Trabajo Final

Curso R aplicado al análisis cualitativo (EP/FCS)

Sofia Pandolfo

17 de febrero de 2023

Fuente de datos

Para este trabajo se utilizaron los artículos misceláneos de la Revista Uruguaya de Ciencia Política (RUCP) en los últimos 13 años (2009-2022). El objetivo es conocer cuáles son los temas predominantes en la investigación del Instituto de Ciencia Política de la Facultad de Ciencias Sociales, Universidad de la República. Se han seleccionado los números misceláneos para no sesgar la muestra con la selección intencional de temáticas.

Los artículos se extrajeron vía web scrapping utilizando los paquetes *rvest*, *stringr*, *purrr* y *pdf_tools* como se muestra a continuación.

```
urls <- list(
  url_RUCP_312 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/44"),
  url_RUCP_302 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/42"),
  url_RUCP_392 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/40"),
  url_RUCP_282 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/38"),
  url_RUCP_272 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/36"),
  url_RUCP_262 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/4"),
  url_RUCP_261 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/5"),
  url_RUCP_252 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/7"),
  url_RUCP_241 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/10"),
  url_RUCP_232 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/11"),
  url_RUCP_232 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/29"),
  url_RUCP_212 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/15"),
  url_RUCP_201 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/16"),
  url_RUCP_191 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/17"),
  url_RUCP_181 = rvest::read_html("http://rucp.cienciassociales.edu.uy/index.php/rucp/issue/view/18"))

pdfs <- list()
for(i in 1:length(urls)){
  pdfs [[i]] <- urls [[i]] %>%
    html_elements(".pdf")%>%
    html_attr("href")%>%
    stringr::str_replace_all(pattern = "view",replacement = "download")%>%
    purrr::map(pdftools::pdf_text)
}
```

Pre-procesamiento de texto

La extracción de datos generó una lista con dos niveles: en primer lugar, un elemento de la lista por cada volúmen extraído. Este elemento era a su vez una lista, con un elemento por artículo. Cada artículo era un vector de caracteres en el que cada elemento constituía una página del artículo.

El primer paso del pre-procesamiento fue re-estructurar la información para poder trabajar con ella. Se eliminó el primer nivel de la lista, el volúmen, ya que el propósito de este análisis era conocer los temas generales tratados en la RUCP. A continuación se unieron las páginas para tener el artículo completo en un mismo elemento. Luego se utilizó la función *unlist* para romper la estructura de la lista y obtener un vector de caracteres con un elemento por artículo. En este paso se utilizó la función *gsub* del paquete base para eliminar la expresión "Revista Uruguaya de Ciencia Política" ya que, al ser el nombre de la revista a analizar, esta expresión se repetiría sesgando los análisis de frecuencia de palabras. Finalmente se asignó un ID a cada artículo convirtiendo el vector en un dataframe.

```
pdfs <- purrr::flatten(pdfs)
pdfs2 <- list()
for (i in 1:length(pdfs)){
  pdfs2 [[i]] <- paste(pdfs[i], collapse = " ")
}

pdfs2 <- unlist(pdfs2)
pdfs2 <- gsub("revista uruguaya de ciencia política", "", pdfs2, ignore.case = TRUE)
pdfs2 <- data.frame(ID = 1:length(pdfs2), articulos = pdfs2)
```

Análisis

Frecuencia de ocurrencia

Para conocer los temas predominantes en la RUCP se utilizó en primer lugar un análisis de frecuencia de ocurrencia con el paquete *quanteda*. En primer lugar el código se corrió utilizando la función *dfm_remove* para eliminar las stopwords comunes en los tres idiomas en los que se puede publicar en la RUCP: español, inglés y portugués. También se eliminaron las palabras de menos de tres caractéres. Se agruparon los términos por artículo y se mantuvieron los términos que tuvieran al menos 6 menciones.

```
dfm_pdfs <- quanteda::dfm(quanteda::tokens(pdfs2$articulos,
                                           remove_punct = TRUE,
                                           remove_numbers = TRUE),
                          tolower=TRUE,
                          verbose = FALSE) %>%
  quanteda::dfm_remove(pattern = c(quanteda::stopwords("es"),
                                   quanteda::stopwords("en"),
                                   quanteda::stopwords("pt")),
                       min_nchar=3)%>%
  quanteda::dfm_trim(min_termfreq = 6)%>%
  quanteda::dfm_group(groups = pdfs2$ID)
```

Para comenzar el análisis se extrajeron los 50 términos más frecuentes.

```
top50 <- as.data.frame(topfeatures(dfm_pdfs,50))
```

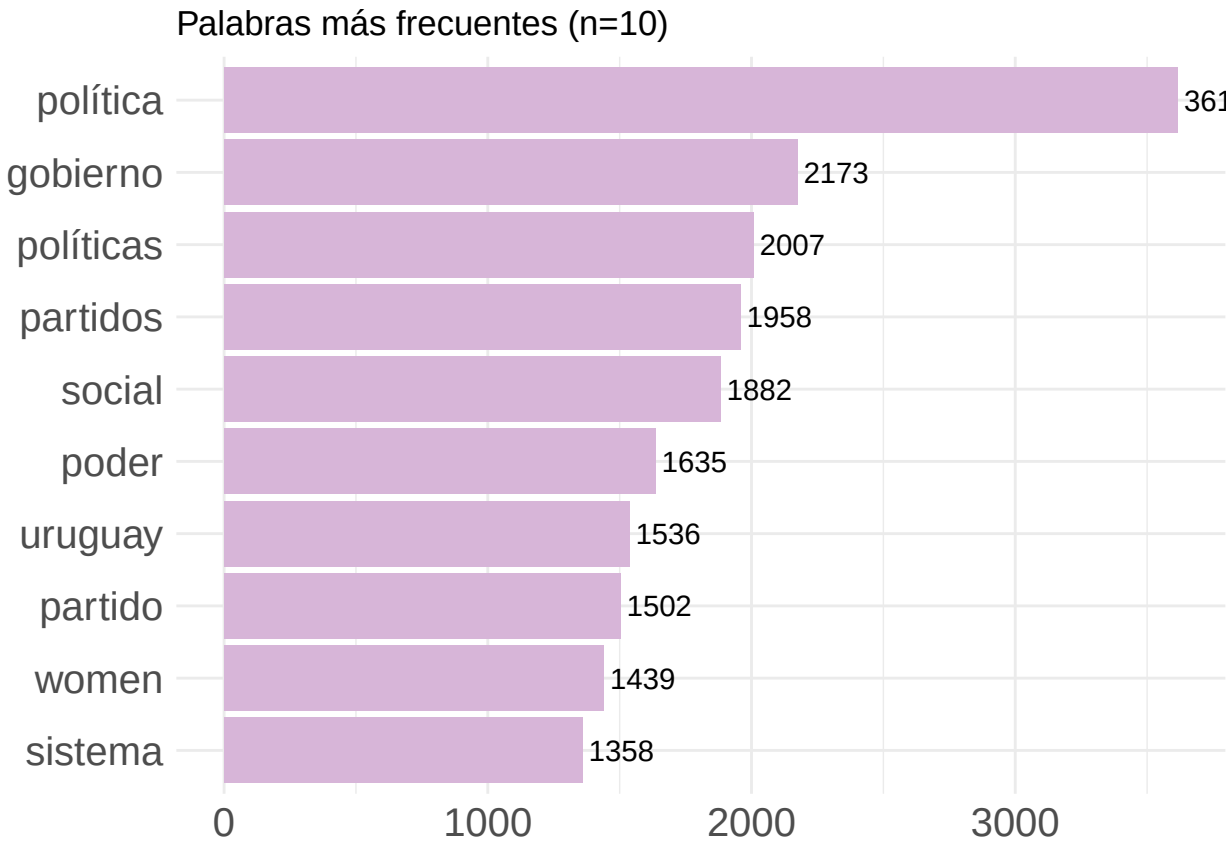
Esta lista arrojó un conjunto de términos que eran en realidad typos, por lo que se volvió a construir el dfm incorporando estos términos en el argumento *pattern* de la función *dfm_remove*.

```
dfm_pdfs <- quanteda::dfm(quanteda::tokens(pdfs2$articulos,
                                           remove_punct = TRUE,
                                           remove_numbers = TRUE),
                          tolower=TRUE,
                          verbose = FALSE) %>%
  quanteda::dfm_remove(pattern = c(quanteda::stopwords("es"),
                                   quanteda::stopwords("en"),
                                   quanteda::stopwords("pt"),
                                   "nde",
                                   "nla",
                                   "nen",
                                   "nque",
                                   "nel",
                                   "nlos",
                                   "nlas",
                                   "ndel",
                                   "ser",
                                   "caso",
                                   "parte",
                                   "puede",
                                   "análisis",
                                   "cada",
                                   "forma",
                                   "nivel",
                                   "mayor",
                                   "años",
                                   "bien",
                                   "así",
                                   "lugar",
                                   "atria",
                                   "ncomo",
                                   "perspec-",
                                   "inés",
                                   "nsocial",
                                   "ndesigualdades"),
                       min_nchar=3)%>%
  quanteda::dfm_trim(min_termfreq = 6)%>%
  quanteda::dfm_group(groups = pdfs2$ID)
```

A continuación se presentan las 10 palabras más utilizadas:

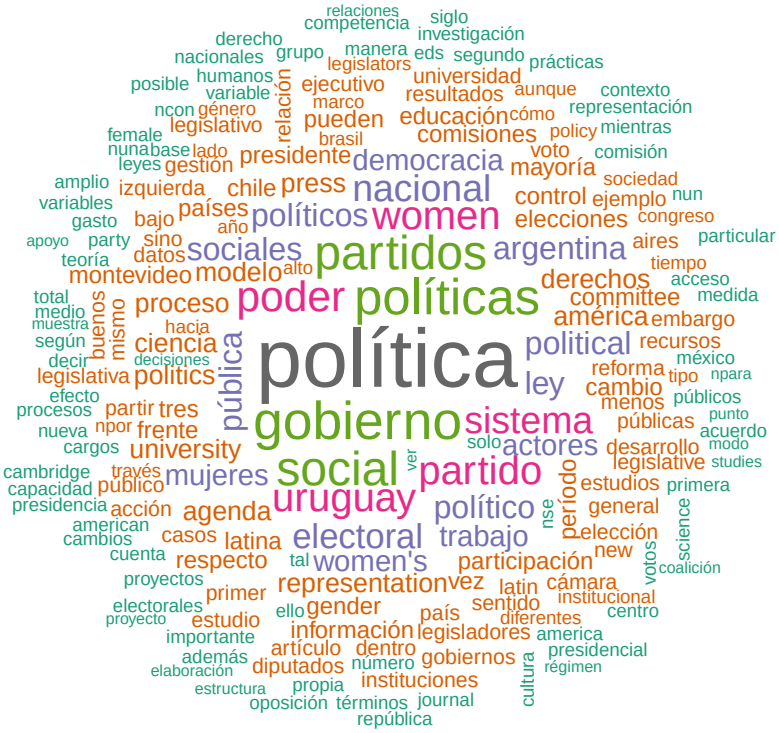
```
top <- data.frame(topfeatures(dfm_pdfs,10))
top$palabra = rownames(top)

#hago el gráfico con ggplot
top <- data.frame(topfeatures(dfm_pdfs,10))
top$palabra = rownames(top)

top %>%
  ggplot(aes(x = reorder(palabra, topfeatures.dfm_pdfs..10.),
             y = topfeatures.dfm_pdfs..10., fill = palabra)) +
```

```
  geom_col(show.legend = FALSE) +
  coord_flip() +
  geom_text(aes(hjust = -0.1, label = topfeatures.dfm_pdfs..10.)) +
  theme_minimal() +
  theme(axis.title.y = element_blank(), axis.title.x = element_blank(), axis.text = element_text(size =
  ggtitle("Palabras más frecuentes (n=10)") +
  scale_fill_manual(values = c(rep("#D7B5D8",10)))
```



A partir de la dfm se realizó una nube de palabras, utilizando también el paquete *quanteda*.

```
quanteda.textplots::textplot_wordcloud(dfm_pdfs,
                                       min.count = 2,
                                       max_words = 200,
                                       random.order = FALSE,
                                       colors = RColorBrewer::brewer.pal(8,"Dark2"),
                                       comparison = F)
```



Estas visualizaciones muestran el lugar predominante de la Ciencia Política orientada a los estudios partidarios, ya que al sumar "partidos" y "partido" la cuestión partidaria se ubica en un indiscutido primer lugar con 3593 menciones.

En tercer lugar se encuentra el término "políticas" que parece indicar la preponderancia de los estudios de políticas públicas. De todas formas, para conocer más sobre el uso del término se realizó un análisis de correlación también con el paquete *quanteda*.

```
quanteda.textstats::textstat_simil(dfm_pdfs,
                                   selection = "políticas",
                                   method = "correlation",
                                   margin = "features")%>%
  as.data.frame()%>%
  dplyr::arrange(-correlation)%>%
  dplyr::top_n(15)
```

```
##         feature1  feature2 correlation
## 1       públicas políticas   0.7123080
## 2     npolíticas políticas   0.6538495
## 3          marco políticas   0.6066046
## 4       sociales políticas   0.6041519
## 5      problemas políticas   0.6005175
## 6            nun políticas   0.5908722
## 7        trabajo políticas   0.5876083
## 8           peso políticas   0.5859601
## 9    implementan políticas   0.5812550
```

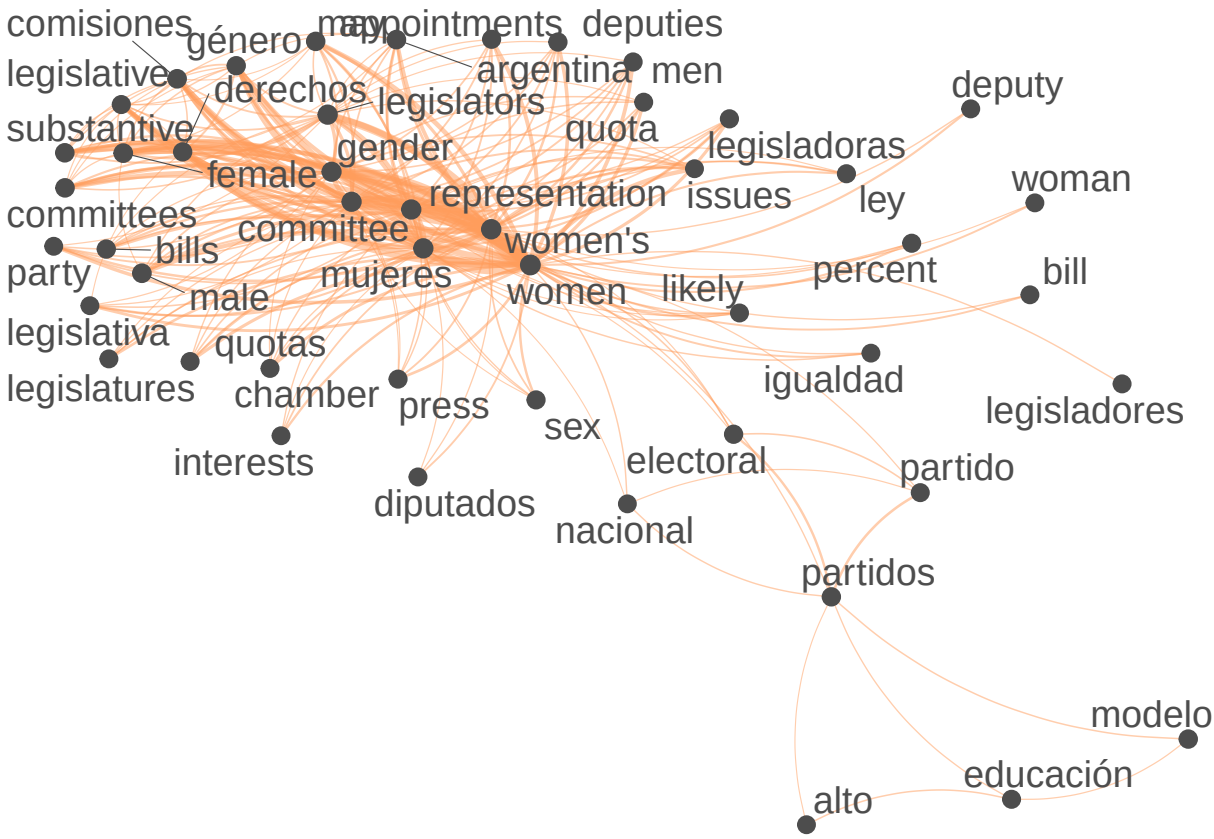
```
## 10        amplia políticas   0.5718055
## 11      política políticas   0.5684024
## 12       consejo políticas   0.5679593
## 13   identificar políticas   0.5671763
## 14     framework políticas   0.5630407
## 15 institucional políticas   0.5557463
```

Efectivamente, el término políticas estaba vinculado a los estudios de políticas públicas, siendo éste último el principal término relacionado a "políticas". A su vez, destacan términos como "sociales", "problemas" e "implementan".

Correlación de temas

Finalmente se realizó un mapa de correlación de temas.

```
base_fcm <- dfm_pdfs%>%
  fcm(context = "document")

feat <- names(topfeatures(base_fcm, 50))
base_fcm_select <- fcm_select(base_fcm,
                              pattern = feat,
                              selection = "keep")
size <- log(colSums(dfm_select(base_fcm,
                               feat,
                               selection = "keep")))

set.seed(144)
quanteda.textplots::textplot_network(base_fcm_select,
                                     min_freq = 0.8,
                                     vertex_size = size / max(size) * 3,
                                     edge_color="#ff9d5c")
```



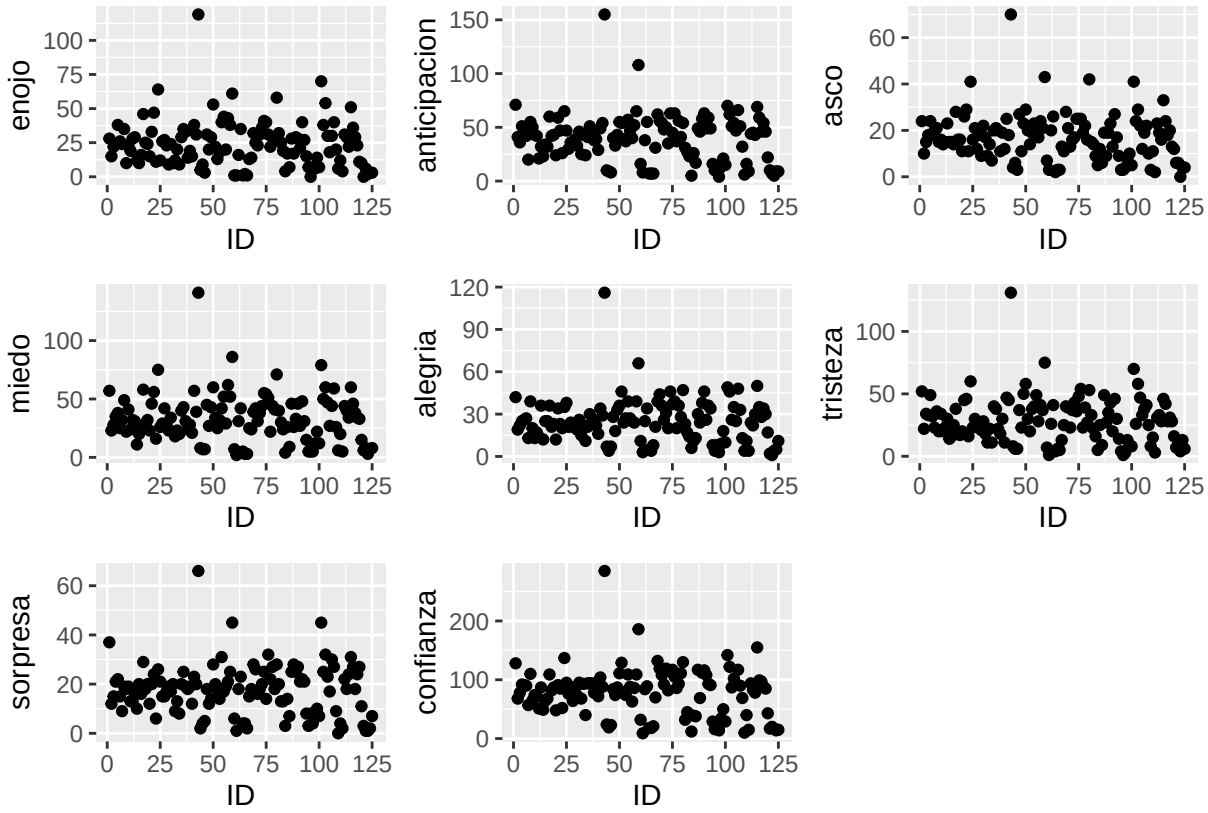
Este mapa muestra una preponderancia de términos en inglés y relativos a cuestiones de género que no surgió con los análisis anteriores.

Análisis de sentimiento

Finalmente se realizó un análisis de sentimiento bajo la hipótesis de que, al tratarse de una revista académica, los artículos no demostrarían gran precencia de sentimientos. Se utilizó el paquete *syuzhet*. Este paquete reconoce en el texto los siguientes sentimientos: enojo, anticipación, asco, miedo, alegría, tristeza, sorpresa y confianza.

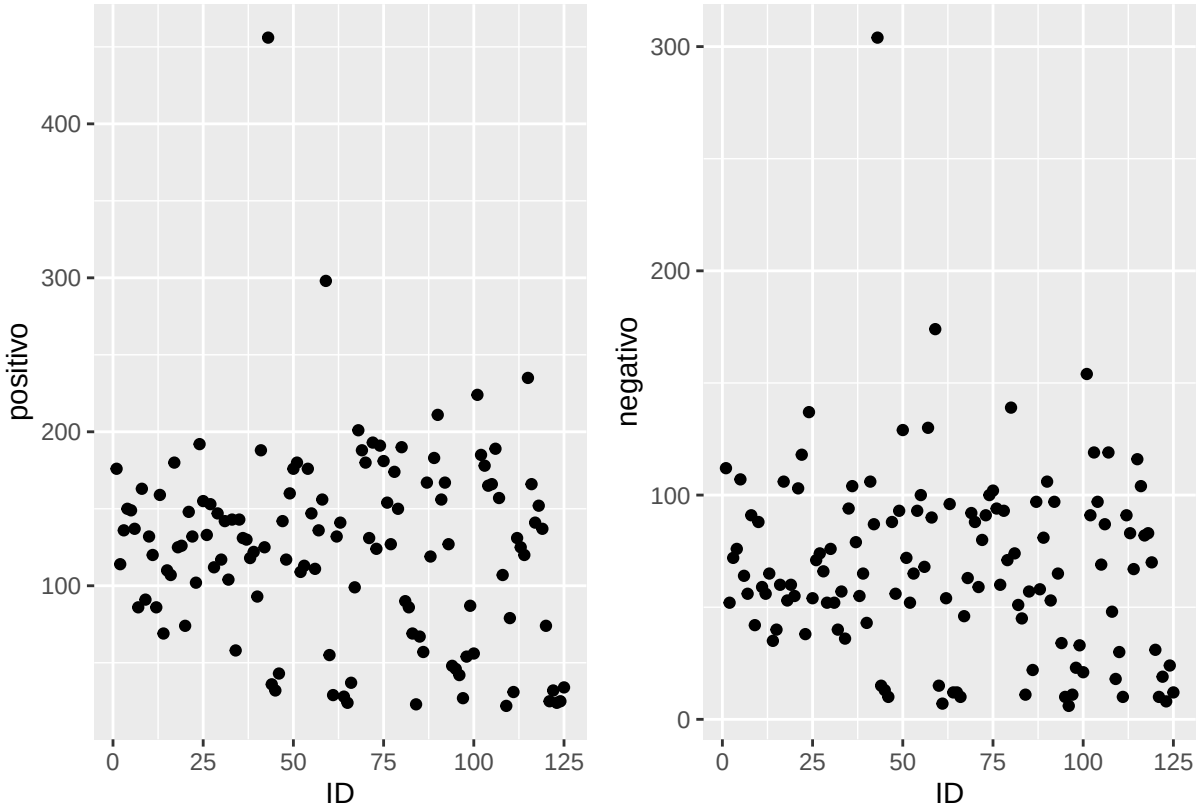
```
sentiment <- get_nrc_sentiment(pdfs2$articulos, language = "spanish")

df_sentiment <- cbind(pdfs2[,1], sentiment)
names(df_sentiment) <- c("ID",
                         "enojo",
                         "anticipacion",
                         "asco",
                         "miedo",
                         "alegria",
                         "tristeza",
                         "sorpresa",
                         "confianza",
                         "negativo",
                         "positivo")
#mantengo solo el ID del artículo
```

```
library(patchwork)
graph1 <- ggplot(df_sentiment, aes (ID, enojo)) + geom_point()
graph2 <- ggplot(df_sentiment, aes (ID, anticipacion)) + geom_point()
graph3 <- ggplot(df_sentiment, aes (ID, asco)) + geom_point()
graph4 <- ggplot(df_sentiment, aes (ID, miedo)) + geom_point()
graph5 <- ggplot(df_sentiment, aes (ID, alegria)) + geom_point()
graph6 <- ggplot(df_sentiment, aes (ID, tristeza)) + geom_point()
graph7 <- ggplot(df_sentiment, aes (ID, sorpresa)) + geom_point()
graph8 <- ggplot(df_sentiment, aes (ID, confianza)) + geom_point()
graph1 + graph2 + graph3 + graph4 + graph5 + graph6 + graph7 + graph8
```



También reconoce términos positivos y negativos. Se mantiene la misma hipótesis.

```
positivo <- ggplot(df_sentiment, aes (ID, positivo)) + geom_point()
negativo <- ggplot(df_sentiment, aes (ID, negativo)) + geom_point()
positivo + negativo
```



Efectivamente, los artículos académicos muestran niveles constantes y relativamente bajos de presencia de términos que demuestren sentimientos y opiniones normativas, con la presencia de algunos valores atípicos.