Jul 12, 2019

Text Mining Approach for Adapting a School-Based Sexual Health Promotion Program in Colombia

Pablo Vallejo-Medina[1], Juan C. Correa[1], Mayra Gómez-Lugo[1], Diego Alejandro Saavedra[1], Eileen García-Montaño[2], Diana Pérez-Pedraza[2], Janivys Niebles-Charris[2], Paola García-Roncallo[2], Daniella Abello-Luque[2], José Pedro Espada[3], Alexandra Morales[3]

[1]Fundación Universitaria Konrad Lorenz, Faculty of Psychology, [2]Universidad de la Costa, Social Sciences Department, [3]Universidad Miguel Hernandez de Elche, Health Psychology Department

1 Works for me dx.doi.org/10.17504/protocols.io.5ddg226

Juan C Correa
Fundación Universitaria Konrad Lorenz

ABSTRACT

A common practice among clinical psychologists and other health professionals is the use of school-based sexual health promotion programs as a means for preventing sexually transmitted infections. A fundamental criterion for the designing and adaptation of these programs is the age of their target populations because limited education and language are the most relevant factors that limit the efficacy of these programs. In this paper, we proposed a methodological approach that facilitates the empirical evaluation of the written materials that accompany the school-based sexual health promotion programs, taken as a case a Spanish-written program used in Colombia. The results showed the empirical adequacy of this program while positing novel insights for scrutinizing the efficacy of similar programs.

GUIDELINES

On July 10, 2019, we performed our analysis in the following machine:

R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: elementary OS 0.4.1 Loki

Matrix products: default
BLAS: /usr/lib/libblas/libblas.so.3.6.0
LAPACK: /usr/lib/lapack/liblapack.so.3.6.0

locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=es_CO.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=es_CO.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=es_CO.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=es_CO.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats graphics grDevices utils datasets methods
[7] base

other attached packages:
[1] koRpus_0.11-5 sylly_0.1-5

loaded via a namespace (and not attached):
[1] Rcpp_1.0.1 assertthat_0.2.1 crayon_1.3.4
[4] dplyr_0.8.3 grid_3.6.0 R6_2.4.0
[7] gtable_0.3.0 magrittr_1.5 scales_1.0.0
[10] ggplot2_3.2.0 pillar_1.4.2 rlang_0.4.0
[13] lazyeval_0.2.2 data.table_1.12.2 rstudioapi_0.10
[16] tools_3.6.0 glue_1.3.1 purrr_0.3.2
[19] munsell_0.5.0 compiler_3.6.0 pkgconfig_2.0.2

```
[22] colorspace_1.4-1  tidyselect_0.2.5  tibble_2.1.3
```

MATERIALS TEXT

Our analyses rely on the written material for each of the five sessions of the COMPAS program.

BEFORE STARTING

Make sure you have downloaded and installed the following softwares:
1) R from https://cran.r-project.org/
2) Rstudio from https://www.rstudio.com/products/rstudio/download/
3) koRpus by running install.packages("koRpus") in the console
4) koRpus.lang.es by running install.koRpus.lang("es") in the console
5) KH Coder from https://khcoder.net/en/

1 Download the following five txt-formatted files. **Sesion1.txt** **Sesion2.txt** **Sesion3.txt** **Sesion4.txt** **Sesion5.txt** and put these files inside a folder with the name "COMPAS".

2 If you are using a Linux distro, you can put that folder in the following path: "/home/juan/Documents/COMPAS". In Windows, you can put it in the following path: "C:\Users\juan\Documents\COMPAS\"

3 Download the following R-script and run it accordingly. **CompasProgramAnalysis.R**
Note, that in line 10 of this script you will see this code:

```
folder <- setwd("/home/juan/Documents/COMPAS")
```

Make sure to replace the path where you downloaded the folder COMPAS and edit the above code accordingly.

4 In line 11 you will see this code:

```
sesiones <- dir(folder)
```

After running that code, check that the following output is true

```
sesiones
[1] "Sesion1.txt" "Sesion2.txt" "Sesion3.txt" "Sesion4.txt" "Sesion5.txt"
```

Otherwise, run line 11 again

5 Run the following code

```
library(koRpus)
ID <- paste("session",1:length(sesiones), sep = "")
results <- vector("list", length(sesiones))
tok <- vector("list", length(sesiones))
Escolaridad <- vector("double", length(sesiones))
```

```
Edad <- vector("double", length(sesiones))
library(koRpus.lang.es)
for(i in seq_along(sesiones)){
  results[[i]] <- tokenize(sesiones[[i]], lang="es")
  tok[[i]] <- SMOG(results[[i]])
  Escolaridad[[i]] <- tok[[i]]@SMOG[[2]]
  Edad[[i]] <- tok[[i]]@SMOG[[3]]
}
DifText <- data.frame(ID, sesiones, Escolaridad, Edad)
hist(DifText$Escolaridad, xlab = "Años de Escolaridad", col = "lightblue")
library(ggplot2)
ggplot(DifText, aes(DifText$ID, DifText$Escolaridad)) + geom_bar(stat = "identity", fill="red")
+ xlab("Session Number") + ylab("Years of Education")

df2 <- data.frame(Variable=rep(c("Scholarity", "Age"), each=5),
                  Session=rep(c("Session1", "Session2", "Session3","Session4","Session5"),2),
                  Years=c(13.12, 12.93, 12.67, 11.62, 14.46, 18.12, 17.93,17.67, 16.62, 19.46))

ggplot(data=df2, aes(x=Session, y=Years, fill=Variable)) +
  geom_bar(stat="identity", position=position_dodge()) +
  geom_text(aes(label=Years), vjust=1.6, color="white",position = position_dodge(0.9), size=3.5)
```

6 A second part of the analyses entails the visualization of the network of words co-occurrence In KH Coder. For this process, we removed Spanish stopwords (e.g., articles, prepositions, etc.) as well as frequent verbs such as: “can,” “to be,” “to have,” and non-informative words such as “case,” “question” and “answer.”. Secondly, we removed regular expressions (i.e., hola, bienvenido, etc.) and ordinal adjectives (firstly, second, etc.) from the corpus analysis (i.e., a total of 8 new words irrelevant for the program). We set the sentences as the unit of analysis and employed a minimum term frequency of 10 as a threshold to depict words co-occurrence, along with a Jaccard similarity index as the filter for connexions between words. We colored our resulting networks following a betweenness centrality criterion, which stands for a metric that quantifies the statistical importance of a word, represented as a node within a network that shows the connexion between words. In addition, we used the following coding rules for merging semantic similar words. **coding rule.txt**