Análisis sobre el idioma español en México, con base en la frecuencia de palabras azules, rojas, obscenas y vulgares en Twitter

Orlando Ramos, Luis Alfredo Moctezuma, Jesús García,
David Pinto, Rodolfo Martínez

Benemérita Universidad Autónoma de Puebla, Puebla
México

{orlandxrf, luisalfredomoctezuma}@gmail.com
gr_jesus@outlook.com, dpinto@cs.buap.mx

Resumen. En este artículo se presenta una comparativa entre estados de la República Mexicana de la frecuencia de palabras azules, rojas, obscenas y vulgares que escriben usuarios de la red social de microblogging Twitter. Se presentan gráficas de los resultados obtenidos. El objetivo es mostrar en mapas del comportamiento de la frecuencia de palabras por cada estado y clasificados por el tipo de palabra analizada. Los experimentos fueron realizados sobre un corpus de tweets.

Palabras clave: Palabras azules, palabras rojas, palabras obscenas, palabras vulgares, tweets.

1. Introducción

Una de las características que definen a México sobre el idioma español es su riqueza lingüística en relación a palabras con connotación sexual y ofensivas que utilizan las personas en la vida cotidiana, muchas veces en doble sentido. Estas expresiones verbales o escritas son consideradas por la sociedad como malas palabras. En cuanto a las palabras azules se utilizan en un contexto positivo, mientras que las palabras rojas son utilizadas en un contexto negativo.

El procesamiento del lenguaje natural en la red social Twitter nos da conocimiento de la dispersión del idioma español a lo largo de la república Mexicana. Al tener los datos se puede buscar información específica sobre temas de interés, en este caso la búsqueda de palabras rojas, azules, obscenas y vulgares. El objetivo de este trabajo es identificar los estados de la República Mexicana en los que se concentra una mayor frecuencia en el uso de ellas y mostrarlas de una manera más clara para el lector.

2. Trabajo relacionado

En esta sección se describen trabajos previos que tratan sobre el análisis de frecuencias de determinadas palabras por ubicación geográfica divididas por entidades federativas.

En [1] se realizó un análisis de la dispersión del idioma español en México con respecto a la frecuencia n-gramas de letras que fueron calculados por regiones geográficas de la República Mexicana y comparados con la media nacional para obtener la frecuencia por entidad federativa, donde se usaron corpus de noticias periodísticas y de tweets. En general, se calcula la frecuencia de cada n-grama (unigrama, bigrama o trigrama) y se ordenan los n-gramas en forma descendente. Se usa una porción de los n-gramas más frecuentes y se calcula el grado de intersección entre los n-gramas calculados a nivel nacional y aquellos calculados a nivel estatal. El grado de traslape indica la cercanía del idioma en el estado con respecto a la media nacional.

En [2] se utilizó un corpus de tweets para clasificar la frecuencia de palabras obscenas y vulgares por entidad federativa de la República Mexicana, se realizó un análisis estadístico sobre los tweets con base en los diccionarios de vulgaridades y obscenidades se obtuvieron las frecuencias. Se realizó un proceso de balanceo quedando la misma cantidad de tweets correspondiente a la clase menos representativa en dicho conjunto. Una vez desarrollado el corpus de entrenamiento, se utilizaron para el desarrollo del modelo de clasificación todas las palabras que aparecen en cada tweet. No se aplicó ningún preprocesamiento al corpus de entrenamiento construido.

En [5] se realizó una propuesta para el estudio de campos semánticos en Twitter, este trabajo se plantea con el fin de dar paso a una serie de investigaciones donde se pueda usar la técnica para extracción de tweets. En este trabajo se usan bases de datos para almacenar los tweets que son extraídos en formado JSON usando la API original de twitter, solo almacenando los datos, el uso que se les pudiera dar es dejado para posibles trabajos futuros. La extracción masiva de tweets no es una opción en este trabajo ya que solo se trabajan sobre perfiles preestablecidos y tomados aleatoriamente.

Se utilizó [3] para complementar los diccionarios de palabras obscenas y vulgares usados en [1], sin embargo, algunas frases que aparecen ya no son utilizadas de manera frecuente en la actualidad por usuarios de la red social de Twitter. En este trabajo se hizo uso de herramientas que permitieron la extracción masiva de tweets y poder analizar una mayor cantidad de perfiles y poder analizar información específica y de relevancia en el procesamiento de lenguaje natural. La contribución más fuerte presentada en este trabajo es que usamos técnicas para tener una representación de las palabras que más se usan por estado de la república mexicana, con esta información podemos ver las tendencias en el uso de ciertas palabras por estado pudiendo comparar con factores ocurridos en cada estado, por ejemplo al analizar un estado con problemas de delincuencia, ver que palabras son las más usadas y proponer una representación de que palabras usan regularmente los delincuentes o gente que convive con ellos.

3. Procesamiento de lenguaje natural

La red social de microblogging Twitter ofrece una API para desarrolladores, que consiste en un conjunto de métodos para poder hacer uso de los datos públicos, en este caso se obtuvieron tweets, que es un mensaje de 140 caracteres compartido públicamente con una comunidad.

A continuación se describe cada tipo de palabras analizadas en datos obtenidos.

- Palabra azul: se refiere a la palabra que expresan en el autor un sentimiento positivo, por ejemplo; alegrar, gozar, respetar, ganar, cuidar etc.
- Palabra roja: este tipo de palabra manifiesta un sentimiento negativo en el autor, por ejemplo; enojar, llorar, juzgar, burlar, gobernar, etc.
- Palabra obscena: se usa generalmente para calificar cierto tipo de lenguaje, sobre todo a las palabras que tienen connotación sexual, por ejemplo; mocos, chichis, culo, huevos, etc.
- Palabra vulgar: hace referencia a palabras de carácter ofensivo, conocidas como palabras altisonantes, por lo regular vistas por la sociedad como malas palabras, por ejemplo; puto, chingada, cabrón, pendejo, etc.

El corpus construido para este experimento consiste de tweets, extraídos por medio de la API mencionada anteriormente, con base en las coordenadas geográficas de cada uno de los estados de la república mexicana, en la Tabla 3.1 se muestra la cantidad de tweets que se recolectaron para este trabajo. Cabe mencionar que la cantidad de tweets depende del estado de la República Mexicana, en algunos estados los tweets que se pueden extraer son menores que en otros, es por ello que se ha tomado al estado con un menor número de tweets y se han recortado los demás estados para que todos tengan la misma cantidad de tweets.

Tabla 3.1. Corpus de Tweets.

Para cada uno de los 32 estados se utilizaron	
Tweets	460
Total	14720

En la Tabla 3.2 se muestran los cuatro diccionarios que se utilizaron para calcular la frecuencia de los diferentes tipos de palabras. Los diccionarios de palabras obscenas y vulgares, se construyeron en base a una encuesta a 20 personas de entre 23 y 34 años, a quienes se les pidió que escribieran las palabras que usaban con más frecuencia. Esta decisión fue tomada porque los diccionarios con los que se contaba sobre palabras obscenas y vulgares estaban desactualizados (pasados de moda), por este motivo se decidió actualizarlos.

Tabla 3.2. Diccionarios utilizados.

Palabras			
Obscenas	**Vulgares**	**Azules**	**Rojas**
158	104	78	94

4. Experimentos

Se realizó una comparativa de los datos guardados en el corpus y los diccionarios descrito en la sección 2, dicha comparativa se realizó contando la frecuencia con la que aparecieron las palabras de los diccionarios en el corpus de cada estado.

4.1 Preprocesamiento del corpus

Para poder comparar las palabras de los diccionarios con los tweets que se extrajeron se realizó un tratamiento a dichos tweets que consistió en eliminación de caracteres especiales, signos de puntuación, cada letra en mayúscula se sustituyó por su correspondiente en minúscula, sustitución de retornos de carro y nueva línea para el reacomodo de los tweets, dejando cada uno en un renglón, ya que la API Twitter proporciona los datos tal como los ingreso el usuario.

Para la comparativa con las palabras vulgares y obscenas se realizó la comparativa de dichas palabras tal y como aparecen en los diccionarios, sin embargo, para las palabras azules y rojas se lematizó el corpus de cada estado, así como los diccionarios correspondientes usando TreeTagger.

4.2 Análisis de palabras

De los diccionarios de palabras se comparó con el corpus de cada estado. El resultado fue un archivo con las palabras encontradas y su frecuencia, lo cual permite hacer una comparativa entre todos los estados.

En la Tabla 4.2.1 se muestra el resultado obtenido de analizar los corpus de palabras obscenas, vulgares, azules y rojas explicadas anteriormente. Para poder comparar los resultados obtenidos entre todos los estados lo que se hizo fue obtener un 100% del total de palabras vulgares encontradas y buscar cuantas palabras del total se encontraron en cada estado. Esto nos permite decir que del total de palabras usadas en todos los estados, las obscenas se usaron en un 0.93% en Aguascalientes y el 0.40% se usaron en Baja california Norte por ejemplo. De igual manera podemos ver que el porcentaje de palabras vulgares para cada estado representa la cantidad de veces que se usaron respecto de los otros estados.

Tabla 4.2.1. Frecuencias encontradas por estados.

Estado	Palabras			
	Obscenas	**Vulgares**	**Azules**	**Rojas**
Aguascalientes	0.93%	1.29%	2.68%	1.42%
Baja California Norte	0.40%	0.89%	1.41%	2.28%
Baja California Sur	1.46%	0.98%	1.09%	1.11%
Campeche	0.67%	1.25%	2.01%	1.92%
Chiapas	2.66%	2.46%	2.42%	0.81%
Chihuahua	0.80%	1.34%	1.30%	1.32%
Coahuila	3.06%	6.03%	8.73%	12.96%
Colima	0.40%	0.58%	1.15%	0.66%
Distrito Federal	7.06%	8.17%	6.43%	6.28%
Durango	1.20%	1.38%	1.97%	1.82%
Guanajuato	2.80%	3.30%	3.05%	1.21%
Guerrero	1.73%	1.47%	1.04%	1.92%
Hidalgo	1.20%	1.92%	1.41%	1.47%

Estado	Palabras			
	Obscenas	**Vulgares**	**Azules**	**Rojas**
Jalisco	8.39%	5.58%	6.92%	6.88%
México	16.38%	9.73%	6.37%	6.02%
Michoacán	0.93%	0.94%	1.45%	1.47%
Morelos	0.67%	0.98%	1.45%	1.82%
Nayarit	0.00%	0.13%	0.62%	0.40%
Nuevo León	0.13%	0.22%	2.14%	0.05%
Oaxaca	0.40%	0.40%	1.15%	0.71%
Puebla	2.26%	1.34%	1.03%	1.87%
Querétaro	3.60%	5.94%	4.48%	4.05%
Quintana Roo	2.80%	2.63%	3.37%	3.24%
San Luis Potosí	2.53%	2.81%	2.50%	4.05%
Sinaloa	3.46%	3.79%	3.88%	2.28%
Sonora	10.65%	11.43%	6.43%	6.02%
Tabasco	7.19%	5.04%	5.34%	6.88%
Tamaulipas	2.53%	3.93%	4.07%	4.00%
Tlaxcala	1.73%	1.29%	1.09%	1.57%
Veracruz	5.19%	5.31%	5.82%	5.21%
Yucatán	5.59%	6.16%	6.17%	6.63%
Zacatecas	1.20%	1.25%	1.01%	1.67%

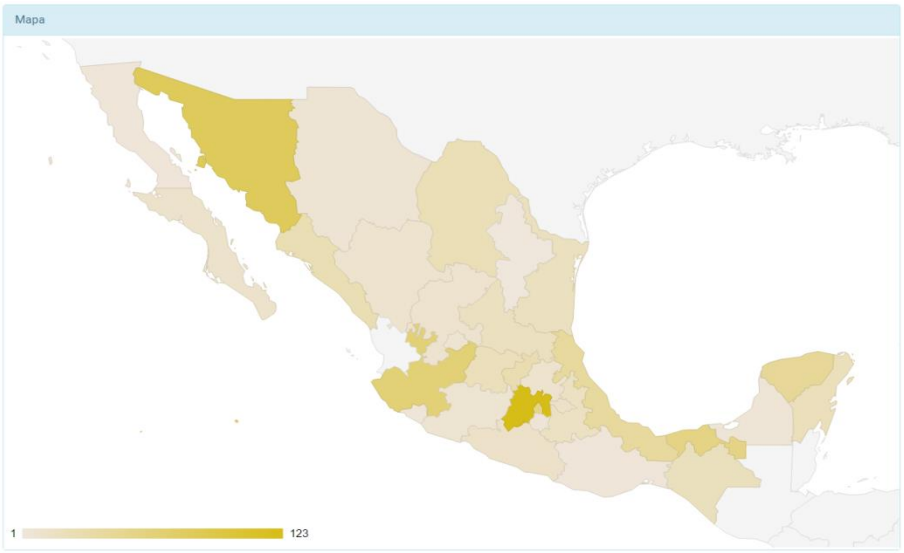

Fig. 1. Mapa para las palabras obscenas.

5. Resultados

Para una mejor comprensión para el lector se muestran los resultados de manera gráfica, en mapas interactivos en los que se aprecian las entidades federativas con la frecuencia con la que se obtuvieron los distintos tipos de palabras. En los mapas se muestra de color más intenso a los estados con una mayor frecuencia, y de color más tenue los de menor frecuencia.

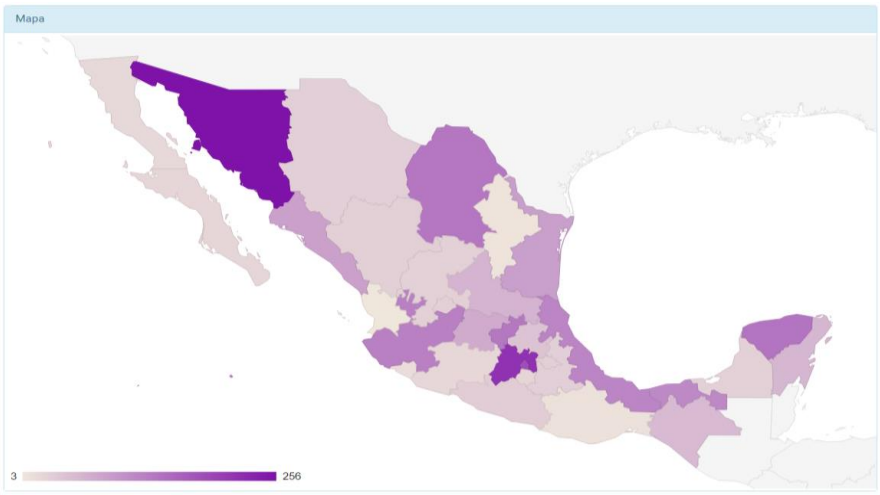

Fig. 2. Mapa para las palabras vulgares.

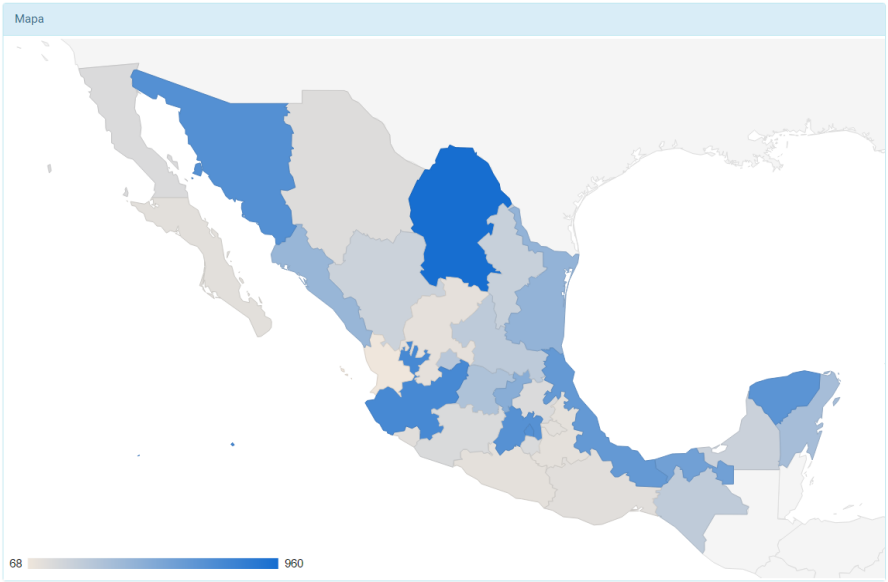

Fig. 3. Mapa para palabras azules.

En la Fig. 1 se aprecia el mapa con la frecuencia obtenida de analizar el uso de las palabras obscenas por cada estado de la República. Los estados de Sonora, Estado de México y Jalisco presentan una mayor frecuencia en el uso de palabras obscenas, mientras que el estado de Nayarit es el único que no presenta incidencias de dichas palabras.

Como se puede observar en la Fig. 2 los estados en los que se encontró un mayor número de incidencias de palabras vulgares es Sonora junto con el Estado de México y los que menos frecuencia obtuvieron fueron los estados de Nayarit, Nuevo León y Oaxaca.

Para la Fig. 3 se muestra la frecuencia de palabras azules en donde se observa claramente que los estados de Coahuila y Jalisco presentan un mayor número de palabras positivas con base en el análisis realizado, mientras que los estados de Nayarit y Zacatecas fueron los que se encontraron dichas palabras con menos frecuencia.

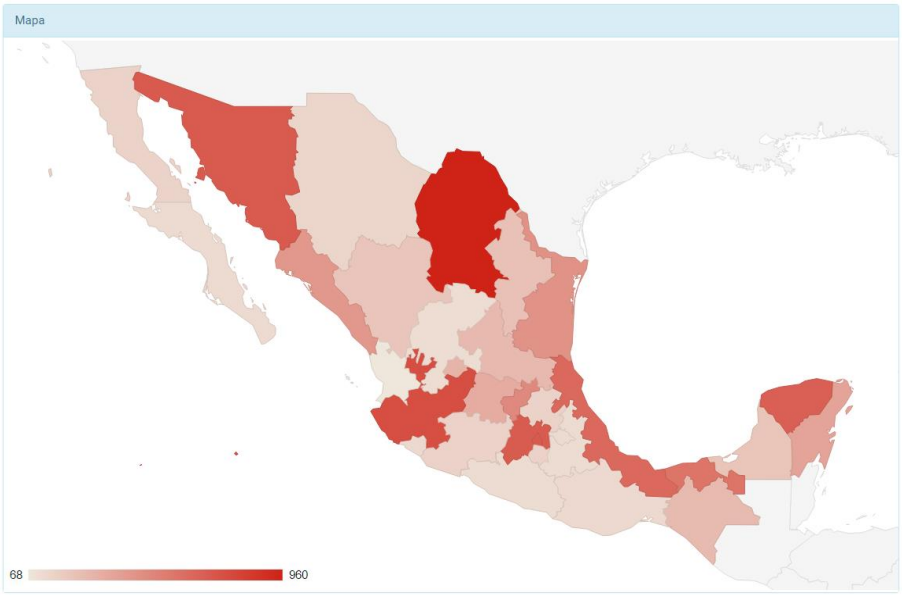

Fig. 4. Mapa para las palabras rojas.

En la Fig. 4 se presentan Coahuila, Sonora y Jalisco como los estados con un mayor número de frecuencias de palabras negativas. Y los estados con menos incidencia de palabras rojas son los estados de Nayarit, Zacatecas y Guerrero.

6. Conclusiones

Los estados de la República Mexicana con una presencia mayor de los cuatro tipos de palabras usados en este trabajo son Sonora, Coahuila y el Estado de México, lo que nos permite concluir que en el norte y centro del país es donde se concentra la mayor frecuencia de palabras azules, rojas, obscenas y vulgares.

Las palabras más utilizadas de los diferentes tipos son:

- Azules: *hacer*, *decidir*, *ganar*.
- Rojas: *ultimar*, *jugar*, *fallar*.
- Obscenas: *verga*, *hueva*, *huevos*.
- Vulgares: *pinche*, *pedo*, *pendejo*.

Como trabajo futuro se pretende incrementar los diccionarios de los cuatro tipos de palabras, así como del corpus de tweets obtenido con base en las mismas coordenadas utilizadas en el presente trabajo, se propone analizar diferentes redes sociales como Facebook, para así comparar la red social en la que se prefiere utilizar los diferentes tipos de palabras.

Referencias

1. Ramos, O., Pinto, D., Priego, B., Olmos, I., Beltrán, B.: Análisis empírico de la dispersión del español mexicano. Research in Computing Science (2014)
2. Guzmán, E., Beltrán, B., Tovar, M., Vázquez, A., Martínez, R.: Clasificación de frases obscenas o vulgares dentro de tweets. Research in Computing Science, Vol. 85, pp. 65–74 (2014)
3. d. l. L. Academia Mexicana: Diccionario de mexicanismos. Siglo XXI Editores México (2010)
4. Gupta, N.K.: Extracting Phrases Describing Problems with Products and Services from Twitter Messages. Computación y Sistemas, Vol. 17, No. 2, pp. 197–206 (2013)
5. Alonso Berroca, J.L.: Propuesta de estudio del campo semántico de los libros electrónicos en Twitter (2012)
6. Fainholc, B.: Un análisis contemporáneo del Twitter. RED – Revista de Educación a Distancia
7. Pla, F., Hurtado, Ll-F.: Análisis de Sentimientos en Twitter. In: Proceedings of the TASS workshop at SEPLN (2013)
8. Alegria, I.: Introducción a la tarea compartida Tweet-Norm 2013: Normalización léxica de tuits en español (2013)