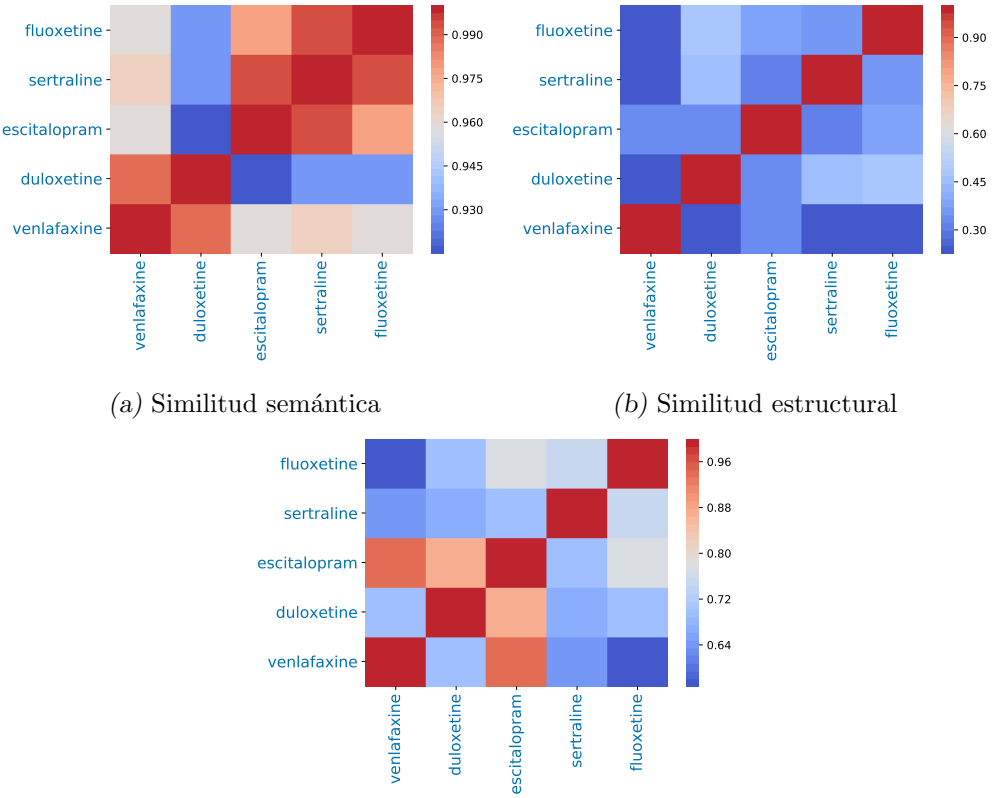


Fig. 4.13: Comparación entre similitudes estructural y de afinidad para todo par de drogas antidepresivas con más de 1000 reseñas.



(a) Similitud semántica

(b) Similitud estructural

(c) Similitud de afinidad

Fig. 4.14: Comparación de similitudes para todo par de drogas antidepresivas con más de 1000 reseñas. La similitud semántica surge del análisis mediante la técnica LSA.

4.5.5. Comparación de drogas prescriptas para una misma condición

4.5.5.1. Trastorno de ansiedad

Comparación de similitudes (corpus TASA)

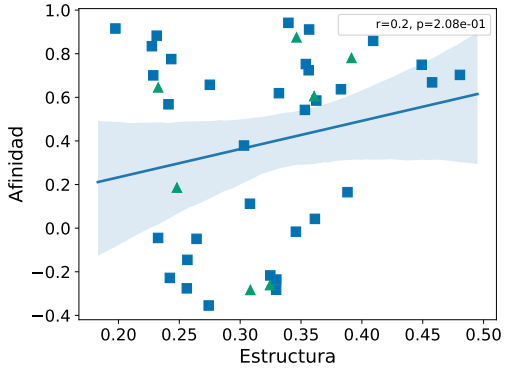
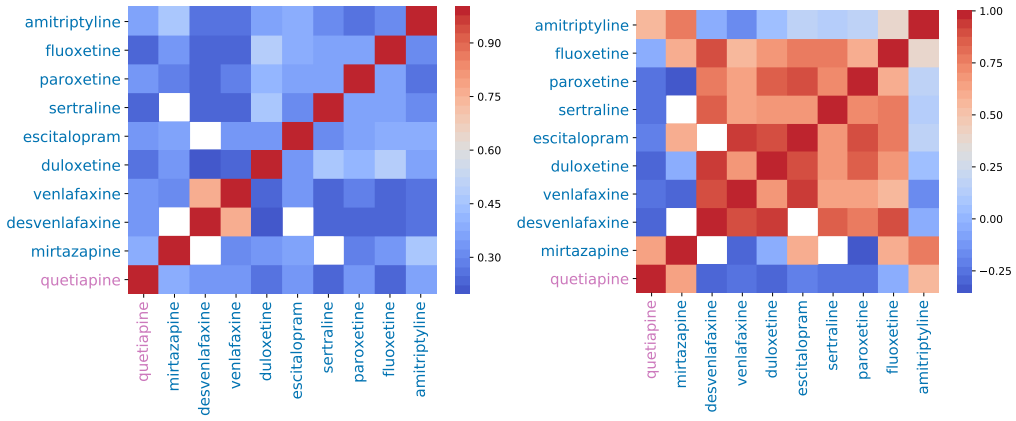


Fig. 4.15: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento del trastorno de ansiedad.



(a) Similitud estructural *(b)* Similitud de afinidad

Fig. 4.16: Comparaciones de similitud para todo par de drogas prescripto para el tratamiento del trastorno de ansiedad.

Comparación de similitudes (LSA entrenado sobre corpus propio)

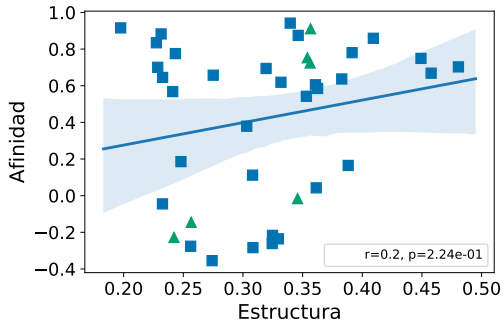
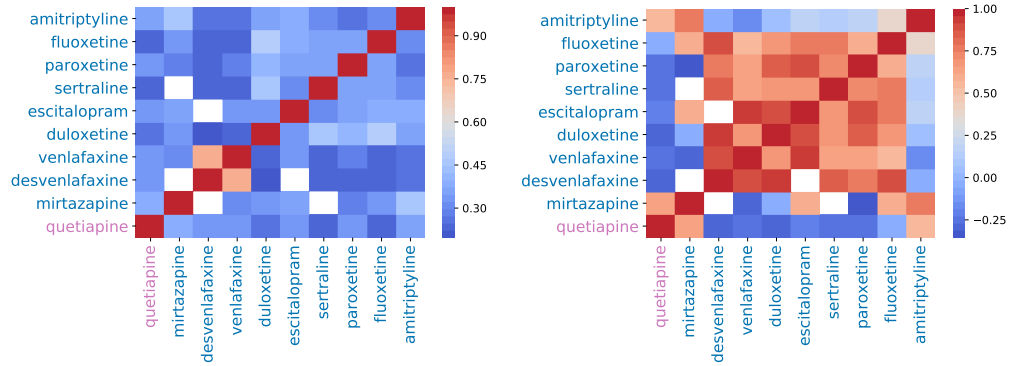


Fig. 4.17: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento del trastorno de ansiedad.



(a) Similitud estructural *(b)* Similitud de afinidad

Fig. 4.18: Comparaciones de similitud para todo par de drogas prescripto para el tratamiento del trastorno de ansiedad.

4.5.5.2. Trastorno Bipolar

Comparación de similitudes (corpus TASA)

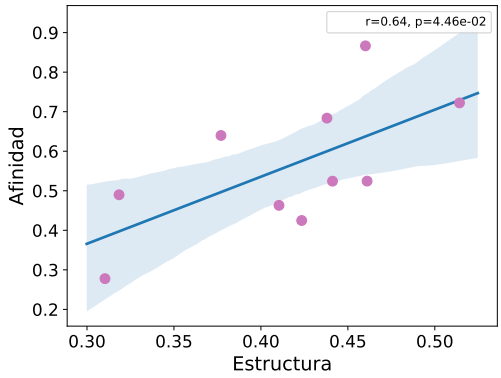
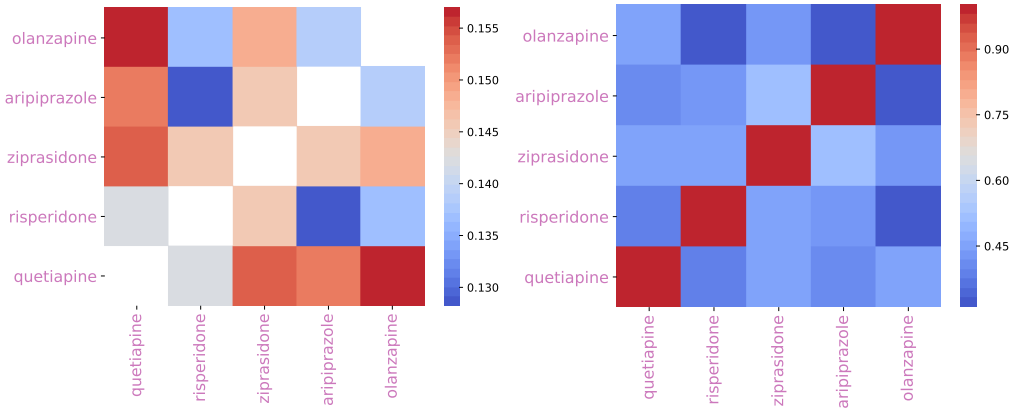
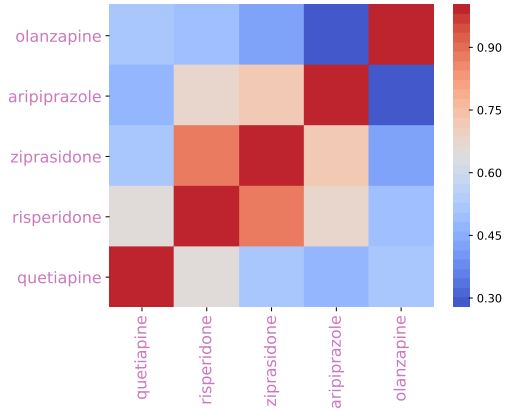


Fig. 4.19: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento del trastorno bipolar.



(a) Similitud semántica

(b) Similitud estructural



(c) Similitud de afinidad

Fig. 4.20: Comparaciones de similitud para todo par de drogas prescripto para el tratamiento del trastorno bipolar. La similitud semántica surge de la utilización del corpus TASA.

Comparación de similitudes (LSA entrenado sobre corpus propio)

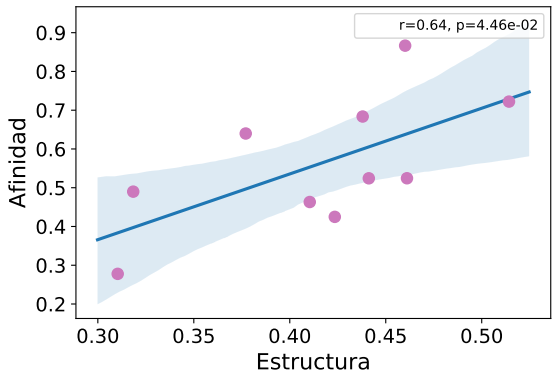
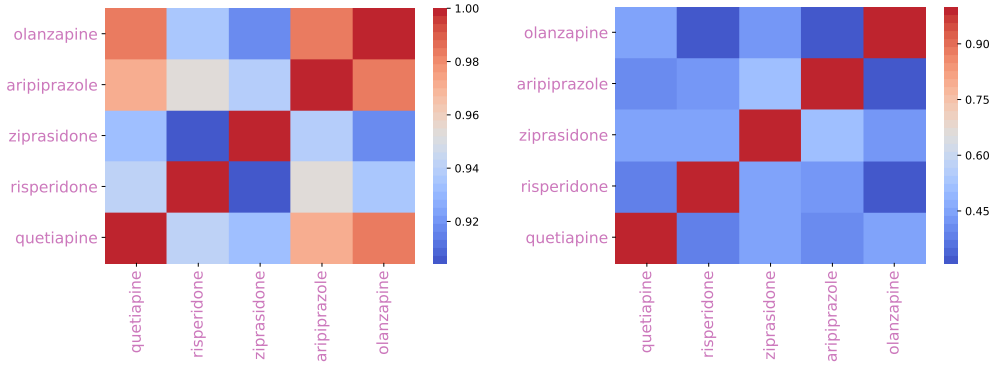
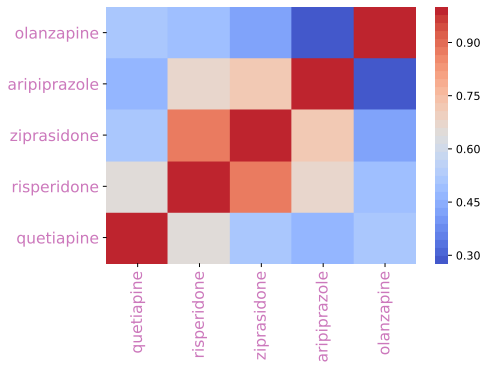


Fig. 4.21: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento del trastorno bipolar.



(a) Similitud semántica

(b) Similitud estructural



(c) Similitud de afinidad

Fig. 4.22: Comparaciones de similitud para todo par de drogas prescripto para el tratamiento del trastorno bipolar. La similitud semántica surge del análisis mediante la técnica LSA.

4.5.5.3. Depresión

Comparación de similitudes (corpus TASA)

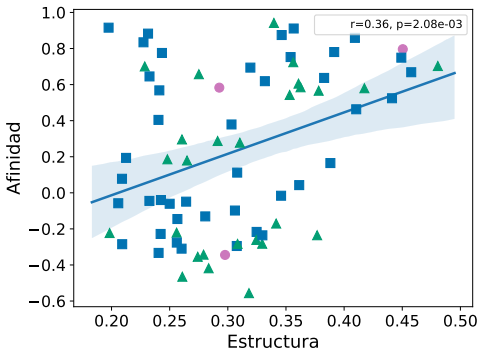


Fig. 4.23: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento de la depresión.

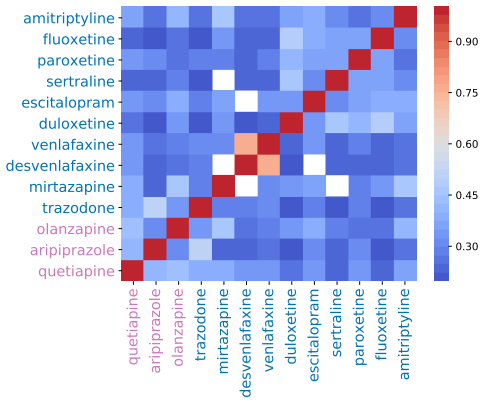


Fig. 4.24: Comparación de similitud estructural para todo par de drogas prescripto para el tratamiento de la depresión.

LSA

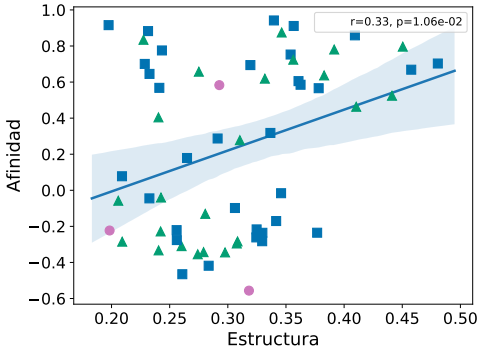
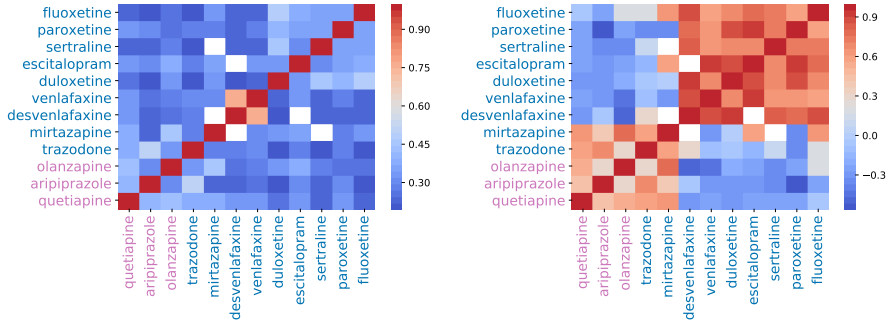


Fig. 4.25: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento de la depresión.



(a) Similitud estructural *(b)* Similitud de afinidad

Fig. 4.26: Comparaciones de similitud para todo par de drogas prescripto para el tratamiento de la depresión.

4.5.5.4. Esquizofrenia

Comparación de similitudes (corpus TASA)

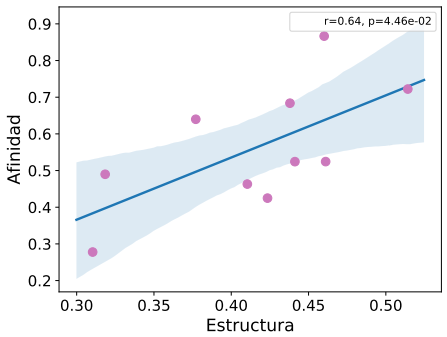
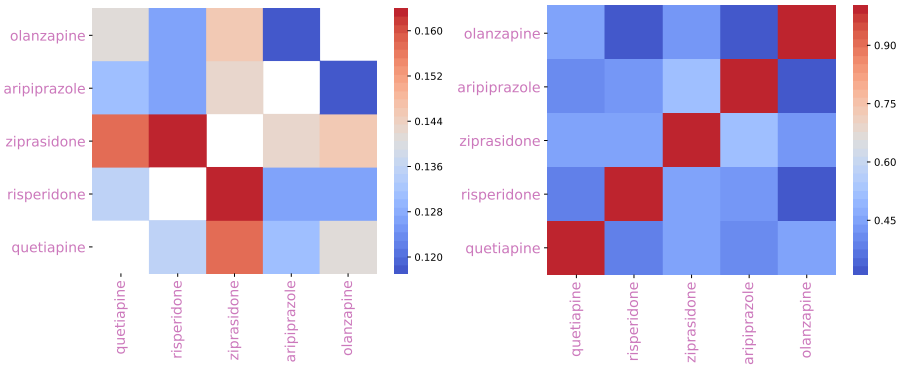
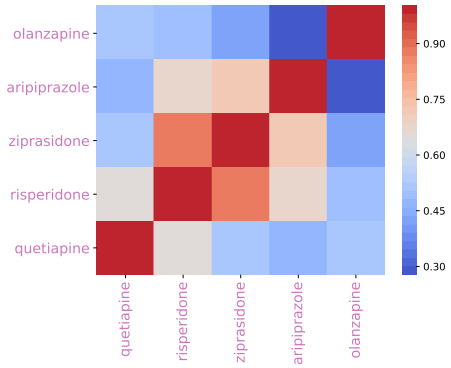


Fig. 4.27: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento de la esquizofrenia.



(a) Similitud semántica *(b)* Similitud estructural

Fig. 4.28: Comparaciones de similitud semántica y estructural para todo par de drogas prescripto para el tratamiento de la esquizofrenia. La similitud semántica surge de la utilización del corpus TASA.



(a) Similitud de afinidad

Fig. 4.29: Comparación de similitud de afinidad para todo par de drogas prescripto para el tratamiento de la esquizofrenia.

4.5.5.5. Subconjunto de condiciones no cognitivas

Comparación de similitudes (corpus TASA)

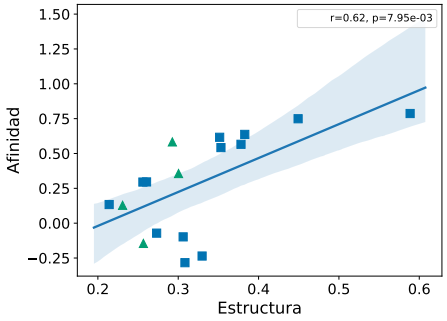
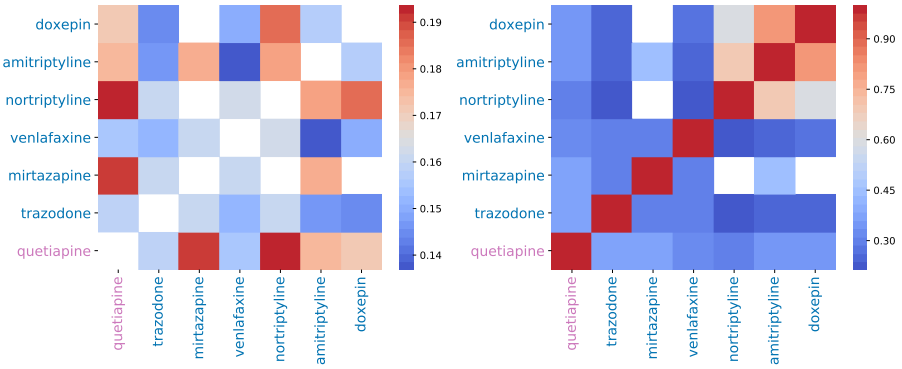
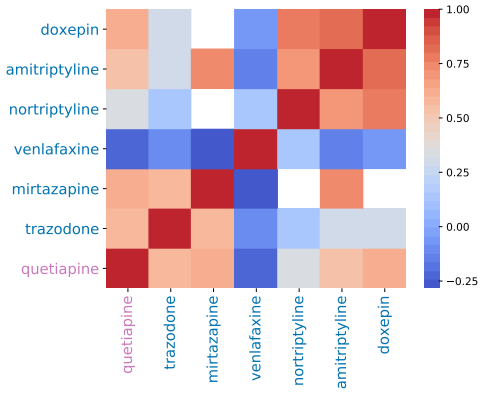


Fig. 4.30: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento de condiciones *no cognitivas*.



(a) Similitud semántica *(b)* Similitud estructural

Fig. 4.31: Comparaciones de similitud semántica y estructural para todo par de drogas prescripto para el tratamiento de condiciones *no cognitivas*. La similitud semántica surge de la utilización del corpus TASA.



(a) Similitud de afinidad

Fig. 4.32: Comparación de similitud de afinidad para todo par de drogas prescripto para el tratamiento de condiciones *no cognitivas*.

Comparación de similitudes (LSA entrenado sobre corpus propio)

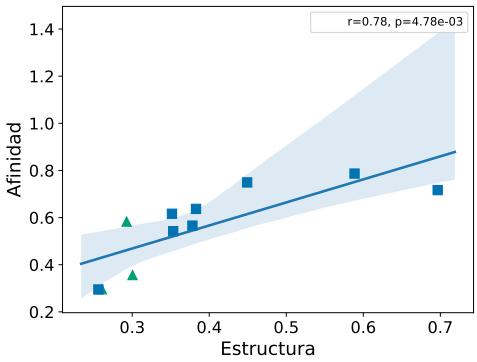
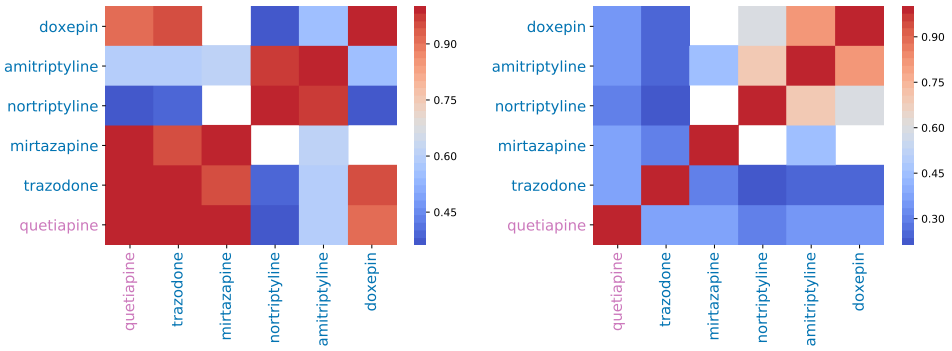
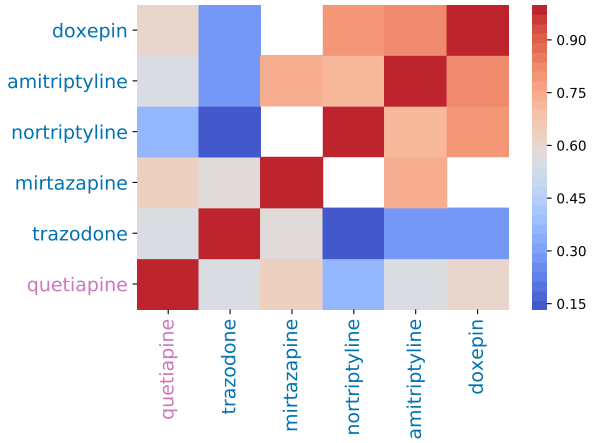


Fig. 4.33: Comparación entre similitudes estructural y de afinidad para todo par de drogas prescripto para el tratamiento de condiciones *no cognitivas*.



(a) Similitud semántica *(b)* Similitud estructural

Fig. 4.34: Comparaciones de similitudes semántica y estructural para todo par de drogas prescripto para el tratamiento de condiciones *no cognitivas*. La similitud semántica surge del análisis mediante la técnica LSA.



(a) Similitud de afinidad

Fig. 4.35: Comparaciones de similitud de afinidad para todo par de drogas prescripto para el tratamiento de condiciones *no cognitivas*.