Clasificación de frases obscenas o vulgares dentro de tweets

Estefania Guzmán, Beatriz Beltrán, Mireya Tovar,

Andrés Vázquez, Rodolfo Martínez

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación
Avenida San Claudio, 14 Sur, Ciudad Universitaria
Puebla, México
fanyguzm@gmail.com, {bbeltran, mtovar, andrex}@cs.buap.mx,
beetho@gmail.com
http://nlp.cs.buap.mx

Resumen. Se propone una metodología para la detección frases obscenas y vulgares en los tweets, debido a que México es uno países donde se usa mucho el doble sentido para comunicarse. La metodología propuesta se apoya en un diccionario de mexicanismos etiquetado manualmente por expertos. Se pudo detectar que las palabras obscenas y vulgares son las que más se emplean y los estados del país que más las utilizan. Además en base al diccionario, se logra clasificar un conjunto de tweets, dichos tweet fueron tomados por zonas geográficas de México, consideramos que dichas personas son mexicanas y por lo tanto pueden llegar a manejar el doble sentido. La intención de esto es poder tener un sistema que ayude a encontrar cuando una persona está utilizando el doble sentido dentro de algún texto corto, como pueden ser los tweets y además lograr hacer mapas de México que nos proporcione información los lugares donde se utilizan con mayor frecuencia el doble sentido.

Palabras clave: Palabras obscenas, palabras vulgares, albur, diccionarios, textos cortos.

1. Introducción

El uso del doble sentido en el habla del mexicano es algo común en todas las esferas sociales en mayor o menor grado. En algunas ocasiones es percibido sin intención por la misma vaguedad intrínseca del lenguaje; pero en otras, se utiliza de manera premeditada para provocar un efecto en el interlocutor, como un recurso humorístico o hasta ofensivo.

Además, en México el albur está en todas partes, en la comida, en los oficios, esto es, cualquier palabra cobra doble sentido y la combinación de verbos y sustantivos, acompañados de gestos y sonrisas, se convierten en un albur. Para los expertos, el uso

del doble sentido puede ser un buen método educativo, porque requiere de un amplio vocabulario, atención y destreza mental en pocos segundos [9].

Algunas personas no entienden un albur, y esto puede llevarlas a una confusión, lo cual en particular suele pasar en redes sociales, donde hay que saber distinguir que es lo que realmente las personas quieren expresar. Hoy en día uno de los principales modos de comunicación sobre todo entre los más jóvenes, es a través de twitter. Para los especialistas en procesamiento de Lenguaje Natural se hace interesante estudiar si en este tipo de mensajes se utiliza palabras consideradas como obscenas o vulgares y detectar cuales palabras son las más empleadas.

Según la RAE[1], se define como **vulgar** a frases, palabras o expresiones; las cuales son groseras u ordinarias, siendo catalogadas como impropias para personas cultas o educadas. Y se define como **obsceno** a frases, palabras o expresiones; las cuales son impúdicas u ofensivas al pudor y en la mayoría de los casos con relación al sexo.

Ambas se denotan como groserías y que no son apropiadas para comunicarse en un ámbito formal, y su aplicación es en el contexto informal para expresar dichos, albures, chistes, etc. Pero en resumen para diferenciar si una frase es vulgar u obscena se deberá identificar primero si es un tipo de ofensa al pudor, para descartar que sea obsceno y si no cuenta con ello se cataloga como vulgar

En la presente investigación se desarrollan dos diccionarios de frases vulgares y obscenidades, los cuales permitieron etiquetar un gran número de tweets identificando el doble sentido, para desarrollar posteriormente modelos de clasificación.

La estructura del artículo, se tiene de la siguiente manera, en la sección 2, se presenta el trabajo que se ha realizado en el área de identificación del doble sentido o humor. En la sección 3 se describe la propuesta de la metodología para la identificación de obscenidades o vulgaridades dentro de textos cortos. En la sección 4 se muestran algunos resultados obtenidos y finalmente las conclusiones que se obtienen hasta el momento.

2. Trabajo Relacionado

México se caracteriza por su riqueza lingüística y en ese aspecto los mexicanos saben aprovecharla para expresarse de muchas formas, desde la manera más inocente hasta la más vulgar, teniendo así muchas características de humor en cada situación que se le presenta. Sin embargo hay veces que es difícil identificar si es un cumplido o una burla. Hasta hace unos años identificar si se estaba usando el albur en textos a través de herramientas computacionales no existía, era más fácil la generación que la detección.

El albur es un acto de someter a prueba la agilidad mental de alguien. Por definición éste nunca puede ser inocente, pero sí involuntario, pues se define como un juego de palabras que lleva implícito o explícito una connotación sexual que se puede desprender de cualquier conversación [10]. El albur es un juego de palabras con doble sentido. Este juego de palabras emplea principalmente dos elementos verbales. El

[1] http://www.rae.es/

primer elemento es la asociación de palabras y expresiones aparentemente inocuas con términos sexuales y/o escatológicos. El segundo elemento que se emplea generalmente en el albur es la deconstrucción de palabras inocuas para crear palabras o expresiones completamente distintas, pero que implican un mensaje sexual o escatológico [3].

A través de los años han surgido varios trabajos utilizando Procesamiento de Lenguaje Natural (PLN), para clasificar y recuperar información, sin embargo no se ha realizado ningún sistema el cual sea capaz de detectar el albur de un texto en español. Existen trabajos que han utilizado herramientas como WEKA[2] para definir si textos cortos contienen albur o no [2].

Ocampo en [2], utiliza un diccionario que contiene alrededor de 300 palabras para detectar albur y humor en textos como chistes y rimas, en el idioma español. Para la clasificación de su modelo desarrollado utiliza la herramienta WEKA.

La mayoría de los trabajos han sido desarrollados por Rada Mihalcea, para el idioma inglés [1], [4], [6]. Por ejemplo, Rada y Attardo en [1] hacen uso de características de estilo de textos humorísticos tales como la aliteración, la antonimia y el contenido adulto; así como de características basadas en contenido y de una combinación de ambas características, para hacer una clasificación.

Los resultados obtenidos en los experimentos de clasificación automática muestran que los enfoques computacionales representan una solución viable para la tarea de reconocimiento de humor, que es posible utilizar estas técnicas para distinguir entre textos humorísticos y no humorísticos, y que se alcanza buen desempeño usando técnicas de clasificación basadas en características de contenido y de estilo.

Así mismo, Mihalcea y Pulman en [7] analizan de forma detallada otras dos características que, según teorías psicológicas, están presentes de manera muy frecuente en el humor: centrado en el humano y la polaridad negativa.

Pero uno de los trabajos más innovadores es el de Rada Mihalcea *et al*. [5], [8], en donde exploran diversos modelos computacionales para la resolución de incongruencias, que es una de las teorías del humor más ampliamente aceptadas; la cual sugiere que el humor se debe a la mezcla de dos cuadros opuestos de interpretación posible para un enunciado.

Hasta el momento se han obtenido resultados significativos en el desarrollo de herramientas computacionales para el idioma inglés, sin embargo para el idioma español, el avance ha sido muy lento. En el presente trabajo solamente se intenta detectar si en ciertos tipos de mensajes en redes sociales se utiliza el doble sentido, para en una etapa posterior desarrollar modelos de clasificación.

3. Confección del corpus

Se formó un corpus de 548,243 tweets a través de un programa realizado en Python y usando el API del Twitter. El proceso de generación del corpus consideró tomar todos aquellos tweets publicados durante un cierto dia y que estuviesen localizados en un

[2] http://www.cs.waikato.ac.nz/ml/weka/

radio no mayor a 10 kilómetros de la capital de cada estado de la república mexicana (geolocalizados). Un análisis del corpus indica que se tiene un promedio de 14.340 palabras por tweet, con un vocabulario total de 581,109 palabras y en el cual interactuaron 173,339 usuarios con un promedio de 3.167 tweets por usuario, esto se puede ver en la Tabla 1.

Tabla 1. Información del corpus de tweets por estado y total

Estado	Tweets	Mínimo de palabras	Máximo de palabras	Promedio de palabras	Vocabulario	Tokens
AGU	14,426	1	32	11.452	23,351	165,200
BCN	13,644	1	33	11.064	23,563	150,961
BCS	11,145	1	30	10.897	18,813	121,443
CAM	15,496	1	36	11.201	22,671	173,569
CHH	12,652	1	39	11.576	21,293	146,458
CHP	12,440	1	31	10.980	19,656	136,590
COA	13,581	1	34	11.667	21,898	158,445
COL	14,498	1	32	12.281	24,728	178,054
DIF	13,452	1	32	11.857	26,305	159,500
DUR	15,837	1	33	11.643	24,401	184,397
GRO	11,012	1	32	10.748	18,502	118,353
GUA	15,021	1	32	11.136	24,385	167,271
HID	14,304	1	33	10.874	21,349	155,546
JAL	14,201	1	31	11.693	25,394	166,056
MEX	13,208	1	36	11.767	25,311	155,413
MIC	14,643	1	31	11.186	23,104	163,797
MOR	12,288	1	32	11.439	21,242	140,560
NAY	12,892	1	33	11.932	22,854	153,828
NLE	14,354	1	33	10.455	22,999	150,072
OAX	11,159	1	36	12.333	19,221	137,623
PUE	13,415	1	33	11.092	21,143	148,797
QRO	16,274	1	32	10.372	23,390	168,798
ROO	14,369	1	33	11.974	21,975	172,055
SIN	16,169	1	32	10.814	23,518	174,858
SLP	16,473	1	32	11.676	26,194	192,341
SON	12,856	1	31	9.210	17,653	118,405
TAB	15,459	1	31	9.969	21,906	154,117
TAM	16,183	1	36	10.156	21,567	164,360
TLA	12,225	1	33	11.168	20,478	136,525
YUC	12,976	1	31	10.924	22,326	141,755
ZAC	15,394	1	32	11.798	24,138	181,614
TOTAL	548,243	1	39	12.340	581,109	6,765,487

A partir del diccionario de mexicanismos, el cual viene con una clasificación de uso, se tomaron los que tienen la categoría de VULG (vulgaridades) y OBSC (obscenidades), refiriéndose a palabras vulgares y obscenas respectivamente, creando así 2

diccionarios, uno de vulgaridades y el otro de obscenidades, con 409 y 361 frases respectivamente. Ejemplos de ambos diccionarios se muestran en la Tabla 2.

Tabla 2. Ejemplos de obscenidades y vulgaridades

Obscenidades	@Maracucho_Serio tweeted: RT @Manuelovsky: Dejarse **coger** no duele, lo que duele es enamorarse. @sswwar tweeted: Upsss **se me para** la verga http://t.co/I7WCUifjaO @boulalaAlex tweeted: Dejarme el bigote es lo más cercano a una **panocha** que he tenido en la cara. @Wilherreraleon tweeted: RT @weezysanchez: Vine a **calzonear** a las embajadoras!!! Raroooooooooo @ChicoCheVive @jorgedelacruz99 tweeted: "@Javii_Noble: Una Princesa bien amada, bien valorada y bien **cogida**, jamás te será infiel. #YoDigo #JaviTip"
Vulgaridades	@ChechoV tweeted: Ya saquen la caguama ps que está **de la chingada** el calor @lVICOCENTENO tweeted: Cuando es amor de verdad, no hay ni puta, ni **pendejo** que los separe... @raulcamangulo tweeted: Que gustarles que las anden manoseando entre el **chingamadral** de gente en el carnaval @iJosee_ tweeted: Pues dicen que no te sirve de nada estar **mamado** si estas feo... #NomasTeInformo @patycarretell0 tweeted: Me tienen hasta la madre con sus selfies

4. Análisis del corpus desarrollado

Se realizó un análisis estadístico sobre los tweets y en base a los diccionarios de vulgaridades y obscenidades se obtuvieron los siguientes resultados: Se realiza análisis por estado para revisar dentro de los tweets que estados tienen dentro de su vocabulario mayor cantidad de vulgaridades, resultando así el estado de Querétaro (QRO), seguido de Tabasco (TAB) y Nuevo León (NLE), y en último lugar se posiciona Sinaloa (SIN), como se puede observar en la Tabla 3, además en la tercer columna de la Tabla 4 se observa que el estado que resulta con mayor ocurrencias, es uno de los que utilizan las palabras con mayor frecuencia.

Tabla 3. Frecuencia de ocurrencia de vulgaridades por estado

Estado	Ocurrencias	Estado	Ocurrencias
QRO	408	MOR	247
TAB	350	MIC	240
NLE	347	ZAC	230
PUE	345	MEX	225
HID	325	YUC	224
TAM	319	GRO	220

CHP	311	DF	217
VER	306	BCN	211
GUA	305	ROO	203
AGU	305	COA	182
CHH	303	COL	181
TLA	284	NAY	179
CAM	275	OAX	171
SON	264	BCS	167
SLP	260	DUR	166
JAL	254	SIN	102

Además se pueden obtener las cinco primeras palabras o frases vulgares que dentro de todos los tweets obtuvieron mayor ocurrencia, las cuales se pueden mostrar en la Fig. 1, y que fueron tomadas a partir de la Tabla 4, donde se muestra los estados que más utilizan dicha palabra o frase.

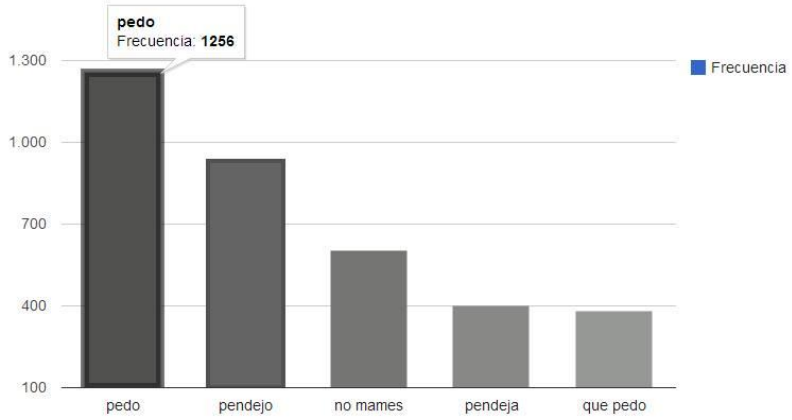

Fig. 1. Palabras o frases vulgares con mayor frecuencia.

Tabla 4. Palabras vulgares con mayor ocurrencia en total.

Frase	Ocurrencias	Estados
pedo	1256	**QRO**, TAB, PUE
pendejo	935	**QRO**, CHH, NLE
no mames	601	HID, TAB, TAM
pendeja	401	CAM, ZAC, HID
que pedo	382	PUE, **QRO**, CHH

Posteriormente, se realiza un análisis para cada estado, con el objetivo de obtener dentro de los tweets que estados manejan dentro de su vocabulario mayor cantidad de obscenidades, en las tres primeras posiciones se encuentra Chihuahua (CHH), Zacatecas (ZAC) y Guanajuato (GUA), respectivamente, y en último lugar se posiciona Sonora (SON).

Se obtienen las cinco primeras palabras o frases obscenas que dentro de todos los tweets obtuvieron mayor ocurrencia las cuales se muestran en la Fig. 2, éstas fueron tomadas a partir de la Tabla 5 y donde se pueden observar los estados que más utilizan dicha palabra o frase.

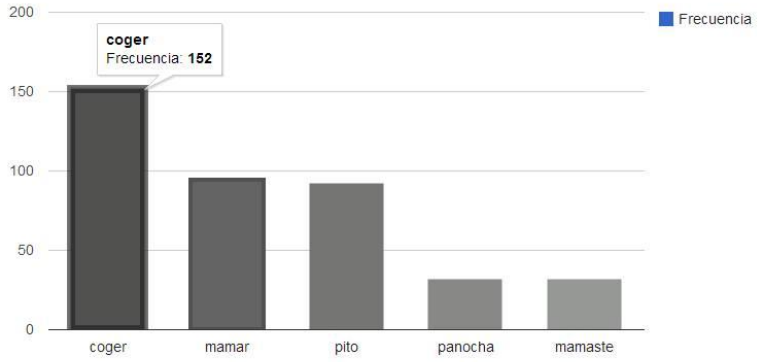

Fig. 2. Palabras o frases obscenidades con mayor frecuencia.

Tabla 5. Palabras obscenas con mayor ocurrencia en total.

Frase	Ocurrencias	Estados
coger	152	MEX, **CHH**, JAL
mamar	95	DF, QRO, VER
pito	92	MIC, **CHH**, NAY
panocha	32	**CHH**, GUA, COL
mamaste	32	**CHH**, TAM,COA

5. Modelo de clasificación

El análisis del corpus y la detección de las palabras obscenas y vulgares más utilizadas en las diferentes regiones, ayudó a categorizar los tweets que forman parte del corpus de entrenamiento. Es decir si el tweet contiene palabras obscenas, dicha frase es etiquetada como **obscenidad**, si el tweet contiene palabras vulgares, se etiqueta como **vulgaridad**, en el caso del que tweet contenga tanto palabras obscenas como vulgares se etiqueta como **mezclado**, y si no contiene ninguna palabra obscena o vulgar se etiqueta como **ninguno**. El conjunto de datos de entrenamiento quedó completamente desbalanceado.

Se realizó un proceso de balanceo quedando la misma cantidad de tweets correspondiente a la clase menos representativa en dicho conjunto. La selección de los tweets que forman parte del conjunto de entrenamiento se realizó de manera aleatoria, quedando al final 130 tweet por cada categoría.

Una vez desarrollado el corpus de entrenamiento, se utilizaron para el desarrollo del modelo de clasificación todas las palabras que aparecen en cada tweets. No se aplicó ningún preprocesamiento al corpus de entrenamiento construido.

Para el desarrollo de los modelos de clasificación se utilizó la herramienta Weka. En particular los algoritmos de clasificación empleados fueron: Vecino más cercano (IBK), Máquina de soporte vectorial (SMO), Bayes Multinomial y Árbol de decisión (J48).

En la Tabla 6 se pueden ver los resultados obtenidos, aplicando validación cruzada con 10 pliegues. Como se puede observar el algoritmo J48 es el que nos brinda mejores resultados con un 76.35%.

Tabla 6. Clasificación de tweets con cuatro clases

Algoritmo	**Correctamente clasificadas**	**Incorrectamente clasificadas**
IBK	50.77%	49.23%
SMO	75%	25%
Bayes multinomial	69.81%	30.19%
J48	**76.35%**	23.65%

Como el número de muestras positivas de cada clase es pequeño, se puede apreciar que los resultados en el modelo de clasificación no fueron muy altos, por este motivo se decidió desarrollar entonces dos corpus de entrenamiento, considerando solamente dos categorías: Obscenidad y Ninguna y Vulgaridad y Ninguna.

Las Tablas 7 y 8, muestran los resultados obtenidos aplicando validación cruzada con 10 pliegues.

Tabla 7. Clasificación con las clases Obscenidad y Ninguna

Algoritmo	**Correctamente clasificadas**	**Incorrectamente clasificadas**
IBK	70.53%	29.47%
SMO	**91.07%**	8.93%
Bayes multinomial	85.14%	14.86%
J48	84.28%	15.72%

Tabla 8. Clasificación con las clases Vulgaridad y Ninguna

Algoritmo	**Correctamente clasificadas**	**Incorrectamente clasificadas**
IBK	75.04%	24.96%
SMO	**98.90%**	1.10%
Bayes multinomial	88.54%	11.46%
J48	97.46%	2.54%

Como puede apreciarse los mejores resultados fueron ofrecidos por el algoritmo de clasificación máquina de soporte vectorial, con los datos por defecto que ofrece la herramienta Weka.

El hecho de que haya mejorado la precisión del modelo de clasificación se debe en gran medida al número de muestras positivas en cada clase, y por el hecho de que es más simple para el algoritmo de clasificación cuando se trabaja con 2 clases, que con 4 clases.

6. Conclusiones

En este trabajo se presenta una primera aproximación para la identificación de frases obscenas y vulgares en mensajes de twitter. Se puede observar que los mensajes de este tipo se emplean más en los estados de: Guanajuato, Estado de México y Jalisco para obscenidades y en el caso de vulgaridades los estados de la republica detectados que usan este tipo de frases son: Querétaro, Puebla e Hidalgo.

Desarrollando los diccionarios de vulgaridades y obscenidades se pudo de alguna manera construir un corpus de entrenamiento para posteriormente desarrollar un modelo de clasificación. A pesar de que el corpus no quedó en primera instancia balanceado, y el proceso de balanceo fue muy simple, los resultados son relativamente alentadores. En una segunda etapa se va a realizar un proceso de balanceo más adecuado.

En estos momentos se está preparando un corpus de prueba, para validar realmente el modelo de clasificación construido.

Como ha sido la primera aproximación, no se ha realizado ningún pre procesamiento a los tweets, aspecto que se está considerando, buscando otro modelo de representación de los mismos, ya que solamente se utilizó bolsa de palabras.

El corpus y el modelo de clasificación desarrollado es nuestro primer acercamiento para intentar educar a las nuevas generaciones en el no empleo desmedido de esta forma de comunicación, así mismo puede servir de apoyo en los sitios web, para que no permitan comentarios de usuarios sin previa revisión.

Referencias

1. Mihalcea, R., Attardo, S.: Making Computers Laugh. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP), pages 531–538, Vancouver, October 2005.
2. Ocampo, R.: Detección automática de humor en textos cortos en español. *Tesis*: Instituto Politécnico Nacional (2010).
3. Hernández, V.: Antología del Albur. *BookSurge Publishing*, Charleston SC. (2006).
4. Mihalcea, R., Strapparava, C.: Making Computers Laugh: Investigations in Automatic Humor Recognition. In *Human Language Technology Conference and Con-*

ference on Empirical Methods in Natural Language Processing. Vancouver, October 2005. Association for Computational Linguistics. 531–538, (2005).
5. Mihalcea, R., Strapparava, C.: Laughter Abounds in the Mouths of Computers: Investigations in Automatic Humor Recognition. In INTETAIN 2005, LNAI 3814, pp. 84–93, Springer-Verlag Berlin Heidelberg (2005).
6. Mihalcea, R., Strapparava, C.: Learning to Laugh (Automatically): Computational Models for Humor Recognition. Computational Intelligence. 22. 126-142 (2006).
7. Mihalcea, R., Pulman, S.: Characterizing Humor: An exploration of Features in Humorous Texts. : CICLing 2007, LNCS 4394, pp. 337–347. Springer-Verlag Berlin Heidelberg, (2007).
8. Mihalcea, R., Strapparava, C., Pulman, S.: Computational Models for Incongruity Detection in Humour. : CICLing 2010, LNCS 6008, pp. 364–374. Springer-Verlag Berlin Heidelberg (2010).
9. García Torres Carmen C.: El doble sentido en la conversación de llamadas telefónicas a programa de radio: análisis semiótico-discursivo. Proyecto de Tesis Universidad Autónoma de Nuevo León (2010).
10. Espinosa Luz: El albur: el mete y saca de la palabra. http://culturacolectiva.com/el-albur-el-mete-y-saca-de-la-palabra/ (visitada: 25/10/2014)