Prediction of Aggressive Comments in Social Media: an Exploratory Study

L. P. D. Bosque and S. E. Garza

***Abstract*— This paper presents a set of techniques for predicting aggressive comments in social media. In a time when *cyberbullying* has, unfortunately, made its entrance into society and Internet, it becomes necessary to find ways for preventing and overcoming this phenomenon. One of these concerns the use of machine learning techniques for automatically detecting cases of cyberbullying; a primary task within this cyberbullying detection consists of *aggressive text detection*. We concretely explore different computational techniques for carrying out this task, either as a classification or as a regression problem, and our results suggest that a key feature is the identification of profane words.**

***Keywords*— aggressiveness, social media, prediction, artificial neural networks, support vector machines.**

I. INTRODUCCIÓN

UNO de los fenómenos que hoy en día afecta gravemente a la sociedad es el llamado *acoso* o "bullying", es decir, el abuso repetitivo y dañino que se da en y por jóvenes de edad escolar [1]. Una modalidad especialmente nociva de este fenómeno es el "cyber-bullying" o *ciberacoso*, el cual consiste en hostigar a la víctima a través de las tecnologías de información [2]. A diferencia del acoso tradicional, el ciberacoso es más violento [3], pues la víctima no es solamente hostigada dentro de la escuela y en un horario específico, sino que puede serlo en cualquier parte y horario; así mismo, ya que la tecnología ayuda a esconder la verdadera identidad del agresor, los ataques son más pronunciados y más sencillos de hacer [4]. Desafortunadamente, para las víctimas, el ciberacoso puede incluso terminar en suicidio [5].

Como primer paso para atacar este problema, proponemos la *predicción (clasificación y regresión) automática de texto agresivo en redes sociales*. Para ello, exploramos diferentes técnicas que puedan cumplir con esta tarea—entre ellas están las redes neuronales, las máquinas de soporte vectorial, los sistemas difusos y la regresión lineal. Nuestro objetivo consiste, entonces, en brindar una perspectiva sobre la detección de agresividad en la comunicación interpersonal en los medios sociales.

Para comparar el conjunto de técnicas propuestas, hemos extraído comentarios de la popular red social Twitter, la cual alberga opiniones públicas y libres. Como veremos más adelante, aquellas técnicas que juntan varias características son las que toman la delantera en la predicción. De igual forma, una característica clave parece ser la cantidad de palabras ofensivas por comentario.

El presente documento está organizado del modo siguiente: la Sección II presenta nuestra metodología, la Sección III muestra experimentos y resultados, la Sección IV presenta trabajos relacionados y, finalmente, la Sección V ofrece conclusiones y trabajo a futuro.

II. METODOLOGÍA

En primer lugar, conviene definir el término "texto agresivo", pues de aquí parte todo nuestro trabajo. Consideramos como *texto agresivo* cualquier *comentario* (donde el comentario es una serie de palabras) que tiene la intención de *ofender* a una persona o grupo de personas. En segundo lugar, conviene definir el término "predicción"; este se refiere a la asociación de un conjunto de datos u observaciones a una variable, de tal manera que cuando se proveen nuevos datos (no observados con anterioridad), se es capaz de establecer la asociación a esta variable de manera correcta. Cuando la variable es nominal o discreta, a la tarea de predicción se le llama *clasificación*, y cuando la variable es continua, se le conoce como *regresión*. Dicho en otras palabras, la tarea de clasificación consiste en asociar un dato a una clase o categoría; la tarea de regresión, por otro lado, asocia un dato a un valor numérico. En este trabajo, exploramos ambas tareas.

Para predecir texto agresivo, proponemos varios enfoques, a saber: enfoques basados en léxicos, sistemas difusos, enfoques supervisados y otros enfoques estadísticos. Algunos de estos enfoques son ya tradicionales para resolver ambos tipos de problemas (clasificación y regresión).

Antes de describir cada enfoque, conviene mencionar que utilizamos una escala [0, 10] para la tarea de regresión; esto significa que los enfoques a utilizar arrojarán un número dentro de este rango para indicar el nivel de agresividad en un comentario. Escogimos esta escala (donde 0 es no agresivo y 10 es muy agresivo), ya que el rango seleccionado no es ni muy apretado ni muy holgado, y permite manejar varios niveles. Ahora bien, para la tarea de clasificación, hemos de trabajar—como típicamente se hace—con dos clases (clasificación binaria): *agresivo (A)* y *no agresivo (N)*. En ese sentido, los enfoques para hacer clasificación determinarán si un comentario dado es agresivo o no.

A. Enfoques basados en léxicos

Un léxico (o *lexicón*) puede definirse como un listado de palabras. Por tanto, los enfoques basados en léxicos reconocen palabras con características especiales; cabe destacar que es común recurrir a estos enfoques cuando se realiza *análisis de sentimiento* (aquí estamos considerando que la predicción de agresividad es una sub-tarea de esta disciplina). En nuestro

L. P. D. Bosque, Universidad Autónoma de Nuevo León, Nuevo León, México, laura.delbosquevg@uanl.edu.mx

S. E. Garza, Universidad Autónoma de Nuevo León, Nuevo León, México, sara.garzavl@uanl.edu.mx

Corresponding author: Sara Elena Garza Villarreal

caso, proponemos tres enfoques de este tipo: *puntaje NS*, *puntaje ANEW* y *puntaje SentiWordNet*.

Puntaje NS. El sitio `noswearing.com` publica (y actualiza constantemente) un extenso listado de palabras altisonantes, junto con sus definiciones y algunas variantes. Ya que este sitio, además de contar con este listado, acepta contribuciones de cualquier persona, nos parece un recurso apropiado para detectar agresividad en comentarios—esto suponiendo que un acto de agresión verbal está marcado por el uso de malas palabras. Para detectar el nivel de agresión utilizando el sitio de noswearing.com (al cual abreviaremos como NS), básicamente lo que hacemos es calcular la *frecuencia relativa normalizada* de palabras ofensivas en el comentario. Para calcular dicha frecuencia, primero obtenemos la frecuencia relativa f_i de palabras ofensivas en el comentario c_i :

$$f_i = \frac{o_i}{n_i}, \tag{1}$$

donde o_i es la cantidad de malas palabras y n_i es la cantidad total de palabras en el comentario c_i. Esta frecuencia se normaliza para mitigar el efecto que podrían tener comentarios muy largos o muy cortos. El puntaje final se calcula como

$$ns_i = (10)\frac{f_i}{f_{max}}, \tag{2}$$

donde ns_i es el puntaje y $f_{\max}$ es la máxima frecuencia encontrada en el repositorio de comentarios. Por ejemplo, supongamos que tenemos un comentario $c_a=\{w_1,w_2,w_3\}$ y w_1 es una mala palabra. Entonces f_i=1/3=0.33, y si $f_{\max}$=0.5 entonces sc_i=(10)(0.33/0.5)=6.6. Note que el rango de este puntaje cae dentro de nuestra escala [0, 10] para regresión. No obstante, también es posible producir una clase con este puntaje (en general, es posible hacer esto para cualquier método que produzca un valor numérico); simplemente utilizamos una función escalonada con umbral mayor o igual a 3. Esto significa que cualquier puntaje 0-3 se mapea a la clase *no agresivo* (N) y cualquier puntaje 4-10 se mapea a la clase *agresivo* (A).

Puntaje ANEW. Este lexicón, cuyas siglas significan "Affective Norms for English Words", es un recurso que contiene palabras con indicadores de felicidad [6]; puesto que consideramos que los comentarios felices y agresivos son mutuamente excluyentes, vemos como conveniente el uso de este lexicón. ANEW consiste de 1,034 términos e incluye, para cada uno de ellos, tres diferenciales (escalas cuyos extremos representan adjetivos opuestos):

- Valencia psicológica (bueno-malo)
- Motivación (pasivo-activo)
- Dominio (débil-fuerte)

Considerando estos tres diferenciales, el valor de felicidad de cada palabra cae en el rango [1, 9], donde 9 es lo más cercano a felicidad. Puesto que entre mayor es el puntaje, más positiva resulta ser la palabra (y por ende el comentario), es necesario invertir este valor; de igual manera, es necesario que el puntaje quede dentro de nuestra escala (0-10). Es por esto que hacemos el siguiente cálculo:

$$a_i = \frac{(10)[(9 - v_i) - 1]}{8}, \tag{3}$$

donde v_i es el valor promedio de felicidad para c_i.

Puntaje SentiWordNet (SWN). WordNet es una base de conceptos y relaciones que ha sido utilizada ya en un sinfín de trabajos [15]; dicha base se compone de grupos de sinónimos (llamados "synsets") que representan un mismo concepto y que se encuentran ligados a otros grupos a través de relaciones como "es un" (*hiperónimo*) o "es parte de" (*merónimo*). Una variante importante que ha tomado fuerza en los últimos años es *SentiWordNet* [7], pues esta base incorpora aspectos de análisis de sentimiento; como ya hemos mencionado, consideramos que esta tarea subsume la predicción de texto agresivo. Más específicamente, SentiWordNet es una base cuyos synsets tienen valores de polaridad, es decir, valores que indican qué tan positivo o negativo es el synset. En ese sentido, son tres los valores asociados a un synset: Pos(s), Neg(s) y Obj(s). Estos valores representan, respectivamente, el grado en que el synset es positivo, negativo y/o neutro. Cada uno de estos valores cae en el rango de [0,1], y la suma de los tres siempre es 1.0. Por tanto, cuando menos uno de estos valores debe ser mayor a 0.

Para calcular el nivel de agresividad de un comentario utilizando SentiWordNet, asociamos cada palabra del comentario a un synset y obtenemos el promedio de los valores Neg(s) de los synsets resultantes (nótese que este promedio se multiplica por diez para hacerlo caer dentro del rango de nuestra escala). Por ejemplo, supongamos que se tiene el comentario $c_i = \{w_1, w_2\}$, y que w_1 se relaciona con el synset s_1 y w_2 con el synset s_2. Si Neg(s_1) = 0.8 y Neg(s_2) = 0.4, entonces el puntaje SentiWordNet para c_2 es (10)(0.8 + 0.4)/2 = 6.

Es importante destacar que una misma palabra podría estar asociada a más de un synset (un tipo de ambigüedad conocida como "polisemia"); por simplicidad, hemos decidido ignorar los puntajes de las palabras que caen en este caso. El procesamiento de estas palabras se deja como trabajo a futuro.

B. Sistemas difusos

Un *sistema difuso* es un tipo de sistema experto que es capaz de trabajar con datos imprecisos o vagos y que, como su nombre lo indica, está basado en lógica difusa [8]; los sistemas difusos son utilizados con frecuencia para tareas de predicción. Este tipo de sistemas produce salidas mediante el uso de un motor de inferencia con reglas difusas. Para poder hacer inferencias con esta base de reglas, las entradas al sistema primero se *fusifican*, donde este proceso consiste en relacionar cada entrada con uno o más *conjuntos difusos*; para obtener el resultado final, las salidas se *defusifican,* donde este proceso consiste en calcular una salida numérica a partir de conjuntos difusos. En cuanto a las reglas difusas, estas tienen una estructura "SI-ENTONCES" que hace uso de *variables lingüísticas*, donde una variable lingüística es una variable que

tiene como valores conjuntos difusos y que está asociada a una variable numérica *x*. Por su parte, un *conjunto difuso* es un conjunto que admite presencias parciales de sus elementos. Esto significa que el grado de membersía de los elementos en este tipo de conjuntos cae en el rango [0,1]. Este grado de membresía es definido por una *función de membresía*. Para mayor información sobre lógica y sistemas difusos, recomendamos el libro de Wang [9].

Utilizando las definiciones anteriores, es posible describir nuestro sistema difuso para la predicción de texto agresivo. Vale la pena mencionar que el diseño que a continuación explicamos corresponde al diseño que, hasta el momento, ha brindado los mejores resultados para nuestra tarea. Este diseño recibe dos entradas: la cantidad de palabras del comentario y la cantidad de malas palabras en el mismo. En cuanto a la salida, es un número dentro del rango [0,1], que más tarde se convierte a nuestra escala. El sistema cuenta con tres variables lingüísticas, donde cada variable cuenta con cinco valores o particiones (es decir, está asociada con cinco conjuntos difusos). Todos los conjuntos difusos están representados con funciones de membresía triangulares, cuyos parámetros (inicio, pico, fin) están dados por la media y la desviación estándar del conjunto de datos utilizado. Mientras que esta decisión implica que estas funciones deban redefinirse cada vez que se cambia el conjunto de datos (lo cual es hasta cierto punto razonable), consideramos que esto es mejor a definir arbitrariamente las funciones.

Con respecto a la base de conocimiento, esta consiste de 25 reglas extraídas a partir de nuestra experiencia en el tema (algunos ejemplos se listan en la Tabla I). El método de defusificación utilizado es el método del centroide.

Debido a que el sistema difuso arroja como resultado un valor dentro del rango [0,1], es necesario multiplicarlo por diez para que quede dentro de nuestra escala. De igual manera, para derivar la clase en la tarea de clasificación, seguimos utilizando la función escalonada con umbral en tres.

TABLA I
EJEMPLOS DE REGLAS DIFUSAS PARA NUESTRO SISTEMA

1. SI el documento es *muy corto* y tiene *algunas* malas palabras, ENTONCES el documento es *positivo*.
2. SI el documento es *largo* y contiene *muchas* malas palabras, ENTONCES el documento es *muy agresivo*.

C. Métodos supervisados

Un método de aprendizaje supervisado es aquel que requiere como entrada una serie de *ejemplos etiquetados* para derivar un *modelo predictivo*. Cada ejemplo consiste en un *vector de características*, y la etiqueta consiste en la clase (o valor numérico) correcto para el ejemplo.

Las tres técnicas supervisadas que hemos considerado son las *redes neuronales (RNA)*, las *máquinas de soporte vectorial (SVM)*, el *clasificador bayesiano* y el *árbol J48*. Las redes neuronales, por su parte, emulan el comportamiento de las neuronas en nuestro cerebro; mientras tanto, las máquinas de soporte vectorial se basan en agregar dimensiones a los elementos para lograr su correcta clasificación, y el clasificador bayesiano utiliza como fundamento el Teorema de Bayes con independencia condicional. [10]. El árbol J48, por último, corresponde a un árbol de decisión basado en ganancia de información.

En nuestro caso, consideramos dos tipos de vectores: de metadatos y de palabras. Con respecto a los vectores de palabras, cada comentario se representa como una serie de pesos donde cada peso indica la importancia de la palabra en el comentario (esta importancia puede calcularse utilizando frecuencia, TF-IDF y métricas similares). En cuanto a los vectores de metadatos, utilizamos seis características:

1. Cantidad de palabras en el comentario
2. Cantidad de palabras ofensivas (maldiciones)
3. Frecuencia de la palabra “tú”
4. Puntaje NS
5. Puntaje ANEW
6. Puntaje SentiWordNet

En cuanto a las primeras dos características, se puede considerar que son básicas si asumimos que la densidad de malas palabras es un indicador de agresión verbal. Ahora bien, con respecto a la frecuencia de la palabra “tú”, se ha observado que en idiomas como el inglés es común utilizar la segunda persona para ofender a alguien (“*you* _#@%”) [11], y por ende es un factor a considerar. Con respecto a las últimas tres características, no descartamos que la combinación de estos puntajes pudiera apuntar en la dirección correcta.

D. Otros métodos estadísticos

Un método particularmente útil para el tipo de tareas que deseamos abordar es la regresión lineal, la cual genera un modelo lineal (de ahí el nombre) que se ajusta a los datos proporcionados. Las características de estos datos son las mismas que las utilizadas para los métodos supervisados anteriormente descritos.

E. Resumen

En resumen, nos encontramos haciendo una exploración de técnicas para obtener una perspectiva inicial sobre la predicción de texto agresivo en redes sociales. La tarea de predicción incluye dos sub-tareas: clasificación y regresión. Algunas de las técnicas propuestas pueden llevar a cabo ambas tareas y otras solamente una; la Tabla II indica esta información.

TABLA II
TAREAS QUE REALIZAN LAS TÉCNICAS SELECCIONADAS

	Clasificación	**Regresión**
Puntaje NS	X	X
Puntaje ANEW	X	X
Puntaje SWN	X	X
RNA	X	X
Clasificador bayesiano	X	
J48	X	
SVM	X	
Sistema difuso	X	X
Regresión lineal		X

III. EXPERIMENTOS Y RESULTADOS

Como ya se ha indicado anteriormente, nuestro propósito consiste en brindar una primera perspectiva sobre la detección de texto agresivo; para ello, exploramos diferentes técnicas y diferentes conjuntos de características a utilizar. Algunas preguntas a las cuales pretendemos dar respuesta con esta exploración son las siguientes: "*¿Qué técnica parece prometedora?*", "*¿Qué representación del comentario es más útil, la de metadatos o la de vectores de palabras?*", "*¿Qué características parecen ser las más prometedoras?*", "*¿Es mejor plantear el problema como regresión o como clasificación?*".

A. Configuración experimental

Nuestro repositorio de comentarios fue extraído de la popular red social *Twitter*, la cual permite a los usuarios publicar comentarios cortos (llamados "tuits") y recibir ("seguir") las publicaciones de otros usuarios; de igual manera, los usuarios pueden enviarse mensajes, ya sean públicos o privados (a este tipo de mensajes se le conoce como "mensajes directos" o "comentarios direccionados"). Escogimos esta red social debido a que los usuarios pueden publicar lo que deseen y esto, desgraciadamente, puede dar pie a malos comportamientos. Cabe destacar que el repositorio descargado está en inglés, pues este idioma requiere menos pre-procesamiento y algunos de los lexicones utilizados solamente existen para este lenguaje.

El repositorio fue descargado a través de buscar la palabra "escuela" en los tuits, pues el ambiente escolar suele ser (desafortunadamente) propicio para que se den casos de acoso y ciberacoso. Aunque el repositorio contó inicialmente con 111,381 comentarios, si consideramos que un caso de ciberacoso se dará con comentarios *direccionados* o dirigidos a un usuario en particular (en Twitter, estos comentarios anteponen el nombre de usuario–que empieza con "@"—al resto del texto), esta base se reduce drásticamente. El conjunto de comentarios direccionados, adicionalmente, fue fraccionado en dos conjuntos de datos: el conjunto F y el conjunto B. En ambos conjuntos existe la presencia de palabras ambiguas (en inglés), donde uno de los significados es ofensivo ("f*ck" para el conjunto F y "b*tch" para el conjunto B). Hemos escogido estas palabras para tratar de asegurar que se contaría con casos positivos (agresivos) y negativos (por la misma ambigüedad).

Para poder evaluar todos los métodos propuestos, fue necesario hacer una clasificación manual de los tuits del repositorio; para ello, se ocupó un grupo de personas. Dicho grupo asignó a cada comentario del conjunto un valor de agresividad dentro de nuestra escala (recordemos que a partir del valor también es posible asignar una clase). Para poder saber si la clasificación de estas personas era confiable y coherente entre una y otra, se utilizó ANOVA. Aquellos tuits con clasificaciones disparejas fueron finalmente eliminados, dejando así un total de 243 tuits utilizables para la predicción automática.

Antes de dar los tuits como entrada a cada uno de los métodos de predicción, se procedió a hacer algunas rutinas de pre-procesamiento. En ese sentido, se eliminaron signos de puntuación, se cambió todo el texto a minúsculas, se expandieron algunas frases representadas con siglas (por ejemplo, "wtf" se cambió por "what the f*ck") y se cambiaron por texto algunos emoticones (por ejemplo, "=)" se sustituyó por "feliz").

En cuanto a la implementación de las técnicas, los puntajes NS, ANEW y SentiWordNet fueron extraídos con scripts en el lenguaje Python, y para el sistema difuso se utilizó el software QtFuzzyLite (disponible en http://www.fuzzylite.com). Ahora bien, para los métodos supervisados y estadísticos, utilizamos la implementación del software WEKA (disponible en http://www.cs.waikato.ac.nz/ml/weka/) con la configuración predeterminada; las redes neuronales, específicamente, fueron representadas por el perceptrón multicapa que incluye el software.

También conviene destacar que para los métodos supervisados y estadísticos—tanto en clasificación como en regresión—se evaluaron distintas combinaciones de características (representación de metadatos):

- Todas las características (seis en total)
- Todas las características excepto la 3 (cinco)
- Características 1, 2 y 4 (tres)
- Características 1, 2 y 3 (tres)
- Características 1 y 2 (dos)

Estas combinaciones buscaron evaluar los comentarios utilizando ya fuera las características que se consideraban las más fuertes o las más básicas. En cuanto a los vectores de palabras (TF-IDF), estos solamente se utilizaron para las técnicas supervisadas de clasificación.

Para evaluar los resultados de clasificación, se tomó en cuenta el porcentaje promedio de aciertos de cada técnica; para evaluar los resultados de regresión, obtuvimos el *Error Medio Cuadrado* (MSE en inglés), el cual se calcula como $(x\text{-}y)^2$, donde x es el resultado obtenido por la técnica y y es el resultado correcto.

Por último, adicional a las técnicas propuestas, se incluyó como base de línea un método aleatorio; para clasificación, este consistió en asignar la clase al azar y, para regresión, consistió en asignar un número aleatorio dentro de nuestra escala [0,10].

B. Resultados

Los resultados para la predicción de comentarios agresivos como un problema de clasificación se muestran en las Figs. 1 y 2; mientras que la primera corresponde a la utilización de vectores de metadatos, la segunda corresponde a la utilización de vectores TF-IDF de palabras.

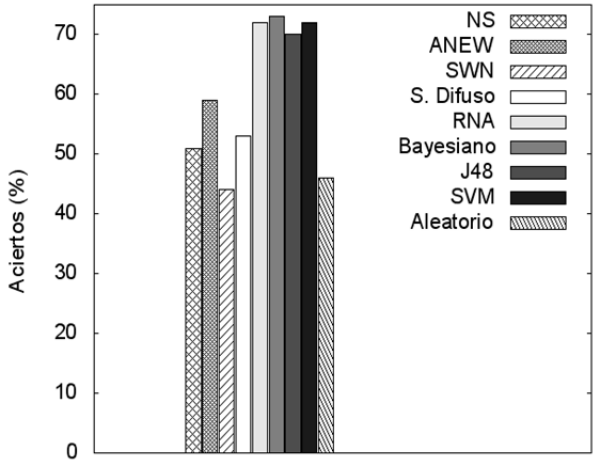

Figura 1. Porcentaje de aciertos para las técnicas propuestas de clasificación usando vectores de metadatos.

Como podemos ver en la Fig. 1, el porcentaje más alto de aciertos fue obtenido por el clasificador bayesiano con cinco atributos (73.25%); cabe destacar que este porcentaje también corresponde al mayor número de aciertos para la clase agresiva (A), la cual es de nuestro mayor interés por tratarse de la clase positiva. En promedio, el clasificador bayesiano también obtuvo el primer lugar (70.94%), seguido por la SVM (70.5%) y las redes neuronales (70.43%). En ese sentido, estas tres técnicas podrían ser utilizadas para la tarea de clasificación con metadatos.

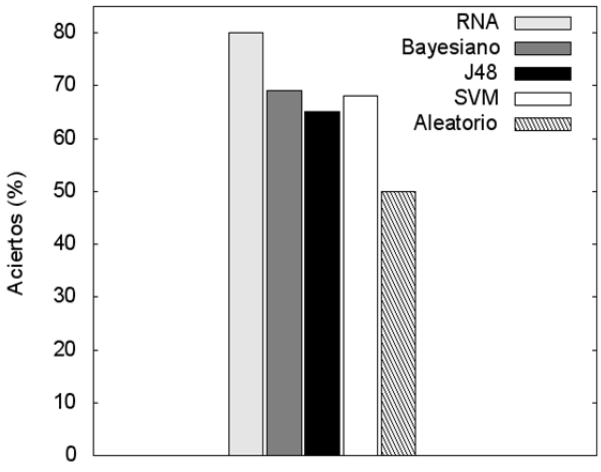

Figura 2. Porcentaje de aciertos para las técnicas supervisadas de clasificación usando vectores TF-IDF de palabras.

Como se aprecia en la Fig. 2, la técnica que obtiene el mayor porcentaje de aciertos con vectores TF-IDF de palabras corresponde a las redes neuronales (80.89%); le siguen el clasificador bayesiano (69.03%) y la SVM (68%). Como podemos ver, las tres técnicas antes mencionadas ocupan los primeros sitios tanto para una representación como para la otra. Otro dato interesante es que el tiempo de entrenamiento para las redes neuronales fue ligeramente mayor, pues fueron aproximadamente 90 segundos en una laptop con procesador AMD de 2.1 GHz contra un segundo que tardaban el resto de las técnicas en esta misma computadora. Por tanto, para conjuntos de datos más grandes estos tiempos podrían alargarse.

Los resultados para la tarea de regresión se muestran en la Fig. 3. Como se puede apreciar en esta figura, el menor error lo obtiene la regresión lineal (4.77), con las redes neuronales en segundo lugar (5.2) y el sistema difuso en tercero (5.45). Es interesante también notar que el puntaje ANEW obtuvo un error más grande que la designación aleatoria.

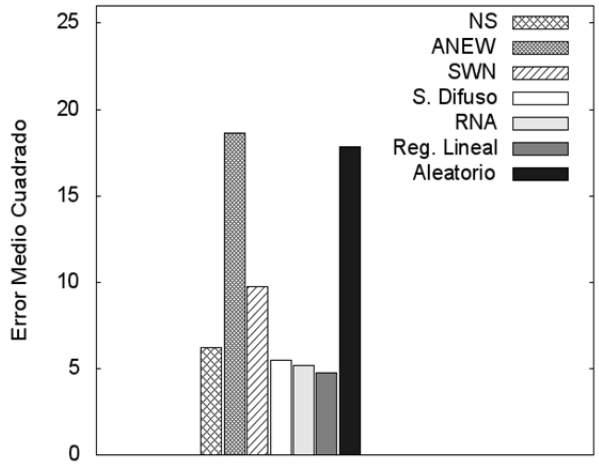

Figura 3. Error medio cuadrado para las técnicas propuestas de regresión.

En cuanto al uso de características, la combinación de tres atributos (1, 2 y 4) fue la que obtuvo en promedio el mayor porcentaje de aciertos con un 70.99% para la tarea de clasificación; le siguió la combinación de cinco atributos (todos menos el 3) con un 70.63%, y finalmente la combinación de todos los atributos con un 69.95%. Para la tarea de regresión, fue la combinación que utiliza los seis atributos la que obtuvo el menor error (3.2), empatada con la que utiliza cinco atributos; les siguió la combinación de los atributos 1, 2 y 4 (4.18).

C. Discusión

Con los resultados obtenidos, es posible ahondar en las preguntas realizadas al principio de esta sección. En primer lugar, conviene resaltar que—tanto para regresión como para clasificación—los métodos supervisados y estadísticos obtuvieron una ventaja considerable sobre el resto de los métodos utilizados (puntajes y sistema difuso); el clasificador bayesiano, las redes neuronales (RNA) y las máquinas de soporte vectorial (SVM), en particular, mostraron un buen desempeño para la tarea de clasificación con las dos representaciones utilizadas (metadatos y pesos TF-IDF); la regresión lineal y las mismas redes neuronales, mientras tanto, mostraron ser las mejores técnicas para la tarea de regresión. Esto nos lleva a considerar que las redes neuronales son un candidato factible, en general, para la tarea de predicción de texto agresivo, con la consideración de que el tiempo de entrenamiento pudiera extenderse para la representación TF-IDF con conjuntos de datos grandes.

En cuanto a la representación más conveniente, el mejor porcentaje de aciertos fue dado por los vectores TF-IDF de palabras, aunque el mejor promedio lo tuvieron los vectores de metadatos; por tanto, hasta ahora se ha visto que ambas representaciones son efectivas. Para trabajos futuros, incluso, podría considerarse una mezcla de ambas representaciones.

Con respecto a las características más prometedoras, la cantidad de palabras en el comentario, la cantidad de palabras ofensivas y el puntaje NS mostraron los mejores resultados; en ese sentido, pudiera también ser más eficiente utilizar solo estas tres características en lugar de todo el conjunto (seis). De igual manera, cabe destacar que, aunque el puntaje NS por sí solo no obtuvo tan buenos resultados, en conjunto con otras características y utilizado con técnicas supervisadas puede ser una característica de gran ayuda.

Finalmente, en cuanto a la cuestión de clasificación vs. regresión, en ambas tareas se obtuvieron puntajes considerables. Por tanto, probablemente se necesiten más experimentos para determinar qué tarea es mejor para apoyar la detección de ciberacoso. En ese sentido, pudiera ser conveniente introducir un esquema multi-clase (tres o más clases) para lograr una granularidad utilizable y compacta.

IV. TRABAJO RELACIONADO

La mayoría de los trabajos relacionados, como veremos más adelante, versan sobre la detección automática de ciberacoso sin ahondar en la tarea de predecir agresividad en los comentarios. De igual forma, estos trabajos no consideran la detección de texto agresivo como un problema de regresión (nuestro trabajo sí explora esta posibilidad), sino que están enfocados hacia la búsqueda de características para realizar clasificación binaria.

El trabajo por Dinakar et al. [12] considera que los comentarios agresivos están ligados a *temas sensibles*, tales como apariencia física, sexualidad, raza, cultura e inteligencia. Por tanto, entrenan clasificadores binarios y multi-clase con características variadas; entre ellas se encuentran unigramas TF-IDF, palabras ofensivas, etiquetas morfológicas frecuentes en texto ofensivo y bigramas frecuentes dentro de los temas sensibles. Utilizando estas características, diferentes tipos de clasificadores fueron evaluados, a saber: JRip, J48, máquinas de soporte vectorial (SVM) y el clasificador bayesiano. La evaluación se llevó a cabo con un conjunto de comentarios de Youtube manualmente clasificados y los mejores resultados se obtuvieron con JRip.

Dadvar et al. [13], por su parte, proponen tomar en cuenta aquella información relacionada con el género para detectar documentos agresivos; es por esto que entrenan dos clasificadores diferentes (uno por género). Las características utilizadas incluyen pronombres en segunda persona, palabras ofensivas (las más frecuentes por género) y valores TF-IDF. Para evaluar el enfoque propuesto, los autores utilizaron una SVM para clasificar comentarios de MySpace. Los resultados mostraron una mejora en la precisión debido al uso de información de género.

Otro trabajo sobresaliente es el propuesto por Nahar et al. [11], el cual extrae características semánticas mediante *Latent Dirichlet Allocation* (LDA), palabras ofensivas, valores TF-IDF y pronombres en segunda persona para entrenar una SVM. Al probar este enfoque, los autores utilizaron un conjunto de datos proveniente del taller *Content Analysis for the Web* (CAW); dicho conjunto de datos incluye comentarios de Twitter, Slashdot y MySpace.

Un trabajo similar corre por cuenta de Sood et al. [14], quienes detectan texto agresivo con ayuda del *Amazon's Mechanical Turk*, el cual emplean para etiquetar manualmente comentarios de un sitio de noticias. De estos comentarios extraen como características bigramas y raíces, con las cuales entrenan una SVM. Su motivación para utilizar una técnica supervisada consiste en superar algunos problemas relacionados con el uso de lexicones, tales como la detección de palabras mal escritas y variaciones específicas de palabras ofensivas.

A diferencia de estos trabajos, el nuestro se enfoca en la tarea de predicción de texto agresivo (los demás están orientados hacia propiamente la detección de ciberacoso). Para ello, evaluamos distintos tipos de técnicas (todas con diseños propios) con el fin de desvelar cuáles son las mejores opciones para tratar este problema específico.

V. CONCLUSIONES Y TRABAJO A FUTURO

En este trabajo hemos presentado una exploración de técnicas (todas con diseños propios) para atacar el problema de detectar texto agresivo en redes sociales. El fin general es la

detección oportuna de ciberacoso en redes sociales. Nuestros resultados sugieren que las técnicas de aprendizaje supervisado con características de palabras ofensivas son candidatos factibles para la predicción de texto agresivo.

Como trabajo a futuro se encuentra el incluir alguna de las técnicas más prometedoras como parte de un mecanismo que detecta automáticamente casos de ciberacoso en redes sociales. De igual manera, se contempla el uso de otros conjuntos de datos de mayor tamaño.

REFERENCIAS

[1] D. L. Espelage and S. M. Swearer. "Research on school bullying and victimization: What have we learned and where do we go from here?", *School Psych Rev*, vol. 32, no. 3, pp. 365-383, 2003.

[2] M. A. Campbell. "Cyber bullying: An old problem in a new guise?", *Aust J Guid Couns*, vol. 15, no. 1, pp. 68-76, 2005.

[3] F. Sticca and S. Perren. "Is cyberbullying worse than traditional bullying? Examining roles of medium, publicity, and anonymity for the perceived severity of bullying." *J Youth Adolesc.*, vol. 42, no. 5, pp. 739-750, 2012.

[4] J. Suler. "The online disinhibition effect", *Cyberpsychol Behav*, vol. 7, no. 3, pp. 321-326, 2004.

[5] S. Hinduja and J. W. Patchin. "Bullying, cyberbullying, and suicide", *Arch Suicide Res*, vol. 14, no. 3, pp. 206-221, 2010.

[6] M. M. Bradley and P. J. Lang. "Affective Norms for English Words (ANEW): Instruction manual and affective ratings", Reporte Técnico, Center for Research in Psychophysiology, University of Florida, 1999.

[7] A. Esuli and F. Sebastiani. "SentiWordNet: A publicly available lexical resource for opinion mining", en Proceedings of LREC, pp. 417-422, 2006.

[8] L. A. Zadeh. "Fuzzy sets", *Inf. Control*, vol. 8, no. 3, pp. 338-353, 1965.

[9] Wang, Li-Xin. *A course in fuzzy systems and control*. Prentice-Hall International, Inc., Estados Unidos, 1996.

[10] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, Springer, New York, 2006.

[11] V. Nahar, L. Xue, and P. Chaoyi. "An effective approach for cyberbullying detection". *Communications in Information Science and Management Engineering*, vol. 3, no. 5, pp. 238-247, 2013.

[12] K. Dinakar, R. Reichart, and H. Lieberman. "Modeling the detection of textual cyberbullying", en AAAI The Social Mobile Web, pp. 11-17, 2011.

[13] M. Dadvar, F. M. G. de Jong, R. J. F. Ordelman, and R. B. Trieschnigg. "Improved cyberbullying detection using gender information", en Proceedings of the 12th Dutch-Belgian Information Retrieval Workshop, pp. 23-26, 2012.

[14] S. Sood, J. Antin, and E. Churchill. "Using crowdsourcing to improve profanity detection", en AAAI Spring Symposium Series, pp. 69-74, 2012.

[15] G.A. Miller. "WordNet: a lexical database for English". *Commun ACM*, vol. 38, no. 11, pp. 39-41.

Laura Patricia Del Bosque received a master's degree in Information Engineering in 2009. She is currently a student from the PhD program in Engineering with Orientation in Information Technologies at the *Facultad de Ingeniería Mecánica y Eléctrica* that belongs to the *Universidad Autónoma de Nuevo León*. Her doctoral thesis concerns automatic cyberbullying detection in social media.

Sara E. Garza received her PhD in Information Technologies and Communications with a minor in Intelligent Systems in 2010. Since then, se has been teaching and researching at the *Facultad de Ingeniería Mecánica y Eléctrica* that belongs to the *Universidad Autónoma de Nuevo León*. So far, she has worked with data mining, artificial intelligence, and complex network analysis. Her current research interests include topic mining, graph clustering, and text mining.