Caso 3: Sentimientos en Twitter

Máster en Big Data

Sebastian Cueva
Pol Gràcia

Todo el Código y desarrollo del caso 3 se puede encontrar en el siguiente repositorio de github:
https://github.com/sebasfire3/Twitter

ANÁLISIS:

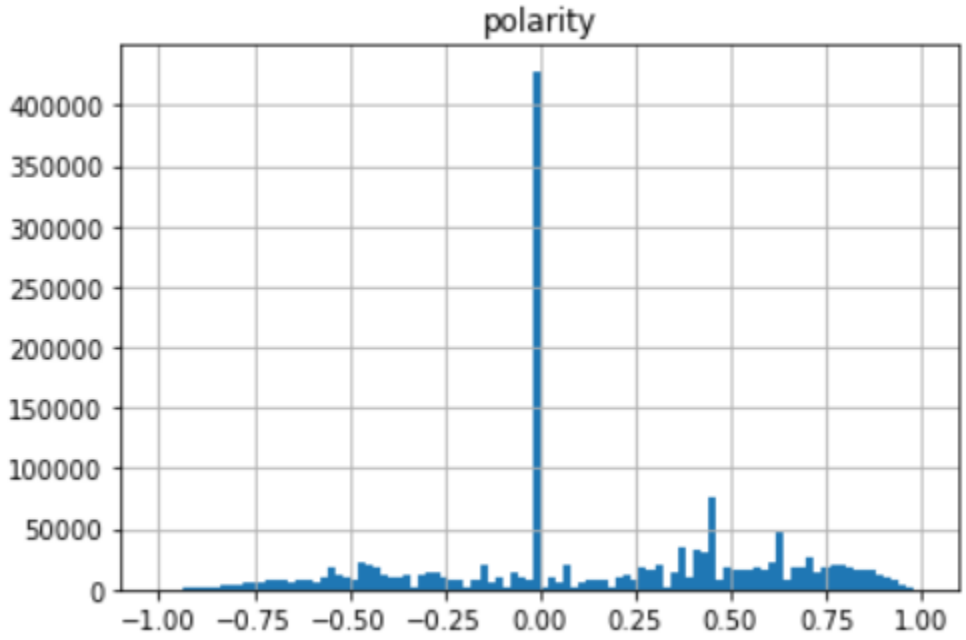
1. **¿Cuál es la distribución de las polaridades y complejidad de lectura/escritura de los tweets en el dataset?**

   En primer lugar se usa la librería nltk para recalcular el sentimiento de los tweets siguiendo los siguientes pasos:

   - Se preprocesa el texto eliminando espacios adicionales, transformando a minúsculas el texto…
   - Se calcula la polaridad de nuevo usando la polarity score con la clase SentimentIntensityAnalyzer.
   - Se calcula la facilidad de lectura.
   - Se calcula la facilidad de escritura.

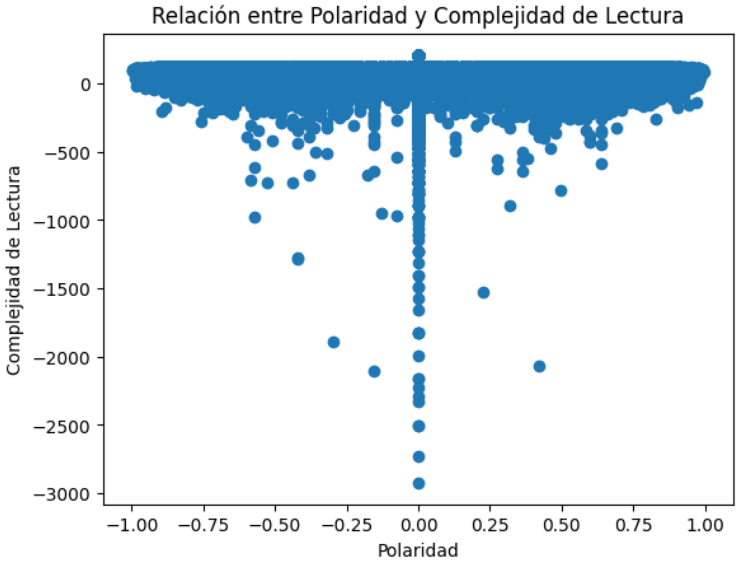
   a. **¿Hay una mayor cantidad de tweets positivos, negativos o neutrales?**

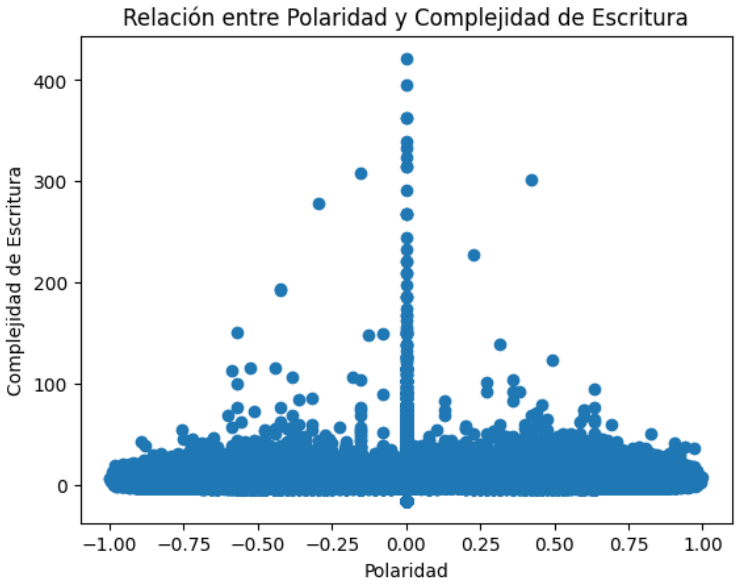
      Haciendo un histograma con la polaridad de los tweets, podemos apreciar que la mayoría de tweets se encuentran en la gamma neutral.



b. ¿Cómo se relacionan las distintas polaridades según la complejidad de lectura/escritura de los tweets?

Mediante dos scatters plots, relacionamos la polaridad con la complejidad de lectura y escritura.





Podemos apreciar que los tweets con neutralidad 0 son los más complejos de escribir, así como unos puntales en complejidad en los puntos medianos de los tweets con polaridad positiva y negativa, respectivamente.

Además, la distribución simétrica en ambos gráficos sugiere que hay una presencia equilibrada de tweets con polaridades positivas y negativas, y que no hay una polaridad dominante en los tweets del dataset.

2. **¿Existen patrones gramaticales o sintácticos comunes en los tweets con polaridad positiva o negativa? Por ejemplo, puede que los tweets positivos tiendan a utilizar más palabras de agradecimiento o elogios, mientras que los tweets negativos utilizan más palabras de crítica o enojo.**

   Para valorar si hay patrones gramaticales o sintácticos comunes según la polaridad de los twits, en primer lugar realizamos un preprocesado para eliminar símbolos de puntuación o stoppers así cómo otros caracteres que no sean palabras.
   A continuación, separamos los twits por positivos o negativos separando en función de si el valor de la polaridad es mayor o menor que 0 y utilizamos la función FreqDist de la librería nltk que nos permite obtener las palabras que se usan más frequentemente en los twits.
   Obtenemos las siguientes palabras cómo más frequentes en los twits positivos:

```
good - 81033
love - 60811
like - 59497
day - 51060
lol - 49467
get - 42140
quot - 39349
thanks - 38081
http - 36479
go - 35296
```

   Y las siguientes para twits negativos:

```
sad - 22566
miss - 21527
get - 21460
bad - 20841
go - 19778
got - 18673
like - 18314
today - 17776
day - 17732
im – 17590
```

   Analizando funcionalmente el contenido de las palabras más utilizados efectivamente podemos ver que los twits con polaridad positiva tienen palabras cómo ‘good’, ‘love’,’like’,’thanks’ con connotación positiva mientras que los twits con polaridad negativa tienen palabras cómo ‘sad’, ‘miss’, ‘bad’ con connatación claramente negativa.

3. **¿Qué usuarios tienden a generar tweets con una polaridad más positiva o negativa?**

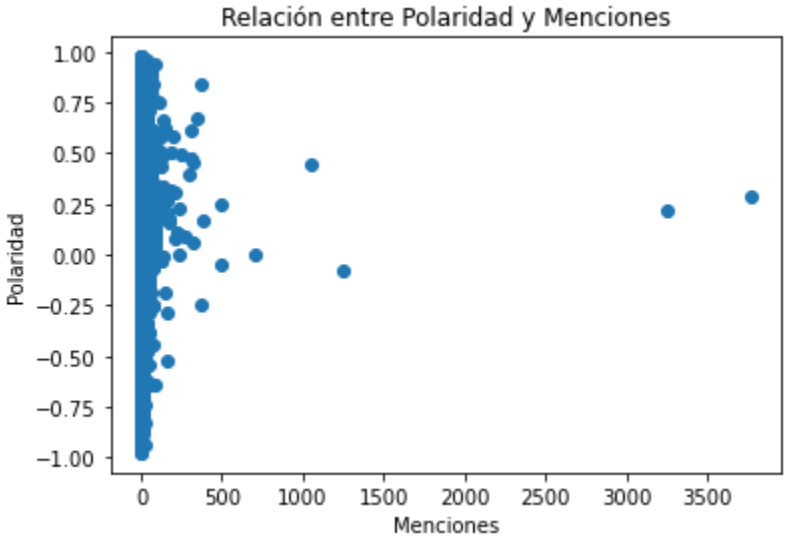
   Buscando los usuarios con los twits con polaridad más positiva y negativa podemos ver que son usuarios con pocos o un único twit.

```
Usuarios con la polaridad promedio más positiva:
       Usuario_tweeteado  lemmatized_text
439011         katekatew             1.0
571846        sammylee15             1.0
571844       sammylacsam             1.0
11128           AlyBabii             1.0
332675      darkmagician             1.0
444682    kelseyfarley95             1.0
```

```
505184   misssarahfrench              1.0
622645         timmytaz7              1.0
240602          ZenZoneMS             1.0
546836        poprincess              1.0

Usuarios con la polaridad promedio más negativa:
       Usuario_tweeteado  lemmatized_text
371767      furtherpeace             -1.0
455940      kumar_amrita             -1.0
102577       Jess_Sumner             -1.0
102590        Jess_ica89             -1.0
34765     CHARL00TTEEEEE             -1.0
419323         jessd1987             -1.0
476215        lozzylol85             -1.0
271470           aritah_             -1.0
455928            kultar             -1.0
455224            ksen22             -1.0
```

¿Hay alguna relación entre la polaridad de los tweets y el número de menciones a un usuario?

Se han recopilado el número de menciones que ha recibido cada usuario y se ha unido con la media de polaridad de los tweets de ese mismo usuario.



Podemos apreciar en la gráfica superior que la mayoría de menciones se puede relacionar con twits de polaridad más alta, es decir, cómo mejores són tus twits más menciones recibes.

4. **¿Hay alguna palabra o conjunto de palabras específicas que estén asociadas con tweets de polaridad extrema?**

Para valorar las palabras o conjunto de palabras asociadas a tweets con polaridad 1 o -1, se han usado el TF-IDF.
Y se han obtenido los siguientes resultados:

```
Palabras con polaridad extrema (positivas) basadas en TF-IDF:
perut: 13765.59349106324
```

```
jsinkeywest: 9850.303732467648
ouwh: 7127.903468067182
peanutparrot: 6704.402451482106
planetabroad: 6397.8742580208045
ronnocnalyd: 6074.220995104297
phony: 6047.532634871702
qucik: 5917.519562009537
peon: 5839.460602186941
nottingham_news: 5576.492101592066

Palabras con polaridad extrema (negativas) basadas en TF-IDF:
miss: 4791.306342569309
sad: 4715.940451676014
bad: 4215.721309570681
work: 3877.945010826876
hate: 3874.0747752575553
go: 3853.405821724651
day: 3822.2964629342805
really: 3809.3431715837387
get: 3771.4724791362664
today: 3752.440726092537
```

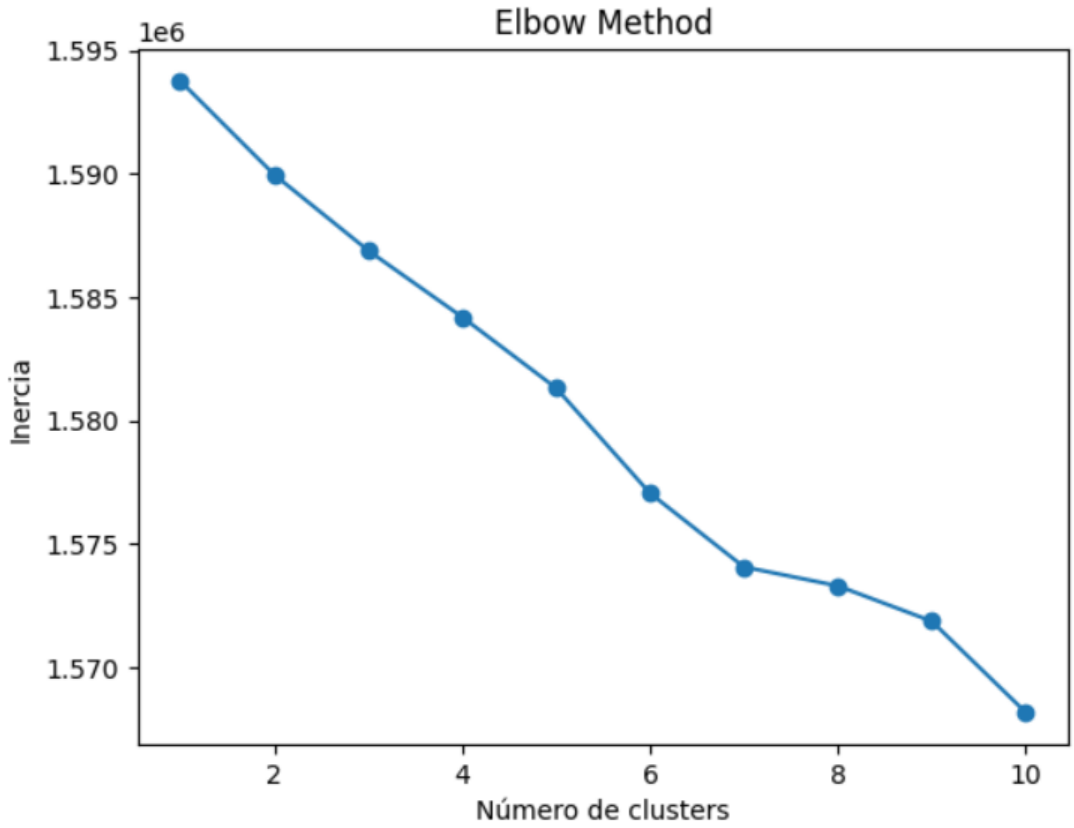
En los tweets positivos, algunas de las palabras con polaridad extrema identificadas son "perut", "jsinkeywest", "ouwh", "peanutparrot", entre otras. Estas palabras podrían ser términos específicos o jerga utilizada en contextos positivos.
En los tweets negativos, algunas de las palabras con polaridad extrema identificadas son "miss", "sad", "bad", "work", "hate", entre otras. Estas palabras suelen estar asociadas con sentimientos negativos o situaciones desfavorables.

a. ¿Estas palabras son más comunes en tweets sobre un tema en particular o están distribuidas en todo el dataset?

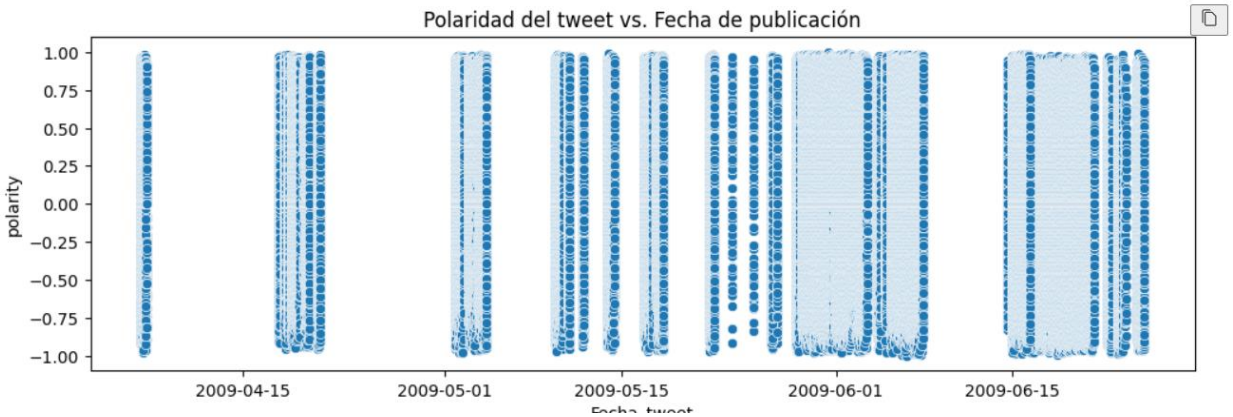
Las palabras están distribuidas en reducidos temas debido a la dificultad de encontrar twits con polaridades extremas.

b. Escoge un tema y clusteriza los usuarios según polaridades.

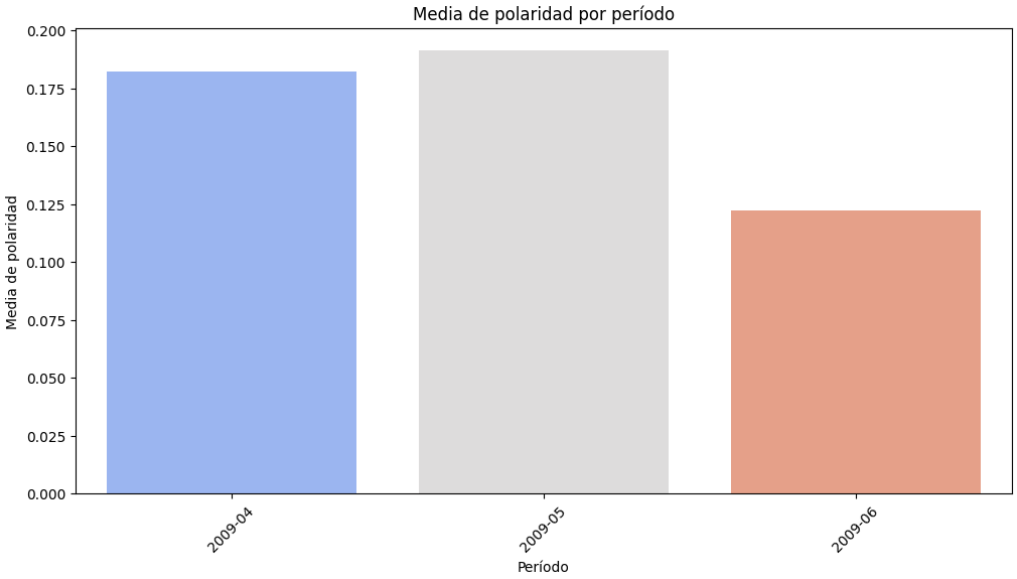
Se han elegido las palabras que contienen la palabra ‘war’ cómo tema. Se ha realizado un gráfico de elbow para intentar encontrar el número óptimo de clústers para clusterizar a los usuarios.
Debido a la complejidad del sistema però, el gráfico de elbow es muy lineal y no se ha podido discernir un número de clústers coherente para hacer la separación según polaridades con exactitud, aunque hemos tomado el 7 cómo valor òptimo.



5. **¿Hay alguna correlación entre la polaridad de un tweet y la fecha en que se publicó?**
   Se ha relacionado la polaridad del tweet con la fecha de la publicación pero no se ha encontrado correlación entre la polaridad y la fecha, ya que se distribuyen muy uniformemente en todos los períodos.



   a. **¿Los tweets publicados durante ciertos períodos de tiempo tienden a ser más positivos o negativos que otros?**
      Se han agrupado mensualmente los tweets y se ha computado la media de la polaridad:



Podemos apreciar cómo efectivamente, los meses de abril y mayo la polaridad de los twits es casi el doble que la media de polaridad del mes de junio.

6. **Identifica los Top 10 Trolls y Top 10 Influencers. Justifica las características de un usuario Troll e Influencer.**

   Para identificar los 10 trolls se han identificado los usuarios con menor polaridad en todos sus twits:

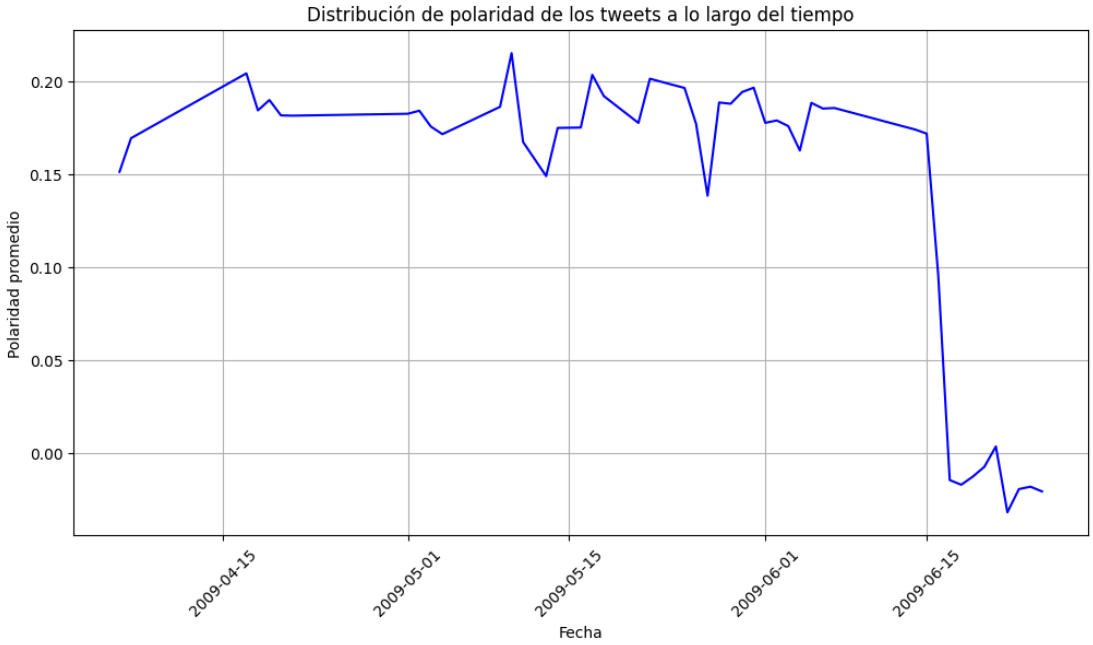
```
Top 10 Trolls:
liveloveperform
erynmariex3
beccalh
Mwissa
krissy25d
otta327
poke__egg
KFay91
mummy2751
HelenaMart
```

   Para identificar el Top 10 de Influencers se han buscado los usuarios con la polaridad más alta:

```
Top 10 Influencers:
mkshine09
jumana_engineer
rupydetequila
Lkaati
cj_mac
silver7
mando_betty
melandnessa
McAnita
```

VISUALIZACIÓN:

1. **¿Cómo se distribuyen los tweets según su polaridad a lo largo del tiempo?**



De manera similar a la visualización del ejercicio 5.a. podemos apreciar cómo durante los meses de abril y mayo, la polaridad de los tweets es generalmente positiva, mientras que a partir del dia 15 de junio la polaridad disminuye brutalmente, indicando probablemente que sucedió algún acontecimiento negativo que creo malestar en todos los tweets.

2. **Visualiza el análisis sintáctico (número de palabras, frase, verbos, nombres…) de los top 10 Trolls e Influencers.**

Se utiliza el modelo de lenguaje de Spacy para realizar el análisis sintáctico de los tweets de los top 10 Trolls e Influencers. Cruzando con los resultados obtenidos en ejercicios anteriores:

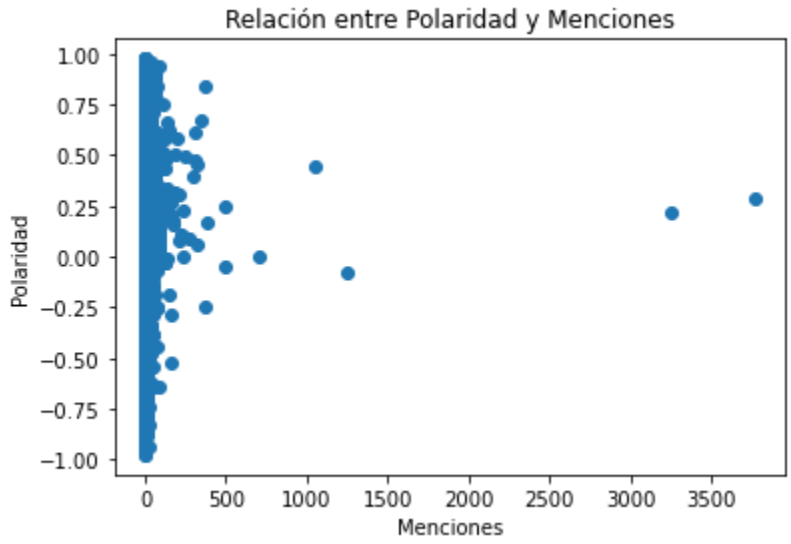
```
Análisis sintáctico de los top 10 Trolls:
Troll 1:
Número de palabras: 26
Número de frases: 3
Número de verbos: 7
Número de nombres: 2
Número de adjetivos: 2
Número de adverbios: 1

Troll 2:
Número de palabras: 29
Número de frases: 4
Número de verbos: 5
Número de nombres: 5
Número de adjetivos: 4
Número de adverbios: 1

Troll 3:
Número de palabras: 15
Número de frases: 3
Número de verbos: 2
Número de nombres: 1
Número de adjetivos: 0
Número de adverbios: 0
```

```
Troll 4:
Número de palabras: 30
Número de frases: 2
Número de verbos: 4
Número de nombres: 5
Número de adjetivos: 2
Número de adverbios: 3

Troll 5:
Número de palabras: 31
Número de frases: 10
Número de verbos: 10
Número de nombres: 10
Número de adjetivos: 0
Número de adverbios: 0

Troll 6:
Número de palabras: 21
Número de frases: 1
Número de verbos: 3
Número de nombres: 8
Número de adjetivos: 2
Número de adverbios: 3

Troll 7:
Número de palabras: 33
Número de frases: 1
Número de verbos: 5
Número de nombres: 5
Número de adjetivos: 3
Número de adverbios: 1

Troll 8:
Número de palabras: 20
Número de frases: 3
Número de verbos: 1
Número de nombres: 3
Número de adjetivos: 5
Número de adverbios: 0

Troll 9:
Número de palabras: 26
Número de frases: 3
Número de verbos: 5
Número de nombres: 3
Número de adjetivos: 0
Número de adverbios: 1

Troll 10:
Número de palabras: 26
Número de frases: 1
Número de verbos: 4
Número de nombres: 4
Número de adjetivos: 0
Número de adverbios: 0
```

```
Análisis sintáctico de los top 10 Influencers:
Influencer 1:
Número de palabras: 18
Número de frases: 1
Número de verbos: 4
Número de nombres: 3
Número de adjetivos: 0
Número de adverbios: 0

Influencer 2:
Número de palabras: 31
Número de frases: 3
Número de verbos: 7
Número de nombres: 4
Número de adjetivos: 0
Número de adverbios: 0

Influencer 3:
Número de palabras: 27
Número de frases: 2
Número de verbos: 0
Número de nombres: 2
Número de adjetivos: 3
Número de adverbios: 0

Influencer 4:
Número de palabras: 23
Número de frases: 1
Número de verbos: 4
Número de nombres: 4
Número de adjetivos: 2
Número de adverbios: 2

Influencer 5:
Número de palabras: 19
Número de frases: 2
Número de verbos: 3
Número de nombres: 8
Número de adjetivos: 0
Número de adverbios: 1

Influencer 6:
Número de palabras: 28
Número de frases: 4
Número de verbos: 5
Número de nombres: 3
Número de adjetivos: 4
Número de adverbios: 2

Influencer 7:
Número de palabras: 28
Número de frases: 5
Número de verbos: 6
Número de nombres: 3
Número de adjetivos: 4
Número de adverbios: 1
```

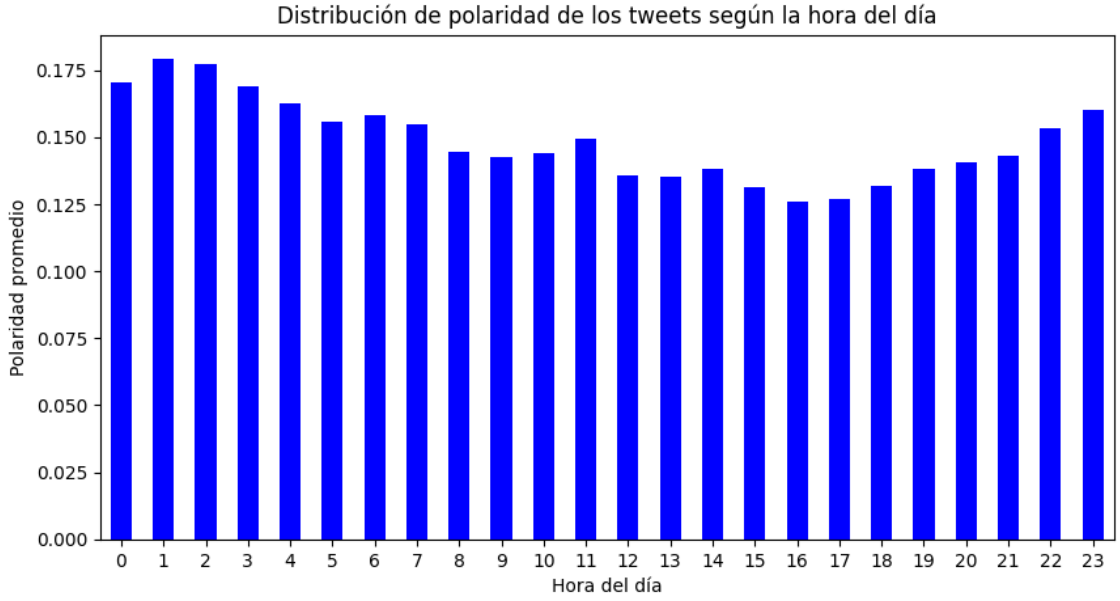
```
Influencer 8:
Número de palabras: 33
Número de frases: 4
Número de verbos: 4
Número de nombres: 3
Número de adjetivos: 3
Número de adverbios: 3

Influencer 9:
Número de palabras: 28
Número de frases: 2
Número de verbos: 1
Número de nombres: 8
Número de adjetivos: 2
Número de adverbios: 1

Influencer 10:
Número de palabras: 26
Número de frases: 5
Número de verbos: 4
Número de nombres: 1
Número de adjetivos: 3
Número de adverbios: 2
```

3. **¿Existe alguna correlación entre el número de seguidores de un usuario y la polaridad de sus tweets? Representa visualmente esta relación.**

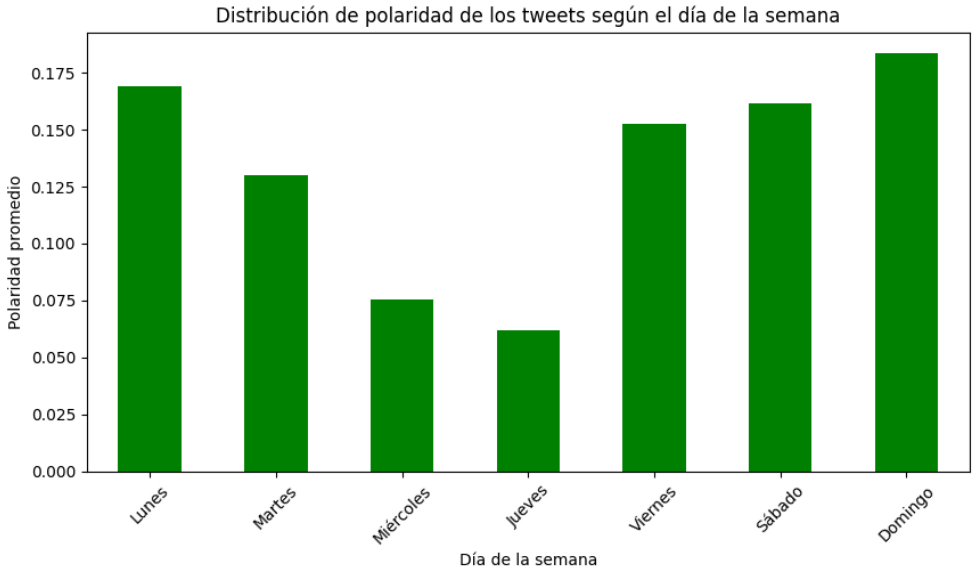


4. **Crea una nube de palabras para cada polaridad.**

5. ¿Cómo se distribuyen los tweets según su polaridad en función de la hora del día o el día de la semana?



MBD – Caso 3


Distribución de polaridad de los tweets según el día de la semana
Polaridad promedio
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Lunes
Martes
Miércoles
Jueves
Viernes
Sábado
Domingo
Día de la semana