Clasificación de acosadores en línea utilizando *Bootstrapping*

Yuridiana Alemán, Darnes Vilariño y David Pinto

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla, México
{candy.aleman,darnes,dpinto}@cs.buap.com.mx

Resumen En este artículo se presenta una propuesta para el etiquetamiento automático de un corpus utilizando la técnica de *Bootstrapping*. El corpus utilizado contiene los diálogos de depredadores sexuales en línea, los cuales se catalogan en dos tipos: Los que quieren material pornográfico y aquellos que buscan un encuentro con su víctima. Se utiliza el algoritmo Naive Bayes Multinomial con trigramas de palabras, etiquetando automáticamente un conjunto de entrenamiento y evaluando con una validación cruzada de 10 pliegues. Posteriormente se utiliza el modelo obtenido para clasificar un conjunto de evaluación.

Palabras clave: Bootstrapping, ofensores sexuales, chats, clasificación, Naive Bayes.

1. Introducción

Las conversaciones en línea o "chats" se han popularizado desde la masificación del internet. Los servicios de mensajería instantánea como *Facebook*, *Yahoo* y *Skype* forman parte de la vida diaria de la mayoría de la personas. Entre sus múltiples ventajas se encuentran el costo, rapidez, la sencillez de su manejo y sobre todo, el hecho de que se trata de una tarea que se realiza en privado. Sin embargo, estas ventajas pueden ser contraproducentes si no se tiene el debido cuidado al utilizarlos. Gran parte de los usuarios de este tipo de servicios son jóvenes e incluso niños, por lo que este medio de comunicación puede ser utilizado por personas para cometer delitos, especialmente depredadores sexuales que buscan a sus posibles víctimas mediante redes sociales, utilizando la información personal disponible en la red.

En este artículo se realiza un análisis de un conjunto de diálogos que se encuentran en el idioma inglés obtenidos de conversaciones donde participan ofensores sexuales. Se utiliza la técnica de *Bootstrapping* para clasificar a los ofensores sexuales en dos categorías, según el objetivo que persigan de sus víctimas. Finalmente, se realiza un análisis de los resultados obtenidos.

La estructura del artículo es la siguiente. En la sección 2 se da una descripción de algunas investigaciones alusivas al tema, posteriormente en la sección 3 se explica la técnica de Bootstrapping con sus variantes. La sección 4 presenta la metodología aplicada para la investigación, mientras que la sección 5 muestra

los resultados obtenidos. Finalmente, en la sección 6 se explican las conclusiones obtenidas y el trabajo futuro.

2. Trabajo relacionado

La mayoría de las investigaciones respecto al tema toman como punto de partida los experimentos realizados en [1], donde a partir de técnicas de clasificación automática de textos se identifican los diálogos de la víctima y el del depredador utilizando *SVM* y *k-NN*. En los experimentos realizados en [2] se intenta reconocer el *grooming attack*, el cual se define como un "Proceso de comunicación por el cual un autor aplica estrategias de búsqueda de afinidad, mientras que simultáneamente adquiere información sobre sus víctimas con el fin de desarrollar las relaciones que resulten en cumplimiento de su necesidad". Se utilizaron técnicas de clasificación de documentos para la creación de patrones y detectar en que fase se encuentra la conversación (ganarse a la víctima, cultivo de una relación amorosa ó petición de favor sexual o abuso). Una propuesta diferente para analizar conversaciones es la de [3], donde se identifica cuando ocurre explotación infantil en una conversación. En este trabajo se realiza una comparación entre el uso de características basadas en términos y las extraídas utilizando la herramienta *Linguistic Inquiry and Word Count (LIWC)*[4] para determinar el tipo de conversación que es (explotación infantil, fantasías sexuales entre adultos o chat general sin contenido sexual).

A partir del año 2012 se incluyó la subtarea *"Sexual Predator Identification"* dentro de la conferencia del *Lab Uncovering Plagiarism, Authorship and Social Misuse (PAN)* la cual consiste en determinar a partir de registros de chat, cual es la persona que trata de convencer a otros participantes para proporcionar algún favor sexual e identificar las líneas de las conversaciones de depredadores que son las mas distintivas del comportamiento del mismo. La precisión mas alta fue alcanzada por [5], donde se propone una metodología en dos etapas, una para clasificar las conversaciones donde interviene al menos un depredador sexual y la otra basado en los diálogos por persona para distinguir a los depredadores de las víctimas o pseudo víctimas.

Es importante mencionar que los trabajos relacionados, detectan conversaciones, parte de una conversación, tipo o si el participante es depredador o víctima, pero todas utilizan técnicas de clasificación supervisada y en ninguna se detecta el tipo de depredador, que es el objeto principal del presente artículo de investigación, ya que hasta el momento no existe un corpus etiquetado que funcione como entrenamiento.

3. *Bootstrapping*

En las investigaciones que se caracterizan por la falta de datos, o por datos no etiquetados, las técnicas mas utilizadas son las de *bootstrapping*[6,7,8], las cuales tratan de obtener una gran cantidad de información partiendo de una pequeña "semilla". Esto es, etiquetar automáticamente un gran número de instancias en

un corpus partiendo de un subconjunto muy reducido de instancias clasificadas manualmente.

Existen muchas técnicas de bootstrapping, las cuales difieren en la manera en que se van agregando instancias al subconjunto etiquetado o las técnicas de selección en caso de utilizarse alguna, sin embargo, todas van de acuerdo al objetivo de la técnica: *"La elevación de un pequeño esfuerzo inicial hacia algo mas grande y más significativo"*.

Algunas de las variantes a esta técnica reportadas en la literatura son:

- ***Self-training*:** Esta técnica se utiliza en [6], donde un corpus es utilizado para crear un modelo que se aplica a un conjunto nuevo de frases que tras ser etiquetadas, pasan a formar parte del corpus original, para volver a generar un nuevo modelo y avanzar iterativamente.
- ***Collaborative-training*:** Se emplea un mismo corpus para obtener diferentes modelos empleando distintas técnicas de aprendizaje. Posteriormente se introduce una fase de selección entre las diferentes opiniones que surgen de aplicar estos modelos al conjunto de frases nuevas y las etiquetas seleccionadas sirven para aumentar el corpus original y proseguir con la siguiente iteración.
- ***Co-training*:** Dos corpus inicialmente iguales sirven para crear dos modelos de diferentes características y los resultados de aplicar estos modelos a un conjunto de frases nuevas se "cruzan", es decir, las frases etiquetadas por el primer modelo sirven para aumentar el corpus que sirvió para crear el segundo modelo y viceversa. De esta forma un modelo no se alimenta únicamente de su percepción del corpus sino que recibe información de otro modelo que imprime otro punto de vista diferente a la resolución del mismo problema.

En este artículo se utiliza la técnica de *Self-training* para el etiquetamiento del corpus, ya que de las reportadas en la literatura, es mas sencilla al utilizar sólamente un modelo para el etiquetamiento de las instancias. Esta técnica se implementó utilizando la herramienta *WEKA*[9].

4. Metodología

Para la realización de los experimentos, se extrajeron diálogos de depredadores sexuales de dos conjuntos de conversaciones distintos:

1. *Perverted Justice*: Es un sitio web que contiene conversaciones entre depredadores sexuales convictos y voluntarios que se hacen pasar por menores de edad.
2. *Training* y *Test* de la competencia PAN 2012 (`http://pan.webis.de/`): Como se describe en [10], es un conjunto de conversaciones obtenidas de varios repositorios, como son los sitios de `http://omegle.inportb.com/`, `http://www.irclog.org/`, `http://krijnhoetmer.nl/irc-logs/`, además,

de algunas extraídas de `http://www.perverted-justice.com/`. El corpus resultante contiene diferentes tipos de conversaciones de depredadores sexuales, por ejemplo:

- Depredador/Víctima.
- Depredador/Pseudo-Víctima (Voluntario).
- Depredador/Pseudo-Víctima (Policía).

Además, se incluyen conversaciones de otros temas como foros de ayuda, y conversaciones *Adulto/Adulto*, en donde se maneja lenguaje de índole sexual, pero consensuado.

La metodología aplicada para el etiquetamiento automático de los diálogos se muestra en la figura 1.

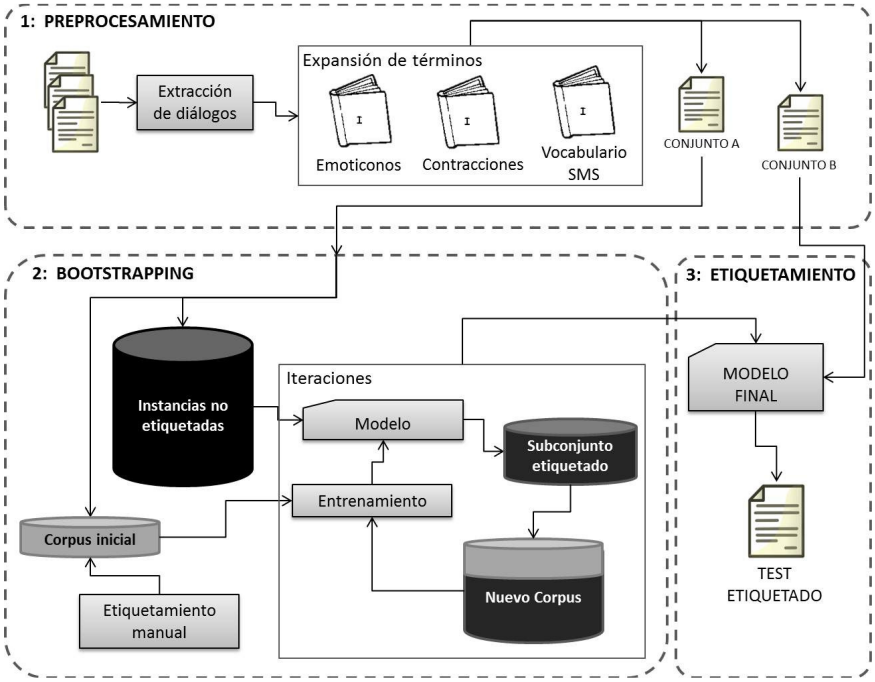

Figura 1. Metodología para el etiquetamiento automático del corpus

Como se puede observar, la investigación se divide en tres fases principales, las cuales se explican a continuación.

4.1. Preprocesamiento

De los conjuntos de conversaciones recopilados, se obtienen solamente los diálogos de los usuarios catalogados como depredadores sexuales, eliminando

el resto de las conversaciones y los diálogos de las víctimas. Finalmente, los diálogos recuperados se unen por usuario (sin importar si son de conversaciones distintas), obteniéndose un conjunto A de 700 diálogos para realizar la clasificación automática, y un conjunto B de 20 diálogos para evaluar el modelo final obtenido.

Gran parte de las diálogos contienen un exceso de términos no reconocidos por un diccionario, además de que abundan los emoticones, cadenas de símbolos raros, que pueden ser URLs, imágenes, entre otros. Dadas estas observaciones, se construyeron 3 recursos léxicos (diccionarios) para ayudar a enriquecer los textos. Estos recursos son:

1. **Emoticones**: Se obtuvo una lista con los emoticones mas comunes (también llamados "smileys"), esta lista fue enriquecida con los emoticones predefinidos de *Windows Messenger*, *Facebook* y *Gmail*. Con esto se lograron recopilar 344 elementos.
2. **Contracciones**: Esta lista contiene alrededor de 65 contracciones mas usadas en los Estados Unidos.
3. **Vocabulario SMS**: Es una lista obtenida de [11], la cual contiene 820 abreviaciones o simplificaciones mas usadas, especialmente por los jóvenes. Estos términos se han ido popularizando en SMS y chats, donde el tiempo de respuesta es importante y generalmente no se toman en cuenta reglas ortográficas y gramaticales.

Tanto en el conjunto A como en el B, todas las ocurrencias de alguno de estos recursos léxicos son sustituidas por su correspondiente significado.

4.2. Aplicación de la técnica

Para utilizar la técnica *Self train* para el etiquetamiento automático, se etiqueta un subconjunto de A manualmente para obtener $A_{etiquetado}$. El etiquetamiento fue realizado por una sola persona, y consistió en leer detenidamente toda la conversación (especialmente los diálogos del depredador) y en base a las peticiones o favores que pedía el depredador a la víctima se asignó una de las siguientes categorías:

- **cat1**: Depredadores que se contentan con obtener y comerciar imágenes de pornografía infantil.
- **cat2**: Depredadores que buscan un encuentro cara a cara con los niños.

De esta forma, el conjunto $A_{etiquetado}$ se conforma de 10 diálogos de depredadores pertenecientes a la categoría 1 y 10 pertenecientes a la categoría 2, estos elementos se eliminan del conjunto A para evitar su doble clasificación. Posteriormente, los 680 diálogos de A se etiquetan con las siguientes reglas:

- Para todas las clasificaciones se utilizan trigramas de palabras, utilizando *LovinsStemmer*.

- Después da hacer experimentos preliminares, se determinó utilizar el clasificador *Naive Bayes Multinomial* para todas las iteraciones.
- Se realizan un total de 6 iteraciones, en cada una se crea un modelo utilizando $A_{etiquetado}$ para clasificar una cantidad de diálogos extraídos aleatoriamente de A.
- Se evalúa la probabilidad de clasificación proporcionada por *WEKA*, si esta es mayor o igual a 90 %, la instancia se elimina de A y se agrega a $A_{etiquetado}$ con su categoría asignada, si esto no se cumple, la instancia no se agrega y continúa en el conjunto A. Al final, se realiza una última clasificación con todas las instancias que no alcanzaron esta probabilidad y se vuelven a clasificar para ser agregadas a $A_{etiquetado}$.
- Las instancias de A extraídas en cada iteración son 30, 50, 100, 200 y 300 respectivamente.

Como se mencionó anteriormente, no se cuenta con un corpus etiquetado, por lo que no hay forma de obtener métricas para la evaluación de las clasificaciones realizadas, sin embargo se realizan evaluaciones en cada iteración con $A_{etiquetado}$ utilizando validación cruzada de 10 pliegues y obteniendo para cada categoría las siguientes métricas:

- *Razón de Falsos Positivos (FPR)*: También denominado ratio o *fall-out*. Se calcula como $FPR = \frac{FP}{FP+VN}$.
- *Razón de Verdaderos Positivos (VPR)*: También denominado recuerdo en recuperación de información, representa la fracción de datos recuperados que son positivos. Se calcula como $VPR = \frac{VP}{VP+FN}$.
- *Exactitud (ACC)*: Representa la fracción de datos evaluados correctamente sobre el total. Se calcula como $ACC = \frac{VP+VN}{Total\ de\ instancias}$.

Donde VP, FN , VN y FP son los datos de la matriz de confusión retornada por *WEKA* (instancias verdaderas positivas, falsas negativas, verdaderas negativas y falsas positivas respectivamente). Con las métricas obtenidas se construye una curva ROC (*Receiver Operating Characteristic*) para la clasificación.

4.3. Clasificación del conjunto *B*

Una vez etiquetadas todas las instancias de A han sido agregadas a $A_{etiquetado}$, se genera un modelo utilizando nuevamente trigramas de términos y *Naive Bayes Multinomial* para etiquetar los 20 diálogos del conjunto B.

5. Resultados

En la tabla 1 se muestra como se fue etiquetando el conjunto A, tomando como *training* el conjunto $A_{etiquetado}$. En la iteración 1, $A_{etiquetado}$ cuenta con 20 instancias, 10 de cada categoría. Se toman 30 instancias de A como *test*, las cuales se clasifican con el modelo creado por $A_{etiquetado}$. De las 30 instancias clasificadas, 15 tienen una probabilidad de predicción menor al 90 %, por lo que

no se agregan a $A_{etiquetado}$. De las 15 restantes, 9 se agregan a la categoría 1 y 6 a la categoría 2, al término de esta iteración, $A_{etiquetado}$ cuenta con 19 instancias de la categoría 1 y 16 de la categoría 2, las cuales se utilizan como *training* en la siguiente iteración.

En la iteración 2, se extraen 50 instancias de A, de las cuales 20 obtienen una probabilidad de predicción menor y 30 son agregadas al conjunto $A_{etiquetado}$ (17 en la categoría 1 y 13 en la categoría 2), por lo que en la iteración 3, se cuenta con un *training* de 65 instancias. Este proceso se repite hasta etiquetar todo el conjunto A.

Tabla 1. Instancias agregadas al conjunto $A_{etiquetado}$ en cada iteración realizada

	$A_{etiquetado}$		Diálogos	Agregadas		No
Iteración	Categoría1	Categoría2	de A	Categoría1	Categoría2	Agregadas
1	10	10	30	9	6	15
2	19	16	50	17	13	20
3	36	29	100	24	33	43
4	60	62	200	45	61	94
5	105	123	300	63	93	144
Final	168	216	316	37	279	-

El conjunto $A_{etiquetado}$ quedó conformado por 205 depredadores de la categoría 1 y 495 de la categoría 2. Como se puede observar el tamaño del test crece considerablemente en cada iteración, así como el número de instancias agregadas a la categoría dos en las últimas iteraciones, este último dato en cierta medida es aceptable, ya que en la realidad existen mas depredadores sexuales que buscan un encuentro con la víctima, además de que son mas peligrosos.

En la tabla 2 se muestran los resultados obtenidos al aplicar las distintas métricas de evaluación.

Tabla 2. Resultados obtenidos por clase en cada iteración

	Cat 1			Cat2		
Iteración	FPR	VPR	ACC	FPR	VPR	ACC
1	0.00	0.80	0.90	0.20	1.00	0.90
2	0.25	0.74	0.74	0.26	0.75	0.74
3	0.25	0.82	0.79	0.18	0.75	0.79
4	0.02	0.85	0.92	0.15	0.97	0.92
5	0.06	0.83	0.89	0.17	0.94	0.89
Final	0.04	0.82	0.92	0.18	0.96	0.92

En todas las iteraciones se observa que las dos razones calculadas son muy similares, y la ACC está arriba del 90 % en la mayoría de los casos. En las

iteraciones 2 y 3 la ACC desciende hasta 74 y 79 % respectivamente, sin embargo en las siguientes iteraciones se incrementa. Además, la razón de falsos positivos es muy baja, especialmente en la categoría 1.

En la figura 2 se muestran las curvas ROC para las dos categorías en la primera y la última iteración. Como se puede observar, en todos los casos el área de la curva está por encima del nivel medio y sobre todo, no existe una diferencia significativa entre las curvas de la iteración inicial con sólo 20 instancias a las curvas de la iteración final, con 700 instancias.

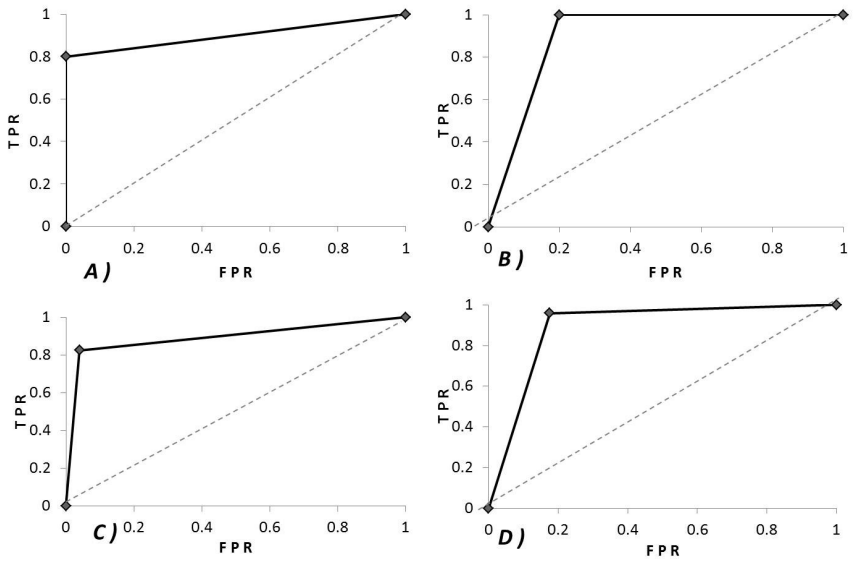

Figura 2. Curvas ROC para el conjunto $A_{etiquetado}$: Categoría 1 (A) y 2 (B) en le iteración 1 y categoría 1 (C) y 2 (D) en la iteración final

Finalmente el etiquetamiento del conjunto B con el modelo creado por los elementos de $A_{etiquetado}$, clasificó 14 diálogos como categoría 1 y 6 como categoría 2. Se eligieron al azar conversaciones de cada categoría para analizarlas manualmente, en las etiquetadas como categoría 1, contienen frases como "i have a pic for you", "call me", "send movie", entre otras. En la categoría 2 se encuentran frases mas relacionadas a sexo explícito. Sin embargo, en algunas conversaciones se crea confusión al momento de realizar la clasificación, ya que el depredador habla de fotos, o hacer video llamadas y posteriormente intenta convencer a la víctima para encontrarse personalmente.

6. Conclusiones y trabajo futuro

En este artículo se presentó una metodología para el etiquetamiento automático de diálogos pertenecientes a depredadores sexuales. Se utiliza la técnica de *Bootstrapping* basado en un pequeño subconjunto etiquetado manualmente. Los resultados obtenidos dan una propuesta de un corpus utilizable para futuros experimentos concernientes a la detección y clasificación de depredadores sexuales en redes sociales.

Como trabajo futuro se busca perfeccionar el etiquetamiento automático utilizando otras versiones de la técnica o el uso de otros conjuntos de características.

Referencias

1. Pendar, N.: Toward spotting the pedophile telling victim from predator in text chats. In: Proceedings of the International Conference on Semantic Computing. ICSC '07, Washington, DC, USA, IEEE Computer Society (2007) 235–241
2. Michalopoulos, D., Mavridis, I.: Utilizing document classification for grooming attack recognition. In: Proceedings of the 2011 IEEE Symposium on Computers and Communications. ISCC '11, Washington, DC, USA, IEEE Computer Society (2011) 864–869
3. Miah, M.W.R., Yearwood, J., Kulkarni, S.: Detection of child exploiting chats from a mixed chat dataset as a text classification task. In: Proceedings of the Australasian Language Technology Association Workshop 2011, Canberra, Australia (December 2011) 157–165
4. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., Booth, R.J.: The Development and Psychometric Properties of LIWC2007. This article is published by LIWC Inc, Austin, Texas 78703 USA in conjunction with the LIWC2007 software program.
5. Villatoro-Tello, E., Juárez-González, A., Escalante, H.J., y Gómez, M.M., Pineda, L.V.: A two-step approach for effective detection of misbehaving users in chats. In: CLEF (Online Working Notes/Labs/Workshop). (2012)
6. Clark, S., Curran, J.R., Osborne, M.: Bootstrapping pos taggers using unlabelled data. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4. CONLL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 49–55
7. Mihalcea, R.: Bootstrapping large sense tagged corpora. In: LREC. (2002)
8. Mihalcea, R.: Co-training and self-training for word sense disambiguation. In: Proceedings of CoNLL-2004, Boston, MA, USA (2004) 33–40
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11**(1) (November 2009) 10–18
10. Inches, G., Crestani, F.: Overview of the international sexual predator identification competition at pan-2012. In Forner, P., Karlgren, J., Womser-Hacker, C., eds.: CLEF (Online Working Notes/Labs/Workshop). (2012)
11. Symens, B.: Acronyms Dictionary for Texting Chatting E-mail. Rebecca J Symens