Tarea 2

Jose Adrian Castillo Sierra

May 2022

1 Introducción

Reddit es una red social de discusión gestionado por la propia comunidad. Una plataforma social en la que los usuarios envían publicaciones que otros usuarios pueden votar según sus preferencias. Si una publicación recibe muchos votos, sube en la clasificación de Reddit y, aumentando su alacance al publico de esta manera; si recibe votos negativos, su alcance se reduce y desaparece de la vista de la mayoría de los usuarios.

Reddit se gestiona en grupos o subreddits. Cualquier usuario puede crear subreddits sobre cualquier tema, ya sea un asunto general, como tecnología, o específico, como una simple broma. Cada subreddit pasa a formar parte de la lista completa de envíos de Reddit, lo cual significa que una publicación en cualquier subreddit puede llegar a la página principal del sitio web.

Para la elaboración de esta tarea se hizo un análisis de sentimiento sobre las palabras mas usadas en la ultima semana en los post mas populares del subreddit 'TIFU' *Today I F*cked Up*. Un subreddit donde diariamente la gente cuenta desde el anonimato los errores que han cometido así como los usuarios opinan sobre el suceso. El análisis sobre este subreddit se escogió principalmente debido a que su contenido suele tener una inclinación hacia lo negativo y al uso de groserias este puede ser el indicador sobre el desempeño de nuestro análisis.

2 Desarrollo

2.1 Obtención y Pre-procesamiento de la Información

La información utilizada para la realización del análisis de texto fue la misma utilizada para la realización de la tarea anterior. Por lo tanto, lo único que se realizó como diferencia fue el acomodo de cada comentario del subreddit dentro de una estructura de DataFrame y posteriormente aplicar las técnicas de limpieza realizadas con anterioridad a cada registro.

2.2 Análisis de Sentimiento

El análisis de sentimientos es una minería de texto que identifica y extrae información subjetiva en el material de origen y ayuda a una empresa a comprender el sentimiento social de su marca, producto o servicio.

Comprender cómo se sienten los clientes acerca de una marca o productos es esencial. Esta información puede ayudar a mejorar la experiencia del cliente o identificar y solucionar problemas con sus productos o servicios.

Para el análisis de texto se hizo utilización de tres librerías distintas de Python, siendo estas, TextBlob, Vader y SentiWordNet. Cada una siendo utilizada para clasificar los mismos comentarios realizados en Reddit. Dado que cada una de las tres librerías evalúa internamente de forma distinta las proporciones de las clasificaciones, positiva, negativa o neutral, son también muy distintas, tal y como podemos observar en la Figura 1.

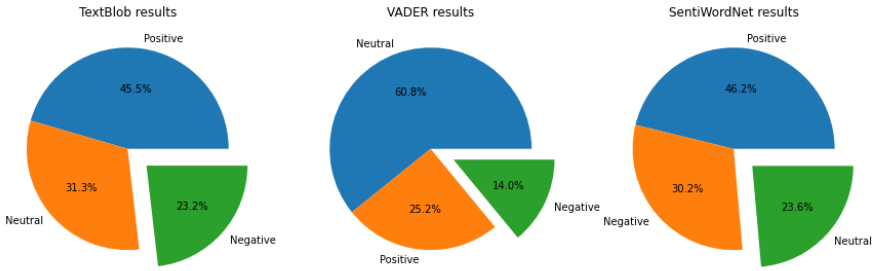

Figure 1: Clasificación de texto por librería

Siendo completamente diferentes las clasificaciones no podemos evaluar cual de las tres opciones que tenemos es la que mas se adecua a nuestra información. Es por eso que tendremos que utilizar otra métrica para evaluar las clasificaciones obtenidas; en este caso revisaremos las palabras mas comunes, positivas y negativas, en cada modelo con el objetivo de detectar algún comportamiento anormal dentro de la clasificación.

Como podemos ver la Figura 2 en la clasificacion de las palabras positivas para el modelo realizado con SentiWordNet la palabra *fuck* se encuentra entre las palabras mas utilizadas, algo que no se encuentra correctamente debido a que es una palabra normalmente negativa, por lo que podemos descartar esta opción.

TextBlob

	Common_words	count
0	get	3306
1	like	3097
2	go	2303
3	good	2062
4	make	2040
5	one	1914
6	time	1880
7	say	1869
8	think	1790
9	know	1784
10	would	1705
11	thing	1514
12	people	1475
13	take	1321
14	want	1292
15	work	1204
16	need	1095
17	really	1094
18	even	1076
19	lol	1046

Vader

	Common_words	count
0	like	2569
1	get	2130
2	go	1479
3	make	1429
4	good	1372
5	say	1267
6	one	1242
7	time	1193
8	know	1156
9	would	1144
10	think	1125
11	thing	1001
12	want	959
13	people	929
14	take	886
15	feel	835
16	well	814
17	work	794
18	love	754
19	need	737

SentiWordNet

	Common_words	count
0	get	3657
1	like	2767
2	go	2141
3	good	1937
4	one	1815
5	make	1790
6	say	1751
7	know	1687
8	time	1646
9	think	1629
10	would	1619
11	want	1447
12	people	1392
13	thing	1357
14	well	1245
15	really	1196
16	take	1172
17	even	1127
18	work	1066
19	fuck	1066

Figure 2: Palabras Comunes Positivas

Continuando con el análisis de frecuencias pero ahora con las negativas, Figura 3, podemos observar que la segunda palabra mas frecuente en el modelo de TextBlob es *like*, una palabra comúnmente con conotación positiva y que para el mismo modelo este era la segunda palabra mas utilizada para los comentarios positivos; en base a estas observaciones podemos descartar también el modelo de TextBlob, quedándonos con el modelo de Vader.

TextBlob

	Common_words	count
0	get	1710
1	like	1443
2	fuck	1145
3	go	1110
4	bad	922
5	say	916
6	make	903
7	would	867
8	one	851
9	know	826
10	think	807
11	time	790
12	people	753
13	shit	716
14	take	641
15	thing	611
16	mean	592
17	feel	555
18	want	499
19	kid	492

Vader

	Common_words	count
0	get	1258
1	fuck	960
2	go	819
3	shit	693
4	like	674
5	bad	665
6	say	619
7	make	615
8	one	595
9	would	577
10	people	572
11	know	564
12	time	555
13	think	542
14	take	456
15	thing	437
16	feel	371
17	even	356
18	someone	347
19	could	343

SentiWordNet

	Common_words	count
0	like	1795
1	get	1769
2	go	1364
3	make	1194
4	say	1095
5	one	1044
6	think	1033
7	time	1029
8	would	1019
9	know	958
10	people	891
11	bad	880
12	never	850
13	take	837
14	thing	767
15	shit	709
16	feel	681
17	day	680
18	work	647
19	use	630

Figure 3: Palabras Comunes Negativas

3 Conclusión

Los resultados obtenidos sobre el análisis del subreddit 'TIFU' *Today I F*cked Up* fueron algo sorpresivos, ya que debido al nombre de este esperaba que la proporción de resultados negativos fuera la mayor. Y en base al análisis realizado con el modelo Vader se pudo observar que en su mayoría son comentarios neutrales y positivos, componiendo con un 14% los comentarios negativos.

Finalmente, observando la lluvia de palabras de la Figura 4 podemos darnos cuenta de las palabras que mas resaltan tanto positivas como negativas y confirmando los resultados del análisis del texto. En base a que algunas de las

palabras que resaltan en la nube positiva se incluyen palabras como *love, like y happy*, mientras que en la negativas destacan insultos en las mas utilizadas, confirmando así una de las hipótesis que se tenían.

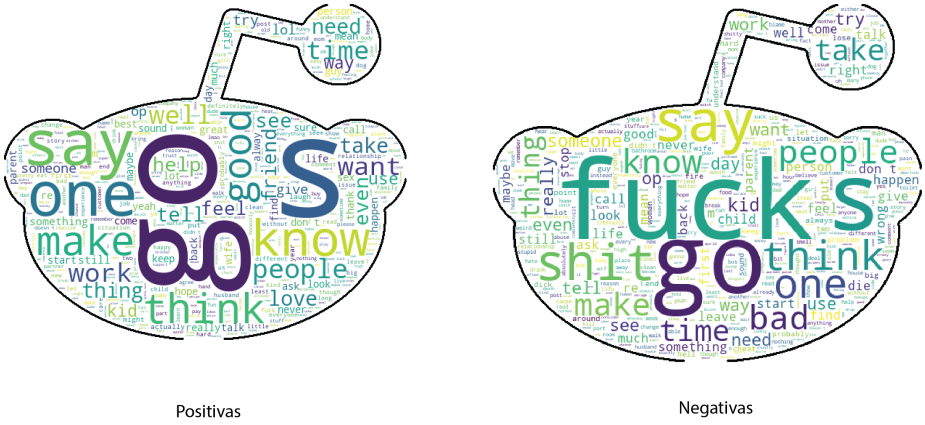

Figure 4: Wordcloud

References

[1] *R/mexico*. URL: `https://www.reddit.com/r/mexico`.

[2] Steelfenix. *Procesamientodatos/Tarea 2 at main · steelfenix/procesamientodatos*. URL: `https://github.com/Steelfenix/ProcesamientoDatos/tree/main/Tarea%202`.

[3] *The python reddit api wrapper*. URL: `https://praw.readthedocs.io/en/stable/`.