Proceso:

Para el desarrollo del proyecto se usaron como base dos grandes colecciones de texto para estudios de depresión se trata de publicaciones (posts o comentarios) realizadas por un conjunto de personas en la red social Reddit -uno con casos clínicamente diagnosticados positivamente de depresión, y otro con casos diagnosticados negativamente-.[1]
En base a esto se extraen características como numero de palabras, numero de caracteres, numero de stopwords, promedio de caracteres por palabra, entre otros factores, así mismo como porcentajes relacionados a estas estadísticas.
Para hacer los análisis de coherencia, se usa CODRA [2], el cual es un framework en el cual se generan arboles con la estructura semántica de cada entrada. La naturaleza de procesamiento por medio de un contenedor docker es muy alta, y se necesita un mínimo de 6 GB de RAM para su correcto procesamiento, si se usa en estas condiciones mínimas el tratamiento de cada dato dura tres minutos – para lo cual es tratamiento de un corpus de mas de 1000000 de datos es improbable en tales condiciones. Debido a esto, se hace uso de la infraestructuta del Centro de Supercomputación de Galicia, por medio del Finisterrae se hace uso de la instalación del contenedor de CODRA, en este caso CESGA soporta únicamente Singularity que es una solución que replica el entorno basado en contenedores que utiliza Docker.
Usando este entorno se tuvo la capacidad de procesar 55892 datos del corpus negativo y 41104 datos del corpus positivo, de los cuales se pudieron extraer arboles de coherencia de 26797 y 18229 datos respectivamente de cada corpus.
En base a los arboles resultantes, se extraen varios tipos de estadísticas para cada corpus, las cuales señalan la estructura semántica que extrae el parser de codra entre las cuales están; la presencia de frases de núcleo, de contraste, de satélite, de habilitación, entre otros.
En una segunda parte de este análisis, también se extraen cuales son las estructuras semánticas mas comunes dentro de las frases, y asimismo como el nivel de profundidad que pueden resultar de los arboles semánticos resultantes para cada corpus de sentencias.
Asimismo, debido a la alta tasa de sentencias procesadas que no resultaron en un árbol de coherencia, se hace el re análisis para poder mirar las causas y razones tales frases no pudieron ser analizadas.

[1] Losada David E., Crestani Fabio, A Test Collection for Research on Depression and Language Use. In *International Conference of the Cross-Language Evaluation Forum for European Languages,* pages 28-39*,* 2019.

[2] Joty Shafiq, Carenini Giuseppe, Ng Raymond T. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. In *Computational Linguistics, Volume 41, Issue 3,* pages 385–435 2015.