Aplicación del análisis datos textuales a un estudio en Salud Mental

Silvina Rodríguez[1]; Daniel Alessandrini[1]; Ramón Álvarez[2]; Laura Viola[3]

1. Licenciatura en Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República. **2.** Profesor adjunto, Unidad de Biometría, IESTA, Facultad de Ciencias Económicas y de Administración, UDELAR. **3.** Cátedra de Siquiatría Pediátrica, Facultad de Medicina, UDELAR

Palabras clave: Análisis Estadístico de Textos, Análisis de preguntas abiertas, Postcodificación, Salud Mental. Correo electrónico de contacto: **analisistextual@gmail.com**

Introducción:

La **lingüística** y la **estadística**: dos ciencias que cruzaron sus lazos hacia mediados del siglo XX, proporcionaron a la ciencia importantes ramas de investigación, tales como la búsqueda de información, el análisis del discurso, estimación del volumen de vocabulario, entre otros.

Objetivos:

- **introducir** en modo general el **Análisis Estadístico de Datos Textuales**
- mostrar una aplicación particular; **comparar la técnica de post-codificación** para el análisis de respuestas a preguntas abiertas, **con la información obtenida** a través de técnicas estadísticas descriptivas tales como el **análisis de correspondencias sobre una tabla léxica.**

Datos utilizados:

Estudio que se centra en obtener información epidemiológica respecto a la salud mental infantil en nuestro país. Realizado por un comité interdisciplinario de investigación, formado por miembros de la Clínica de Psiquiatría Pediátrica (CPP), el Depto. de Métodos Cuantitativos (Fac. Medicina, UDELAR); el Instituto de Estadística de la Fac. de Ciencias Económicas y una extensa colaboración de equipos de investigación extranjeros.

Breve reseña del trabajo realizado por la CPP sobre la Salud Mental de los niños uruguayos

El **instrumento de screening** utilizado fue el Child Behavior Checklist (**CBCL**), (Thomas Achenbach y Leslie Rescorla). Formulario autoadministrado a los padres, que evalúan aspectos relacionados al comportamiento y las relaciones sociales de sus hijos, valorando los últimos seis meses previos a la entrevista.

En la Introducción de la versión escrita de la primera presentación pública de este trabajo, se sintetiza que **"la Salud Mental es un componente básico de la salud integral del individuo y la sociedad".** El enfoque del trabajo citado se dirige a establecer políticas en salud mental infantil.

Objetivo primordial: medir la utilidad y aplicabilidad en nuestro país de un instrumento de screening, y así, obtener información epidemiológica en salud mental y dejar las puertas abiertas para la formulación de nuevas hipótesis de trabajo en el tema.

Conclusiones: el estudio permite mostrar una alta correlación entre las dificultades en el aprendizaje y la presencia de problemas emocionales y de comportamiento en los escolares de todo el país, y dar cuenta al mismo tiempo de la escasa derivación de los casos patológicos en centros especializados.

Resultados:

ACM Preguntas Abiertas Post-Codificadas

Postcodificación de las preguntas abiertas del cuestionario **CBCL**.

Pregunta A: *¿Qué es lo que más le preocupa acerca de su hijo/a?* se obtuvieron 15 códigos.

1 - Nada. **2** - Futuro. **3** - Conducta. **4** - Salud. **5** - Violencia. **6** - Educación. **7** - Inquietud. **8** - Carácter. **9** - Inmadurez. **10** - Timidez. **11** - Distracción. **12** - Sentimientos. **13** - Responsabilidad. **14** - Otros. **15** - No contesta.

Al realizar ACM, con las variables *Grado* escolar*, Quién contestó el formulario,* si el/la niño/a *Repitió* algún año*,* si el/la niño/a es *Saludable, Síndrome Global* y la variable cruzada *Sexo-Edad* como activas; y utilizando las preguntas post-codificadas *A* y *B*, *Trabajo del padre, Trabajo de la madre, Instrucción del padre e Instrucción de la madre* como suplementarias; las variables *Sexo-Edad* y *Grado* son las que más contribuyen a la formación del primer eje factorial: la contribución absoluta es casi el 80% en ese eje factorial y poco más del 61% en el segundo.

La inercia explicada por los cinco primeros ejes factoriales es de 36 %. Se usó como ponderadores los índices de Benzécri y Greenacre, obteniendo porcentajes un tanto mayores 69% y 40 %, respectivamente

Considerando el plano formado por los ejes factoriales 1 y 3, se puede decir que el eje 1 muestra, siguiéndolo de derecha a izquierda, el desarrollo en edad y grado del niño/a; y el eje 3 hace de divisor entre niños sanos y enfermos.

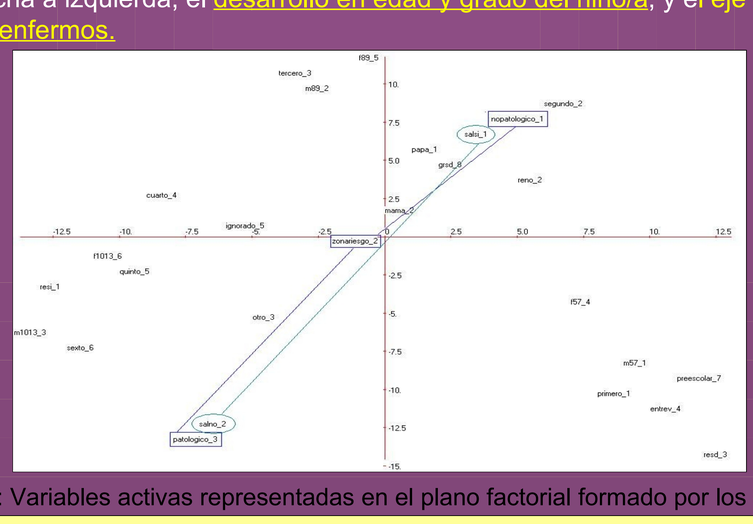


Figura 1: Variables activas representadas en el plano factorial formado por los ejes 1 y 3

Análisis Estadístico de Datos Textuales:

Análisis de Datos Textuales: conjunto de técnicas estadísticas que permite la exploración y análisis de textos.

El método estadístico se basa en medidas y recuentos sobre objetos a comparar. Para aplicar este principio, hay que definir las unidades mínimas del texto, denominadas unidades léxicas.

Unidades Léxicas Simples

Forma gráfica: secuencia de caracteres no delimitadores comprendida entre caracteres delimitadores
Lema: vocablo que cuenta con una misma raíz y significado equivalente.

Lematizar: reagrupar distintas ocurrencias del texto que corresponden a una misma raíz. Proceder así:

- ❖ transformar verbos a su infinitivo (Ej: comí, como, comeré → comer)
- ❖ transformar los sustantivos al singular (Ej: casas → casa)
- ❖ transformar adjetivos al masculino singular (Ej: buenas, buen, buena → bueno)

Unidades léxicas complejas

Para tener en cuenta las palabras compuestas, modismos y expresiones estereotipadas, existen los siguientes criterios:

- **Inseparabilidad:** imposibilidad de insertar unidad léxica en el interior. Ej: pasta de **muchos** dientes.
- **Conmutación:** permite suplantar un elemento por otro. Ej: "la **pasta de dientes** se vende en las farmacias"; la expresión pasta de dientes puede cambiarse por champú, jabón...

En función de lo anterior se definen:

Segmento repetido: secuencia de dos o más palabras, no separadas por un delimitador de secuencia y que aparecen más de una vez en un corpus. Ej: crema de enjuague.
Cuasisegmento: palabras que aparecen en una determinada secuencia, existiendo entre ellas una distancia máxima de separación fijada. Ej: hacer (...) deporte, incluye las siguientes secuencias: **hacer** algo de **deporte**, **hacer** un poco de **deporte**,...

El "CORPUS" P

A	B	C	D	A	E,	B	C	D.	B	A	F	D	A	C	F.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

- $V_1 = 1 \quad V_2 = 1 \quad V_3 = 3 \quad V_4 = 1$
- $\sum_{i=1}^{Fmax} V_i = V = 6$
- $\sum_{i=1}^{Fmax} V_i \times i = T = 16$

La letra **T** designa el tamaño del corpus, en el corpus P, T=16, y el vocabulario, **V**, se compone por seis palabras distintas.

Se denota $\mathbf{V_i}$ al el efectivo de palabras de frecuencia *i*

La suma de efectivos por frecuencias es igual al tamaño del vocabulario, **T**.

Documentos Lexicométricos

Es una reorganización de las ocurrencias del texto. El orden puede ser:

Alfabético	Jerárquico
① A 4	① A 4
② B 3	② B 3
③ C 3	③ C 3
④ D 3	④ D 3
⑤ E 1	⑤ F 2
⑥ F 2	⑥ E 1

diferentes contextos que rodean una palabra (forma-polo o palabra pivote) elegida según criterio de interés.

CONCORDANCIA DE LA PALABRA *A* EN EL CORPUS P

Contexto anterior	Palabra - pivote	Contexto posterior
	A	B C D A E, B
A B C D	A	E, B C D. B
B C D. B	A	E D A C F.
B A E D	A	C F.

Tablas Léxicas

- **Tabla léxica:** unidades léxicas simples por "individuos"
- **Tabla léxica agregada:** unidades léxicas simples por grupos de "individuos"
- **Tabla de segmentos repetidos:** se comporta igual que las anteriores,
- sustituyendo unidades léxicas simples por complejas.

Elementos Característicos y Respuestas Modales

Las representaciones espaciales que resultan de las proyecciones en los planos factoriales se pueden enriquecer mediante resultados de naturaleza más probabilística: los elementos o unidades léxicas características. Además, se pueden caracterizar los grupos de respuestas mediante las denominadas respuestas modales. Las respuestas modales contienen un gran número de palabras características de la parte del corpus a la que pertenecen.

Elementos Característicos

Elementos léxicos sobre o subutilizados en cada parte del corpus en comparación con la frecuencia global en todo el corpus.
Se supone que las palabras tienen una distribución hipergeométrica y de esta manera se compara la diferencia entre la frecuencia global de una palabra y su frecuencia en la parte estudiada.

Respuestas Características

Las respuestas características no son respuestas artificiales que dan un resumen de lo respondido por cada grupo, sino respuestas auténticas, seleccionadas por su representatividad para una categoría de individuos.

Análisis Estadístico de Textos aplicado a las preguntas abiertas

Visualización de los datos

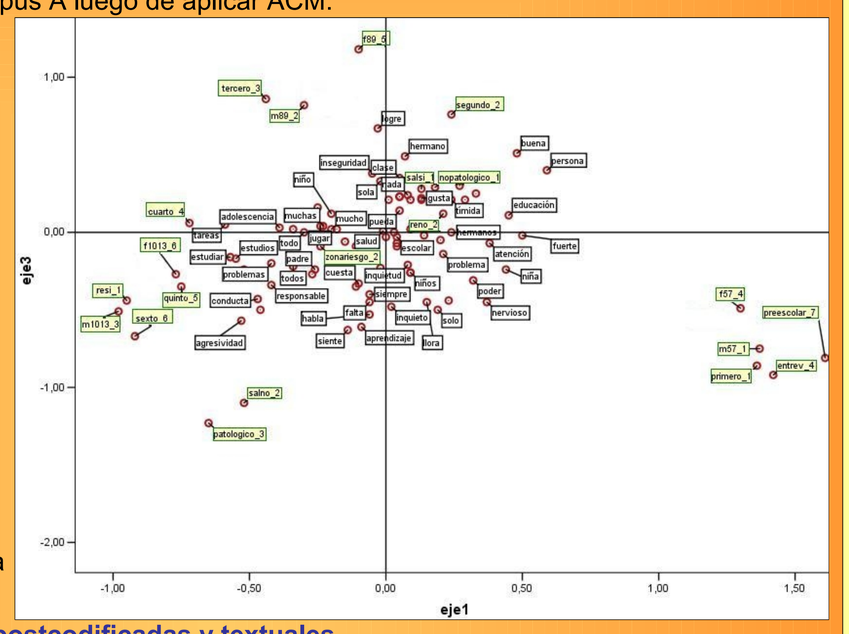
El número total de ocurrencias asciende a T = 24543, con un vocabulario –es decir, cantidad de palabras diferentes– V = 2600, con lo cual el número de palabras distintas asciende a 10,6 %. Mientras que para el subcorpus A: $T_A = 11023$ $V_A = 1771$ el porcentaje de palabras distintas es 16,1% En una etapa posterior, al tomar un umbral de frecuencias de 10 → $T^*_A = 7848$ y su respectivo vocabulario $V^*_A = 148$.

Correspondencia múltiple

La siguiente gráfica muestra la proyección, en el plano factorial formado por los ejes 1 y 3, de las variables activas y de las palabras del subcorpus A luego de aplicar ACM.

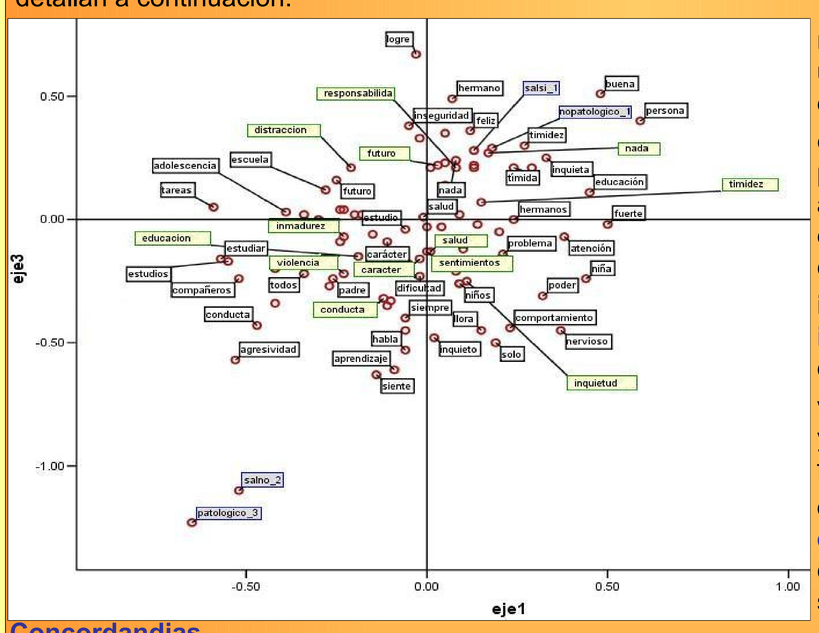
Cuadrante 1: se destaca la presencia de ***nada***, asociada a las categorías *Saludable* y *No patológico* de las variables activas Saludable y Sindrome global, respectivamente; que son las variables que mejor explican el eje 3, donde nada posee un valor-test significativo.

Cuadrante 3: muestra a la palabra ***agresividad*** cercana a las categorías No saludable y Patológico de las respectivas variables activas, cercana a la palabra ***conducta***. La primera es significativa para ambos ejes factoriales, mientras que la segunda lo es sólo para el primer eje.



Comparación de las respuestas postcodificadas y textuales

La siguiente figura muestra la proyección de las palabras del subcorpus A y las categorías de la pregunta A poscodificada. En ella se observan algunas asociaciones entre las palabras y categorías, que se detallan a continuación.



nada: tanto la categoría como la unidad léxica **nada** son muy próximas entre sí.

conducta: la palabra homónima posee valores-test significantes en ambos ejes factoriales, además de estar muy próxima en el gráfico a la categoría.

inquietud: las unidades léxicas **inquieto** y **comportamiento**, al igual que esta categoría, presentan valores test significativos para el eje 3 y sus proyecciones en el plano factorial 1-3 son cercanas.

educación: las palabras **estudiar** y **estudios** son cercanas a esta categoría, y poseen valores-test significativos.

Concordandias

La palabra **nada** aparece 161 veces en el subcorpus A. El 59% de los casos la respuesta que dan los padres sobre qué les preocupa acerca de sus hijos es **nada**. Sólo en un 8%, el entorno de esta palabra tiene significados diferentes: "momentos como que no le importa **nada**", "no quiere **nada** con nadie". En estos casos el sentido semántico se aparta de lo que podría denominarse "nada" como "exento de".

```
Concordance of words equivalent with:      nada
-----------------------------------------
          frequency of repetition  161                          response

                                                      nada      - 0008
 demasiado dada con todas las personas y no tiene miedo a nada  - 0010
                                 por ahora no me preocupa nada  - 0014
                                       no me preocupa nada      - 0020
                                             no tengo nada      - 0021
                                          por ahora nada        - 0022
                                                      nada      - 0024
                                          por ahora nada        - 0029
                    me preocupa que no tiene temor a casi nada  - 0036
```

Conclusiones:

Las herramientas presentadas del **Análisis Estadístico de Textos**, permiten realizar un análisis **más rico y con una menor pérdida de información**. Por ejemplo, en el caso de análisis de preguntas abiertas, el analista no interviene sino hasta la interpretación final de los datos; es decir que no se presenta sesgo en la preparación de los mismos.

Por otra parte, en el caso en particular presentado no se encontraron grandes diferencias en la utilización de las diferentes técnicas (ACM, ADT). A priori, esto puede ser consecuencia de como se codificaron las preguntas abiertas.

Bibliografía:

- Lebart, L.; Salem, A.: *Statistique Textuelle*; Dunod ed., París, 1994.
- Lebart, L.; Salem, A.; Bécue, M.: *Análisis Estadístico de Textos*; Ed. Milenio, Lleida, 2000.
- Viola, L.; Garrido, G., Varela, A.: *Estudio Epidemiológico sobre la Salud Mental de los niños uruguayos*; Clínica de Psiquiatría Pediátrica, Facultad de Medicina, UDELAR, 2007.