Universidad de Puerto Rico
Recinto de Río Piedras
Facultad de Ciencias Naturales
Departamento de Ciencias de Cómputos

Revisión Literaria para Proyecto Final:

# Reddit Sentiment Analysis

Clase CCOM3031-0U1 Introducción a las ciencias de Datos
Profesora. Patricia Ordonez

Trabajo realizado por:
JASIEL A RIVERA-TRINIDAD
Michael H Terrefortes Rosado
José G. Portela González
ELIAM S RUIZ-AGOSTO

# Abstracto:

Las redes sociales proveen información invaluable para entender el comportamiento social de las generaciones modernas. Usar esta información para dar información valiosa a los propios usuarios sobre los grupos a los que se están uniendo presenta un gran beneficio a la hora de explorar el mundo de las redes digitales. En la actualidad paquetes de procesamiento de lenguaje natural y análisis de sentimiento como lo son NLTK y VADER nos permiten analizar estos datos de manera rápida y eficiente. Por esta razón, creemos que herramientas como la que se desarrolló para este proyecto deben ser impulsadas y desarrolladas y que nos pueden ayudar a entender problemas sociales tan importantes como el proceso electoral.

# Palabras Clave:

Análisis de sentimiento, procesamiento de lenguaje natural, NLTK, VADER, Redes sociales, Política

# Introducción:

Las redes sociales han tenido un impacto muy significativo en el comportamiento humano. Además,no parece que hayan detenido su crecimiento pues año tras año surgen nuevas plataformas sociales en el internet y crece el número de usuarios para las existentes. Este hecho implica que cada vez es más pertinente que se analicen los datos de las interacciones humanas sociales que ocurren en estas plataformas. La información que estos análisis pudieran proveer es muy valiosa y en este trabajo nos concentramos en discutir dos de ellas. La primera es como el analizar la naturaleza de los sentimientos expresados en las publicaciones digitales puede ayudar a los usuarios a decidir en cuales redes o comunidades dentro de las plataformas desea navegar. En este trabajo específicamente se discute el uso de Vader para analizar el valor sentimental de las publicaciones en las redes sociales. Con el fin de que los usuarios puedan tener información de cómo será el espacio social digital (positivo,negativo o neutral) antes de adentrarse en el mismo. El segundo tema que discutiremos es el uso de las redes sociales para discutir temas políticos. ¿Serán positivas, negativas o neutrales las discusiones políticas que se llevan a cabo en estas redes? ¿Son las redes sociales un medio apropiado para la discusión de temas políticos? creemos que esto son solo dos ejemplos de una cantidad inmensa de información que se podría obtener al analizar todas las publicaciones que se hacen a diario en las redes sociales.

# Uso de las redes sociales en la actualidad

Los noticieros modernos (en especial para las generaciones de los de los últimos 30-40 años) son las redes sociales. Un gran porcentaje de las parejas hoy en día se conocieron a través de una plataforma social digital o mantuvieron comunicación por un largo tiempo a través de una. En Puerto Rico se cuenta que hay personas que se enteran de asesinatos, terremotos, tormentas, nuevas leyes, protestas y etc. por facebook antes que a través de cualquier plataforma noticiera tradicional. Al observar la gran mayoría de las compañías hoy

en día exitosas a nivel mundial todas tienen un perfil digital sólido que se expande a través de múltiples plataformas. No podemos ignorar el poder de la información que fluye a través de las redes sociales. Según Hanlon & Bullock (2021), este año se reportaron 4.55 billones de usuarios activos de las redes sociales lo cual constituye un 57% de la población mundial con un crecimiento promedio de 9.9% por año. También ellos reportaron un promedio de 2:30 hrs al día y de interacción con 6 plataformas digitales sociales por usuario de internet(Según Hanlon & Bullock ,2021). Facebook es la plataforma más activa en términos de usuario con 2853 millones de usuarios, seguido de youtube con 2291 millones(Según Hanlon & Bullock ,2021). Reddit que es la aplicación que se seleccionó para este proyecto ocupa la posición número quince con 430 millones de usuarios (Según Hanlon & Bullock ,2021). Con un crecimiento de 10% anual no podemos ignorar el flujo de información que se da en estas plataformas por lo que se debe usar esta información para analizar el comportamiento humano y a su vez para crear guías/marcadores claros para que los usuarios sepan en que se envuelven a la hora de unirse a una red o a una comunidad dentro de ellas.

# Uso de NLTK y VADER

Para este proyecto se utilizó un paquete de herramientas llamado “Natural Language Tool Kit (NLTK)” por su nombre en inglés. El mismo provee herramientas para procesar con una computadora el lenguaje natural. En este caso se define el lenguaje natural como el conjunto de sintaxis y símbolos literarios dados a fonemas (letras, palabras) por los seres humanos para dar sentido a la comunicación escrita. Dentro de NLTK residen varias bibliotecas con una variedad de propósitos y para esta investigación se utilizó VADER. El mismo significa “Valence Aware Dictionary” por su nombre en inglés(Adarsh et.al,2019). El mismo utiliza un acercamiento de “lexicon” lo que esencialmente significa que existe un diccionario de palabras inmenso donde las palabras se les asigna un valor/característica con el cual están relacionados (Adarsh et.al,2019). El mismo puede ser de orientación semántica, polaridad y etc. En el caso de este proyecto específico se le asignan valores de sentimentalidad. Es decir, si las mismas son de carácter positivo, negativo o neutral. Este paquete y modelo es usado con muchos fines a través de la industria en la actualidad para realizar muchos tipos de análisis y es una herramienta muy útil.

En el contexto de nuestra investigación tener herramientas de procesamiento de lenguaje natural es muy útil. Esto se debe a que las publicaciones en  Reddit (la red social seleccionada para realizar este trabajo) están altamente compuestas por lenguaje natural (texto). De haber seleccionado otra plataforma como tal vez instagram (cuyas publicaciones en su mayoría son de tipo fotográfico) su utilidad (aunque no anulada) hubiera sido más limitada. El analizar el texto dentro de las comunidades (subreddits) dentro de Reddit en un contexto sentimental (positivo, negativo o neutral) nos permite realizar un análisis para determinar si en una comunidad domina un sentimiento sobre otro o si la misma es balanceada. Esto permitiría a un usuario entender en que se está envolviendo antes de entrar a la comunidad. Además permite analizar si, por ejemplo, el lenguaje en comunidades políticas es cortes o vulgar lo cual a su vez permite evaluar que tipo discusiones (cívicas o anárquicas) se llevan a cabo en estas plataformas. Esto podría ser una herramienta clave para entender el comportamiento electoral y la influencia de las discusiones en las redes sociales sobre las

mismas. En fin este paquete es uno actualizado y una herramienta clave para el desarrollo de este proyecto.

## EL uso de redes sociales para discusiones políticas

Como visto el uso de Reddit y el uso de la herramienta Vader en la programación han sido de gran ayuda para el proceso de analizar e identificar el uso de palabras y como percibirlas a través de las redes sociales. Ahora en la sociedad progresivamente polarizada de hoy, los usuarios de las redes sociales están cada vez más expuestos a comentarios flagrantes y descortés, opiniones disonantes y contenido de noticias controvertido. Se ve un auge de cómo los ciudadanos hoy tienen más oportunidades de adquirir o encontrar información de contenido noticiero o político. (Goyanes,2021) Esto es de suma importancia y relación al análisis sentimental trabajado por que como previamente explicado y discutido Reddit es a base de publicaciones de los mismos usuarios registrados a la página web. Lo que quiere decir que las ideas, sugerencias, planteamientos o debates que se discute son a base de una idea que en la actualidad es mayormente adquirida a través de las redes sociales. La percepción general de la política asociada es negativa. (Gil de Zúñiga,2017) Aunque la disponibilidad de dicha información puede beneficiar a la democracia o cualquier otro tipo de gobierno de varias formas, en nuestro caso a través de las publicaciones de subreddits, también puede contribuir a la percepción errónea de los ciudadanos que pueden informarse adecuadamente sobre la política sin mucho esfuerzo político.

# Resultados

En el subreddit de Política pudimos ver cómo había más publicaciones negativas en general. La muestra se terminó de descartar las 500 publicaciones de la parte superior caliente. Esto significa las publicaciones que están en tendencia en este momento. Como se ve en la Figura 1, hubo más publicaciones negativas que positivas o neutrales. Analizamos las palabras frecuentes de la publicación que procesamos mediante análisis de sentimiento. La frecuencia de las palabras principales utilizadas positivamente se muestra en la Figura 2. Según este diagrama de barras, la palabra Biden es la palabra más utilizada positivamente. Mientras Trump viene después, luego president, news y joe. Más adelante en el estudio también analizamos las palabras negativamente más utilizadas en este subreddit. Los resultados se muestran en la Figura 3. Las palabras más utilizadas negativamente fueron Trump, twitter, news, capitol y house. En este subreddit el 27.8% de las publicaciones fueron positivas, el 49% negativas y el 23.2% neutrales.

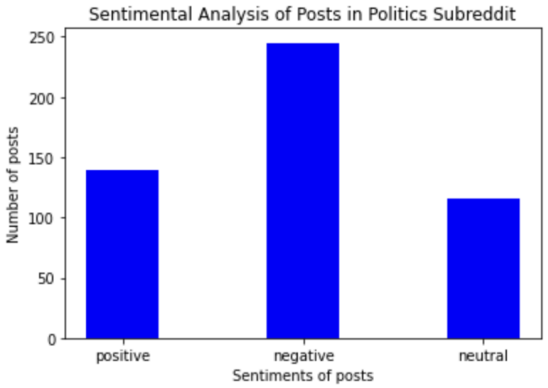


***Figura 1***

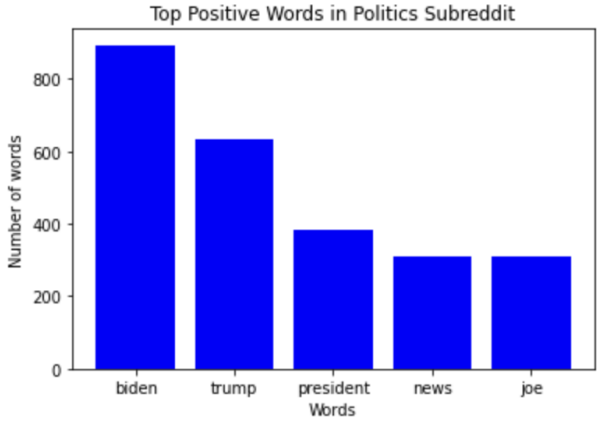


***Figura 2***

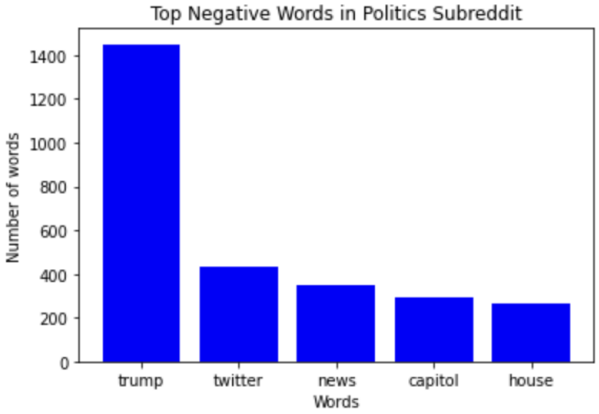


***Figura 3***

El siguiente subreddit analizado fue World News. En este subreddit pudimos ver cómo la mayoría de las publicaciones analizadas eran negativas. Usamos la misma cantidad de publicaciones, 500, y también las publicaciones de tendencia en este momento. El número de publicaciones negativas, positivas y neutrales se muestra en la Figura 4. Cuando se trata de la palabra positiva más utilizada, los resultados fueron Trump, coronavirus, new, us, covid. Esto se ve en la Figura 5. Las palabras de uso negativo más frecuentes, en la Figura 6, fueron Trump, us, china, says y world. Como podemos ver, en este subreddit la mayoría de las

publicaciones fueron negativas. En este subreddit el 27.6% de las publicaciones fueron positivas, el 52.4% negativas y el 20% neutrales.

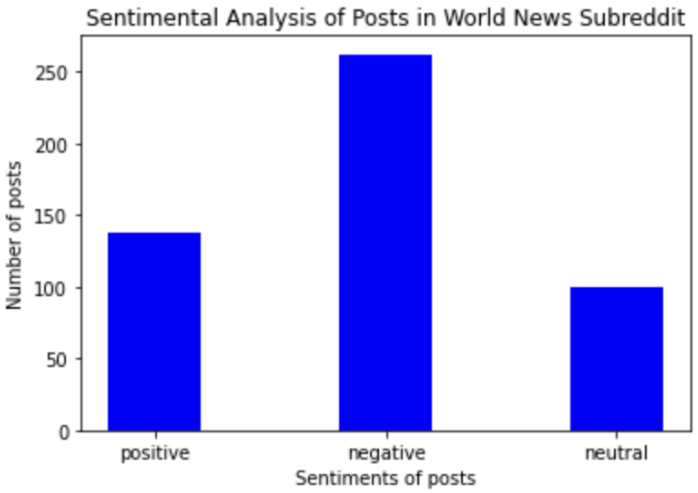


***Figura 4***

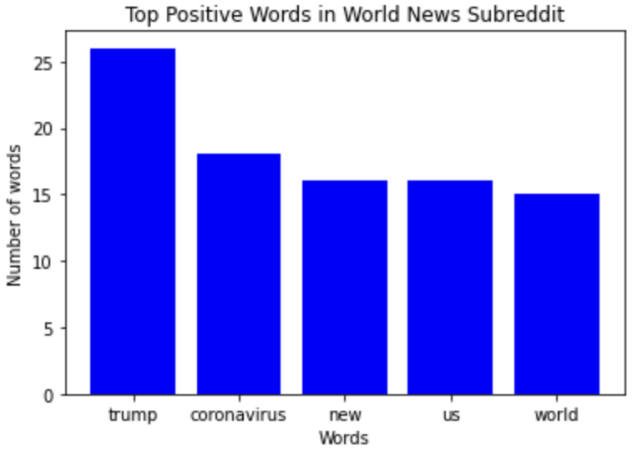


***Figura 5***

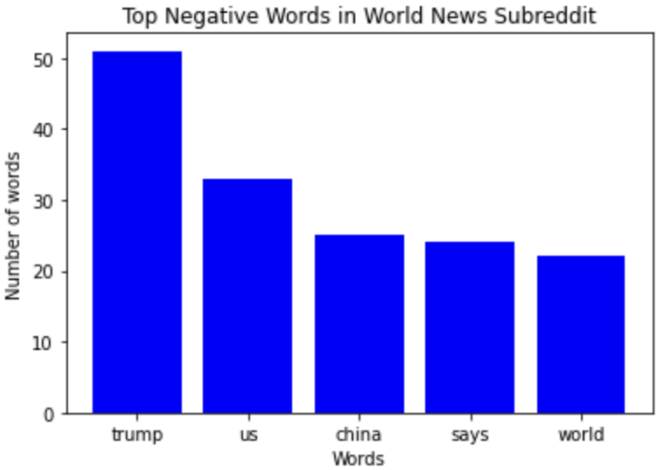


***Figura 6***

Más tarde decidimos analizar subreddits que estaban más afiliados políticamente en lugar de un subreddit de política general o noticias. El subreddit demócrata tiene más publicaciones neutrales que negativas o positivas. Esto se ve en la Figura 7. Las palabras de uso frecuente positivo, como se ve en la Figura 8, fueron Trump, president, vote, biden y one. Esta información es importante para ver qué tan negativos pueden ser ciertos subreddits, y también tal vez sesgos en ellos. Ahora en la Figura 8 graficamos las palabras frecuentemente negativas. Estas palabras fueron: triunfo, republicano, twitter, pueblo y presidente. En este subreddit el 28.2% de las publicaciones fueron positivas, el 31% negativas y el 40.8% neutrales.

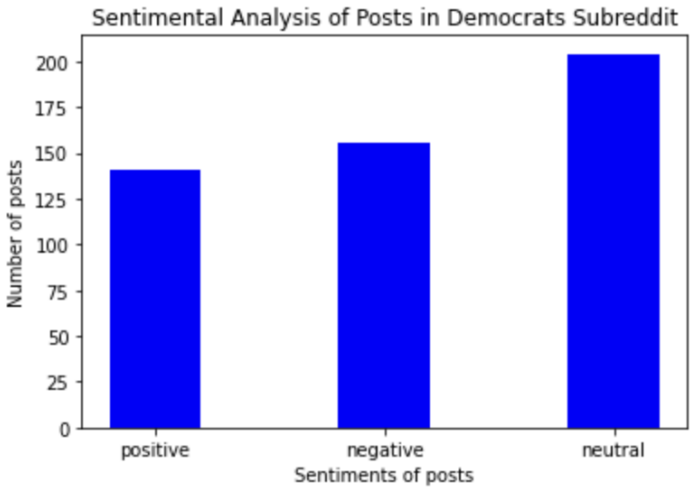


***Figura 7***

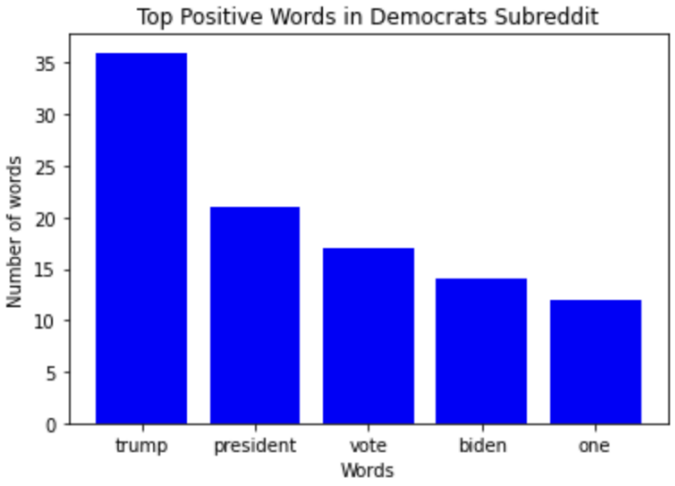


***Figura 8***

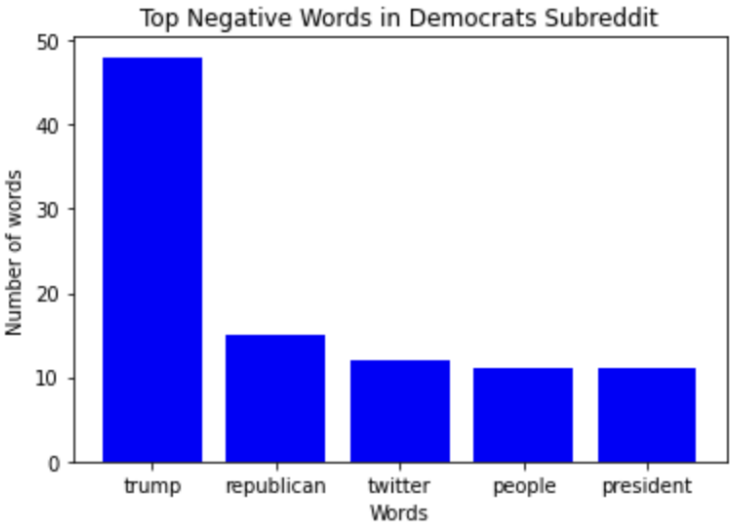


***Figura 9***

Otro subreddit políticamente afiliado que analizamos es Conservative. En la Figura 10, podemos observar cómo tanto el negativo como el neutral se ven casi idénticos. En realidad había 171 puestos neutrales y 170 negativos, por lo que había más puestos neutrales. Cuando se trata de las principales palabras frecuentemente positivas, en la Figura 11, podemos ver cómo trump, like, bill, people y biden se usaron más positivamente. Mientras que en la Figura 12 podemos ver las palabras principales utilizadas de manera más negativa. En esta gráfica, trump, people, ban, biden y hate fueron los más utilizados en las publicaciones negativas. En este subreddit el 31.8% de las publicaciones fueron positivas, el 34% negativas y el 34.2% neutrales.

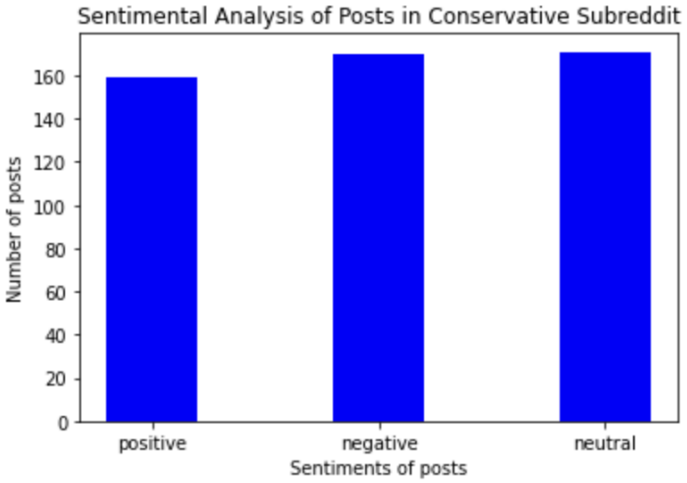


***Figura 10***

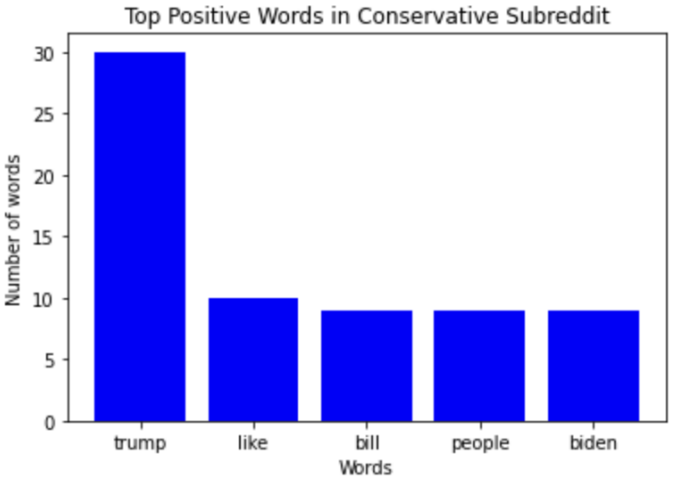


***Figura 11***

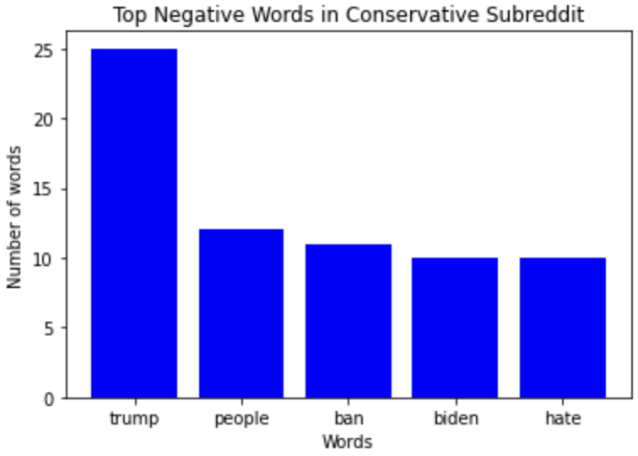


***Figura 12***

Ahora, en nuestro último subreddit políticamente afiliado, elegimos Libertarian. En la Figura 13 vemos cómo lo positivo y lo negativo se ven casi idénticos. Pero en realidad hubo 172 publicaciones positivas y 175 negativas. Entonces, en este subreddit hubo más publicaciones negativas. Hubo un 35% de publicaciones negativas, mientras que el 34.4% fueron positivas y el 30.6% fueron neutrales. Fue muy interesante analizar este subreddit y ver cómo se compara con los subreddits anteriores.

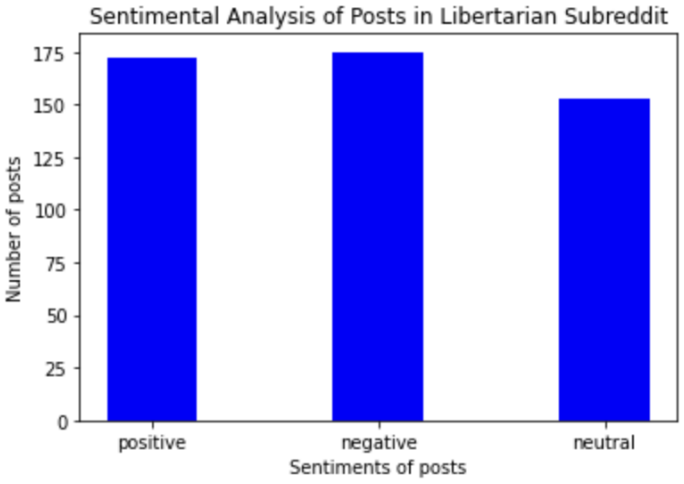


***Figura 13***

Ahora graficamos cada resultado de lado a lado para analizar cuál fue más negativo o positivo. En la Figura 14, vemos cómo tanto los subreddits de Política como los de World News fueron en su mayoría negativos. El subreddit más neutral fue Demócratas y el más positivo fue Libertario. Este análisis muestra que los subreddits que están menos afiliados políticamente, los subreddits generales, como Politics y World News, fueron los más negativos. Esto podría explicarse porque estos subreddits son más accesibles en línea y no están restringidos a ciertas partes de diferentes países. También nos dice cómo debemos tener cuidado con esos subreddits por lo negativos que son.

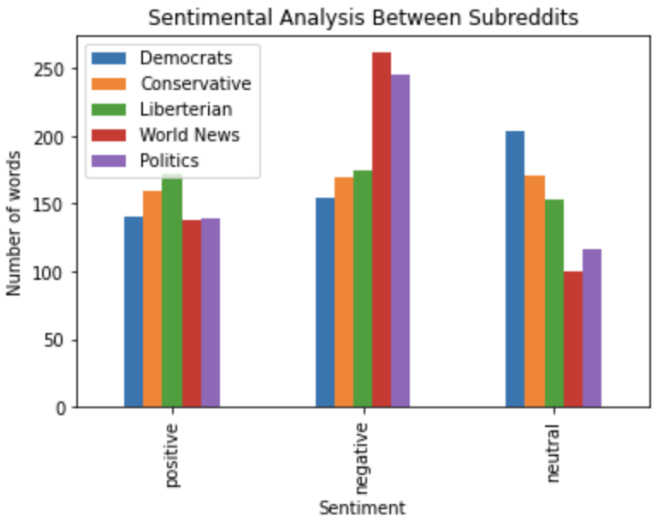


***Figura 14***

# Conclusión:

Como parte de este trabajo creamos una herramienta que provee datos sobre cualquier comunidad en Reddit. Estos datos se pueden utilizar para hacer análisis más complejos como fue demostrado en nuestro trabajo investigativo. Dado este análisis podemos ver cuales comunidades políticas son más negativas y qué comunidades son más positivas. Podemos ver

cuales son los temas de conversaciones de las mismas y en que contextos aparecen. Cabe recalcar que esta herramienta no está limitada a comunidades políticas, cualquier persona puede usar esta herramienta para crear otros tipos de análisis con cualquier comunidad. Es importante comprender que en las palabras positivas o negativas de uso frecuente superiores, el contexto no se usa cuando se analiza. Por ejemplo, en los subreddits políticos tal vez un político pueda aparecer con una nota positiva, pero esto no significa que las personas en el subreddit están hablando positivamente de esa persona. Tal vez una de las publicaciones dice cómo se elimina o cambia una ley de cierto político y luego se analiza como positiva o negativa.

Referencias

Goyanes, M., Borah, P., & de Zúñiga, H. G. (2021). Social media filtering and democracy: Effects of social media news use and uncivil political discussions on social media unfriending. Computers in Human Behavior, 120, 106759.

Adarsh, R., Patil, A., Rayar, S., & Veena, K. M. (2019). Comparison of VADER and LSTM for sentiment analysis. International Journal of Recent Technology and Engineering, 7(6), 540-543.

Hanlon, A., & Bullock, L. (2021, December 6). *Global Social Media Statistics Research Summary 2022*. Smart Insights. Retrieved December 19, 2021, from https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

Gil de Zúñiga, H., & Liu, J. H. (2017). Second screening politics in the social media sphere: Advancing research on dual screen use in political communication with evidence from 20 countries. *Journal of broadcasting & electronic media*, *61*(2), 193-219.

Perrin, A. (2015). Social media usage. *Pew research center*, *125*, 52-68.

Miller, P. R., Bobkowski, P. S., Maliniak, D., & Rapoport, R. B. (2015). Talking politics on Facebook: Network centrality and political discussion practices in social media. *Political Research Quarterly*, *68*(2), 377-391.