Trabajo para Minería de datos de texto

Tema: Tópicos ocultos en torno a la temática de obesidad

Introducción

En la actualidad se está haciendo uso de las plataformas de redes sociales como fuente de información en diversas investigaciones en salud. Las mismas han producido resultados que de alguna forma revolucionaron los procesos de vigilancia y gestión en salud, entre otros, generando un cambio paradigmático en esta área y favoreciendo la medicina preventiva y la promoción de la salud (Organización Panamericana de la Salud [OPS], 2016). La información contenida en redes sociales ha demostrado ser de utilidad para predecir y explicar las características y estados de brotes de enfermedades (Ginsberg J, 2009). Si bien al comienzo estos estudios se orientaron al análisis de enfermedades transmisibles (Boon-Itt S, 2020; Ginsberg J, 2009), actualmente se está atendiendo a las enfermedades crónicas de naturaleza multicausal (Tsuya A, 2014; Lui CW, 2021). La plataforma de Twitter es una de las redes sociales más utilizadas para extraer información a partir de minería de datos de texto. Esto se debe a que les permite a los investigadores obtener un gran caudal de información generado por los usuarios y, además, contiene una API propia que facilita y controla el proceso (Tweepy, 2009). Desde la minería de datos de texto se trabaja con el procesamiento del lenguaje natural (PLN) (Cortéz Vasquez A, 2009), el cual abarca parte de la ciencia de datos, el aprendizaje automático (Machine Learning) y la lingüística. El PLN permite que las computadoras interactúen con el lenguaje humano, dándole significado para que pueda ser usado de manera práctica en el desarrollo de asistentes conversacionales, traducciones, análisis de sentimientos, modelado de tópicos, entre otros (Cortéz Vasquez A, 2009).

En el presente proyecto, herramientas de la ciencia de datos como son el PLN o el aprendizaje automático serán utilizadas para la recopilación, análisis y síntesis de un gran caudal de datos (tecnología Big Data), que permitirán identificar tópicos de discusión en torno a la temática de la obesidad. Se estaría de este modo prestando atención a las concepciones y representaciones de las personas en torno a esta patología, lo cual desde una perspectiva holística se considera de importancia para comprender las prácticas o resultados de salud de las poblaciones y los factores sociales que la condicionan.

Motivación: Específicamente, asumimos que existen tópicos sobre obesidad predominantemente discutidos en las plataformas sociales virtuales, cuyo conocimiento y análisis aportaría a la comprensión de los determinantes sociales de esa patología. Además, se espera que a partir de este proyecto se implementen herramientas de análisis como lo son el PLN y el aprendizaje automático para el estudio de los condicionantes sociales de la salud de las poblaciones, en este caso en relación a obesidad, pero potencialmente aplicable a otras enfermedades de naturaleza multicausal.

Planteamiento del problema: Ante la primera etapa del proyecto nos topamos con el primer problema a resolver: ¿Cuáles serán las palabras clave utilizadas para recolectar tweets?

La selección de las mismas debería ser sistemática y metodológica. Esto permite que otras personas puedan reproducir las técnicas aplicadas en este proyecto y además garantizar la representatividad de la base de datos a obtener. Esto se debe a que existen palabras características en ciertos discursos sociales, y la idea es poder abarcar todos esos discursos sin que quedé ninguno por visibilizar.

Meta: Generar una base de tweets que sea completa y relevante en torno a una temática particular, a partir de una palabra clave y sus derivadas.

Organización: Este repositorio se organizó de forma tal que en la carpeta “1. Metodología aplicada a la selección de palabras clave” se encuentran las técnicas aplicadas explicadas junto con las notebooks y sus respectivas salidas. En la carpeta “2. Conclusiones” se podrá encontrar las principales conclusiones de todo el trabajo elaborado y los problemas futuros a tratar.

**Selección de palabras clave para recolección de tweets en torno a la temática de obesidad**

Para comenzar la selección de *keywords* en relación a la obesidad se tuvo que realizar un diseño sistemático y metodológico, permitiendo asegurar la reproducibilidad de la técnica empleada y la representatividad en cuanto a términos de la base de tweets final. Para ello se indagaron y aplicaron distintas técnicas hasta finalmente seleccionar la más adecuada.

**1.1 Term frequency-inverse document frequency Tf-idf**

En una primera instancia se realizó la selección basándose en *Term frequency-inverse document frequency (*Tf-idf). Se trata de una técnica que mide con qué frecuencia aparece un término o frase dentro de un documento determinado, y lo compara con el número de documentos que mencionan ese término dentro de una colección de documentos. Asimismo, en la segunda parte de la denominación de estas siglas (idf) hablamos de *frecuencia inversa de documento,* donde se disminuye el peso de los términos que son muy frecuentes en varios documentos que se estén teniendo en cuenta. De esta forma, se pueden determinar cuáles son los términos más frecuentes y exclusivos en un documento en particular.

Para aplicarlo a la selección de *keywords* se decidió generar una base de datos con tweets provenientes de usuarios de Argentina, mientras que por otro lado ya se contaba de manera previo con un dataset que había recolectado tweets filtrando con palabras clave relacionadas a la temática seleccionadas a priori.

Para llevar a cabo esta metodología primero se eliminaron todos aquellos caracteres que generan ruido en los textos como son los emojis, las etiquetas HTML, las menciones, los hash-tags, los URLs, números, entre otros. Luego se tokenizaron las palabras de los tweets, esto consiste en dividir el texto en las unidades que lo componen, es decir, palabras. Después, cada palabra fue contada y ubicada en una tabla de doble entrada según el dataset de pertenencia. Aquellas ubicadas con mayor ocurrencia de término son las que pueden estar más altamente relacionadas con el tópico. Sin embargo, pueden aparecer términos que no están estrechamente relacionados como lo son las frases comunes utilizadas en un lenguaje.

Se contaron la cantidad de palabras, sin repetir, utilizadas en cada dataset y a su vez, la cantidad de palabras distintas usadas por cada tema. Después se procedió a eliminar las *stopwords*, algo que anteriormente no se había realizado. Las mismas se denominan en español “palabras vacias” y reciben ese nombre ya que no presentan significado como pueden ser artículos, pronombres, preposiciones, entre otras. Una vez eliminadas se determinaron las 10 palabras más frecuentes por dataset. Insatisfechxs con el resultado se decidió usar probabilidades de las palabras, donde la metodología al comienzo es la misma, se utilizan dos dataset donde uno contiene los tweets relacionados a la temática de interés mientras que el otro es general y no responde a una temática específica. La idea principal es comparar las probabilidades de cada palabra en cada dataset. De esta forma se estaría suponiendo que aquellas palabras que presentan mayores probabilidades de ocurrencia en un dataset es porque están altamente relacionadas al tópico.