Exploración de patrones en los comentarios de la ENDIREH

Angélica Nayeli Rivas Bedolla
Licenciatura en Tecnologías para la Información en Ciencias
Escuela Nacional de Estudios Superiores Unidad Morelia
angelica.nayeli@comunidad.unam.mx

Figure 1: Zapatos Rojos visibiliza violencia contra las Mujeres.

KEYWORDS

ENDIREH, descomposición matricial, violencia contra la mujer

ACM Reference Format:
Angélica Nayeli Rivas Bedolla. 2021. Exploración de patrones en los comentarios de la ENDIREH. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

La violencia contra la mujer es un problema que muestra una tendencia creciente a nivel mundial, que reafirma un estatus de subordinación, atenta contra los derechos de las mujeres y disminuye la calidad de vida de las mismas. Esto hace que sea vital prestar atención a este problema.

La Convención Interamericana para Prevenir, Sancionar y Erradicar la Violencia Contra la Mujer, en el artículo 1° define la violencia contra la mujer como "[...] cualquier acción o conducta basada en su género, que cause muerte, daño o sufrimiento físico, sexual o psicológico a la mujer, tanto en el ámbito público como en el privado" [de Belém do Pará 1994].

El Instituto Nacional de Estadística y Geografía (INEGI) por medio de La Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH), contribuye al desarrollo de investigaciones y diseño de políticas públicas orientadas a atender y erradicar la violencia contra la mujer a nivel nacional. Con esta encuesta se obtienen estimaciones estadísticas que detallan los actos violentos

experimentados con mayor frecuencia por las mujeres, en los ámbitos escolar, laboral, familiar y comunitario, los sujetos que ejercen dicha violencia, el espacio físico en donde lo realizan, las relaciones de pareja, las acciones violentas experimentadas a lo largo de estas y la denuncia de estos actos ante las instituciones correspondientes [INEGI 2016b].

La sección IX de la encuesta profundiza en tipo de violencia obstétrica. Las preguntas de esta sección están enfocadas a mujeres entre 15 y 49 años que estuvieron embarazadas durante el tiempo que comprende la encuesta. Las preguntas están representadas en la Figura 2.

2 OBJETIVO

Identificar los tipos de comentarios reportados en la Encuesta Nacional sobre Dinámica de las Relaciones en los Hogares (ENDIREH) 2016, mediante el uso de herramientas de aprendizaje automático, en específico minería de texto.

3 DESCRIPCIÓN DE LOS DATOS

Del conjunto de datos que dispone en CSV la ENDIREH, tiene una sección, la sección 19, que contiene los comentarios y pensamientos finales de la encuestada. El archivo CSV tiene 25 atributos que representan: ID_VIV (identificador de vivienda), ID_MUJ (identificador de mujer), HOGAR (número de hogar de la mujer), DOMINIO (tipo de vivienda), CVE_ENT (clave de entidad federativa), NOM_ENT (nombre de entidad federativa), CVE_MUN (clave de municipio), NOM_MUN (nombre de municipio), COD_RES_MU (si se completó la entrevista), COD_RES (por quién se completó la entrevista), T_INSTRUM (tipo de entrevista hecha), P19_1 (comentarios adicionales), P19_1e (el comentario), P19_2 (cómo se sintió la entrevistada), P19_2e (un sentir diferente al especificado en respuestas), y, O_ACT_VIO (actos violentos diferentes a los especificados en el cuestionario). Entre otras columnas que no tienen explicación de qué información representan.

De los 25 atributos que hay en esta sección, nos interesan en particular 3: P19_1e , P19_2 y O_ACT_VIO.

P19_1e es el atributo que contiene los comentarios hechos por las personas entrevistadas, con esto se hará la minería de texto. Los comentarios tienen un total de 111,256 instancias, de las cuales 108,329 son registros vacíos. Por lo tanto, solo 2,927 de las entrevistadas dijeron un comentario sobre la entrevista.

El atributo O_ACT_VIO, detalla si el comentario dicho por la persona entrevistada tiene que ver con alguna categoría de acto violento. Los valores que esta columna puede tomar son: físico, emocional, discriminación laboral, violencia obstétrica, sexual, otra o nulo. Siendo nulo cuando su comentario no tiene que ver con algún tipo de violencia. El total de registros que tienen una categoría de O_ACT_VIO asignada son 44, tan solo el 0.015% de los datos son sobre violencia.

El atributo P19_2, detalla el sentir de la persona entrevistada sobre la entrevista. Los valores que esta columna puede tomar son: bien, mal, indiferente, otro y no especificado. De los cuales 2360 se sintieron bien, 344 se sintieron mal, 114 se sintieron indiferente y 109 se sintieron de otra manera. Lo cual señala que el 80% de los datos son comentarios positivos hacia la entrevista.

4 CATEGORIZACIÓN

Se eligió hacer la categorización de los comentarios con factorización matricial no negativa. La factorización matricial no negativa se eligió para la fácil exploración de la matriz W (palabras, temas) y así determinar las palabras que caracterizan cada tema.

Para el preprocesamiento del texto se planificaron dos funciones sobre la columna P19_1e, que se aplicará al crear la matriz de término documento. En la primera función sólo hace un tokenizador que elimina signos de puntuación, y la segunda se tokenizan los signos de puntuación y aplica el SnowballStemmer de NLTK. Lo ideal sería aplicar un lematizador para no perder mucha información pero las librerías de Python no tienen un lematizador que funcione bien en español.

Para la categorización se creó una función. Recibe la función que actuará de procesador, la columna de dataframe de pandas que contiene los comentarios, el número de temas a descomponer la matriz término documento y la cantidad de palabras que pertenecen a cada tema. La función hace una transformación a TF-IDF seguida de una factorización matricial no negativa, ambas funciones obtenidas de la librería Scikit Learn, al final se obtienen las k palabras más representativas de cada tema de la matriz W.

Se propusieron 3 números de componentes para la factorización no negativa para cada matriz TF-IDF obtenida con las dos funciones de preprocesamiento: con 2 componentes, con 4 y con método del codo. Para 2 como número de componentes es porque en la columna O_ACT_VIO existen 2 tipos principales de comentarios: de violencia y no violencia, aunque no están representados en misma cantidad, veremos si puede diferenciarlos. Para 4 como número de componentes es porque la columnas P19_2 tiene 4 categorías, tampoco están representados en la misma cantidad pero al menos no es tan abrupto como en O_ACT_VIO. Con el método del codo se elige de acuerdo a la inercia en el método de agrupación de K-Medias, se calcula el codo para $k \in [1, ..., 15]$.

4.1 Análisis y discusión de resultados

Las tablas presentadas serán las 5 palabras más representativas de cada experimento, siendo cada columna un tema/componente distinto y los renglones son sus palabras asociadas.

Con 2 componentes se obtienen los mismos resultados con los dos tipos de preprocesamiento usando 5 como semilla de números aleatorios. Las 5 palabras características resultantes de esta descomposición matricial se ven reflejadas en Tabla 1 y Tabla 2. Las categorías que percibo son:

(1) Lo que las personas encuestadas piensan de la encuesta en sí.
(2) Cómo se sintieron las personas encuestadas durante la encuesta.

Da buenos resultados ya que podemos obtener dos cosas de valor sobre la encuesta, el que tal recibió el público la encuesta y la satisfacción de este. Pero no lo que buscábamos que era clasificar por tipo de mensaje, violento o no.

Con 4 componentes se obtienen los mismos resultados con los dos tipos de preprocesamiento usando 66 como semilla generadora de números aleatorios Tabla 3 y Tabla 4. El significado que le doy a las categorías es:

(1) La vida conyugal y violencia que vive la persona entrevistada.

Table 1: 2 componentes sin stemmer

	0	1
0	tiempo	sentí
1	comenta	parece
2	entrevistada	parecio
3	preguntas	preguntas
4	encuesta	muchas

Table 2: 2 componentes con stemmer

	0	1
0	tiemp	senti
1	coment	parec
2	entrevist	pareci
3	pregunt	pregunt
4	encuest	much

(2) Cómo se sintieron las personas encuestadas durante la encuesta.
(3) Lo que las personas encuestadas piensan de la encuesta, una mezcla entre positivo y negativo.
(4) Lo que las personas encuestadas piensan de la encuesta en sí pero en un pensamiento negativo.

Se supone que de las encuestadas que dieron un comentario, un 80% se categorizó como sentirse bien, pero no salió alguna categoría positiva, solo negativo y mezcla entre positivo y negativo, lo cual hace que sea una repetición innecesaria sobre el mismo tipo de comentarios. Aparte apareció una categoría que representa violencia, a la mayor cantidad de comentarios se ven representados esos 44 comentarios sobre violencia.

Table 3: 4 componentes sin stemmer

	0	1	2	3
0	vida	senti	preguntas	aburrida
1	pareja	sintio	buenas	tiempo
2	hijos	tranquila	gracias	extensa
3	problemas	parece	repetitivas	tediosa
4	violencia	parecio	largo	preguntan

Table 4: 4 componentes con stemmer

	0	1	2	3
0	pregunt	vid	senti	aburr
1	buen	parej	parec	pregunt
2	repetit	hijs	tranquila	tiemp
3	larg	problem	pues	extens
4	graci	violenci	pareci	tedios

Usando el método de codo si hubo diferencias en los tipos de preprocesamiento.

Con un preprocesamiento sin stemmer, el codo indica un primer codo en k = 3 Figura 2, aunque no es tan marcado como en 5, existe. Usando una semilla generadora de números aleatorios de 42 las palabras obtenidas se reflejan en la Tabla 5, las categorías que inferí fueron:

(1) Lo que las personas encuestadas piensan de la encuesta en sí pero en un pensamiento negativo.
(2) Cómo se sintieron las personas encuestadas durante la encuesta.
(3) La vida conyugal y violencia que vive la persona entrevistada.

Esta agrupación tiene buen rendimiento porque tiene de las 3 cosas que queríamos: violencia, pensamiento de la encuesta y su sentir sobre la encuesta, y sin ambigüedad entre la diferencia de las categorías.

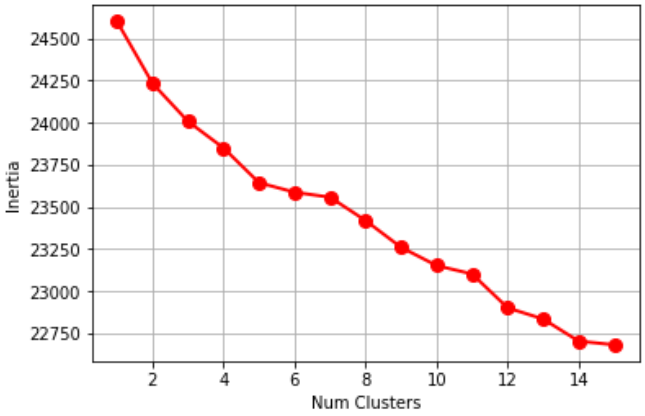

Figure 2: Método del codo sin stemmer.

Table 5: 3 componentes sin stemmer

	0	1	2
0	aburrida	tranquila	hijos
1	pregunta	gracias	problemas
2	demasiado	sintio	mujer
3	extensa	haciendo	esposo
4	tediosa	pues	violencia

Con un preprocesamiento con stemmer, el codo indica un primer codo en k = 5 Figura 3. Con una semilla de números aleatorios de 2 se obtuvieron las palabras reflejadas en la Tabla 6, se obtuvieron estas palabras. Lo que inferí fue:

(1) La vida conyugal y violencia que vive la persona entrevistada.
(2) Lo que las personas encuestadas piensan de la encuesta, una mezcla entre positivo y negativo.
(3) La relación entre violencia, gobierno y hombres.
(4) Lo que las personas encuestadas piensan de la encuesta en sí pero en un pensamiento negativo.

(5) Cómo se sintieron las personas encuestadas durante la encuesta en sí pero en un pensamiento positivo.

Tiene la misma repetición innecesaria de categorías que cuando eran 4 componentes pero esta vez salió una columna interesante. Gobierno, trabajo, hombre y violencia en los primeros puestos. Me atreví a llamarle Patriarcado porque esto es lo que representa.

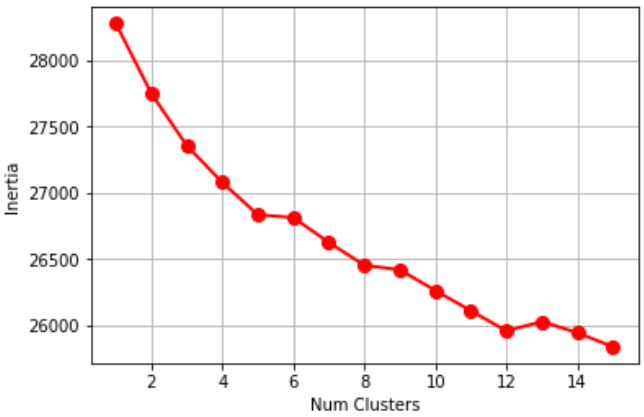

Figure 3: Método del codo con stemmer.

Table 6: 5 componentes con stemmer

	0	1	2	3	4
0	hij	buen	gobiern	aburr	pregunt
1	parej	sinti	trabaj	pregunt	tranquil
2	espos	repit	hombr	demasi	senti
3	problem	repetit	tip	extens	tip
4	violenci	pregunt	violenci	tedios	bien

Los comentarios salieron en su mayoría negativos, no porque el pensamiento en general de la encuesta sea negativo, sino que todas las personas que dijeron comentarios lo hicieron para expresar esta inconformidad.

Al tener más componentes nos arriesgamos a que varios de estos tengan significados similares, como la percepción de los comentarios negativos y neutros, pero también ayuda a obtener categorías que cuando hay menos componentes son asignados a categorías con mayor impacto. Así pudimos ver el tipo de comentarios patriarcales que ahonda en el tipo de violencia sistemática que están comentando las encuestadas.

La categorización de 3 componentes es la mejor categorización para estos comentarios, ya que resuelve el problema de tener demasiadas componentes donde varios de sus componentes tiene significados similares o de diferencia dudosa, pero tiene las suficientes para obtener buena información.

5 REFLEXIÓN

El estudio de los tipos de comentarios y lo que reflejan puede beneficiar al INEGI para tomar decisiones sobre las futuras ediciones de la encuesta. Puesto que la mayoría de estos comentarios mostraban una opinión negativa hacia la encuesta (aburrida, larga duración, repetitiva, entre otros) se podría basar en estos para mejorar la entrevista. Como se dijo en las reflexiones de los resultados, estos resultados negativos se pueden malinterpretar, ya que por cultura mexicana es más probable de expresar su descontento hacia la encuesta que su conformidad. Para resolver esto se proponen dos acercamientos: considerar el contexto y obtener más datos.

Si se considera el contexto de los comentarios, como otras variables obtenidas en la encuesta, se puede saber si tienen correlación con la decisión de decir un comentario o no, y el tipo de comentario. Esto tiene más implicaciones ya que los datos siempre enfatizan ciertos aspectos de la investigación y dejan de lado otros elementos. Por ello, se necesita hacer un análisis ético preciso. Como hacen notar Lauren F. Klein y Catherine D'Ignazio en el libro, *Data Feminism* [Klein and D'Ignazio 2020], es esencial conectar los datos con el contexto en el que fue producido, preguntando sobre las condiciones socioculturales, históricas, institucionales y materiales en los que fueron recopilados y la identidad de quienes lo producieron. En este caso, al ser datos mexicanos, están sujetos a la cultura mexicana, como los tipos de violencia (e.g. discriminación a las muxes), la recurrencia de la violencia, la forma de expresarse (e.g. el regionalismo *pues*), la encuesta en sí, entre otros. La población muestral de la encuesta se seleccionó a partir de las mujeres que denunciaron algún tipo de violencia ante las autoridades correspondientes y de las encuestas de viviendas que hace el INEGI, con un diseño trietápico, estratificado y por conglomerados, dando una media de 4,400 viviendas por entidad federativa dejando un total de 142,363 viviendas encuestadas [INEGI 2016a]. El contexto ayuda a entender las limitaciones de los datos. Por lo cual, aunque se siga la misma metodología en otro país se obtendrán resultados diferentes

Si se tuvieran comentarios de todas las personas que fueron encuestadas durante esta edición o los comentarios de todas las ediciones de la ENDIREH, se podría contrastar mejor los resultados obtenidos, pero esto también tiene implicaciones. Más datos no significa necesariamente mejor, ya que no importa la cantidad que se tiene si su contexto no es lo necesariamente rico para reflejar la realidad. Por ello, los datos se deben racabar en profundidad y no en cantidad. Siendo que la mayoría de los comentarios inician con ËLA ENTREVISTADA DIJO...; ËLA INFORMANTE DIJO...; ËLA ENCUESTADA DESCRIBE...; entre otros, a parte de agregar ruido a los datos, no hay información que inferir con base en esto, y si todos los comentarios siempre son así su profundidad será pobre aunqu su cantidad sea grande.

REFERENCES

Convensión de Belém do Pará. 1994. Convención Interamericana para prevenir, sancionar y erradicar la violencia contra la mujer. *Belém Do Pará* (1994), 11–12.

INEGI. 2016a. Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares 2016 (ENDIREH) - Diseño Muestral. (2016), 1.

INEGI. 2016b. Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares (ENDIREH) 2016. https://www.inegi.org.mx/programas/endireh/.

Lauren F. Klein and Catherine D'Ignazio. 2020. Data Feminism. (2020).