Detección de depredadores sexuales utilizando un sistema de consulta y clasificación supervisada

Yuridiana Alemán, Darnes Vilariño, David Pinto, Mireya Tovar

Facultad de Ciencias de la Computación, BUAP,
Puebla, Mexico

yuridiana.aleman@gmail.com, darnes@cs.buap.mx, dpinto@cs.buap.mx,
mireyatovar@gmail.com

Resumen. En este artículo se realiza un análisis entre técnicas de clasificación supervisada y el uso de un sistema de consultas como clasificador, dicho análisis se aplica a la detección de depredadores sexuales por medio de una metodología de dos fases. Primero, se reduce el corpus seleccionando aquellas conversaciones con más probabilidades de pertenecer a depredadores sexuales para posteriormente clasificar a los usuarios, finalmente se analizan los diálogos de los usuarios clasificados como depredadores. Los resultados obtenidos muestran que se obtienen mejores resultados con la clasificación supervisada, sin embargo, estos se incrementan cuando se fusionan ambas técnicas.

Palabras clave: Sistema de consulta, clasificación, *POStagger*, depredadores.

1. Introducción

Gracias al avance de la tecnología, las redes sociales han avanzado a pasos agigantados y con ello, la forma en que los depredadores sexuales hacen contacto con una víctima. El FBI cataloga a estos adultos como "viajeros", porque van de un lugar a otro para tener relaciones sexuales con niños que conocieron por Internet. Estas personas acechan las salas de chat, buscando convencer a los adolescentes de participar en coqueteos cibernéticos. Con el paso del tiempo, estos adultos desarrollan romances a distancia, para después intentar persuadir a sus víctimas a entablar relaciones íntimas.

Dada esta situación, se han hecho varios avances en la detección automática de depredadores sexuales, empresas como *Facebook* han impulsado este tipo de investigaciones, sin embargo, todavía quedan muchos aspectos que tomar en cuenta para que la detección sea eficaz. Algo que obstruye este tipo de investigaciones es la escasez de material para el análisis, además, se debe tomar en cuenta otros aspectos como el tipo de escritura usado en los chats, así como el hecho de que existen conversaciones que contienen términos obscenos, pero no es propiamente el caso de un depredador sexual, y viceversa, es decir, los depredadores no siempre usan términos de ese tipo, si no un lenguaje mas formal.

Dados los datos disponibles, en el presente estudio se analizarán únicamente conversaciones en el idioma inglés.

En este artículo se realizan diversos experimentos para la detección de conversaciones en donde participan depredadores sexuales e identificar cual de los usuarios de dicha conversación es que el intenta persuadir a los demás. La propuesta consta de 2 etapas, primeramente se realiza una clasificación de conversaciones, para posteriormente detectar los usuarios que son depredadores sexuales. Para ambas fases se utilizan métodos de clasificación supervisada con diferentes conjuntos de características, el sistema de consultas únicamente se utiliza en la primera fase.

El artículo está estructurado de la siguiente manera. En la sección II se detalla es estado del arte referente a este tema de investigación. La sección III muestra la metodología planteada y el preprocesamiento dado a las conversaciones. La sección IV muestra los resultados obtenidos en las fases 1 y 2 respectivamente. Finalmente, la sección V muestra las conclusiones obtenidas y el trabajo futuro para esta investigación.

2. Estado del arte

La mayoría de los trabajos que abordan la temática de detección de depredadores sexuales toman como punto de referencia la investigación presentada en [8], donde se realiza un estudio piloto sobre el uso de técnicas de clasificación automática de textos para identificar depredadores sexuales en línea. Se trabaja con un corpus obtenido del sitio *Perverted Justice*[1]. Este estudio se presenta como una investigación inicial para la detección del *grooming attack*[2]. En los experimentos realizados, se identifican los textos de los depredadores sexuales y los textos de las víctimas (clasificación en dos clases) utilizando Máquinas de soporte vectorial (*SVM*) y *k-vecinos mas cercanos* (*k-NN*). Se hicieron varias pruebas con diferentes características (de 5,000 a 10,000), obteniendo mejores resultados con un conjunto de 10,000 características con *k-NN* (k=30).

En los últimos años se han publicado varias investigaciones sobre esta temática, la mayoría provenientes del congreso *CLEF PAN Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse*[3]. En dicho congreso se abordan temas relacionados a la atribución de autoría, plagio de textos y vandalismo informático. A partir del año 2012 se incluyó dentro de atribución de autoría la subtarea *"Sexual Predator Identification"*.

El *F-score* mas alto en la subtarea fue alcanzado por [12], en esta investigación no se realiza ningún tipo de preprocesamiento, sólo un filtrado para quitar aquellas conversaciones muy cortas, con caracteres no entendibles o donde los

[1] *http://www.perverted-justice.com/*

[2] Definido en [4] como "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, por ejemplo acoso sexual físico"

[3] *http://pan.webis.de/*

participantes tienen pocas intervenciones, con esto se logra reducir drásticamente la dimensión del corpus. Además, proponen una metodología de dos etapas, una para clasificar las conversaciones donde interviene al menos un depredador sexual y otro basado en los diálogos por persona para distinguir a los depredadores de las víctimas o pseudovíctimas. Para la primera clasificación se obtuvo mayor precisión con SVM y *tf-idf*, para el caso del segundo modelo se obtienen mejores resultados con redes neuronales.

Todas las investigaciones mencionadas, tienen en común que utilizan técnicas de clasificación para resolver la tarea, sin embargo, en [1] se propusieron dos metodologías diferentes. Una de ellas está basada en técnicas de recuperación de información, para lo cual se desarrolló un diccionario de términos sexuales con 919 entradas. En cada una de estas entradas se encuentran un conjunto de términos similares a un sentido en particular. Las conversaciones fueron representadas con la técnica de *Posting List*[5]. Cada entrada del diccionario sexual se considera una consulta y se devuelven las primeras 10 conversaciones que de alguna manera incluyen términos alusivos al sexo. Esta propuesta logró detectar mayor número de depredadores, pero también reportó un gran número de conversaciones que no necesariamente están asociadas a individuos que buscan un favor sexual, sino que utilizan términos obscenos para comunicarse.

3. Metodología

Para la realización de los experimentos se utiliza como *training* las conversaciones extraídas del sitio web de la fundación *PJFI.org* (Estados Unidos), la cual pretende contrarrestar a los pedófilos que intentan acercarse a una víctima por medio de la red. En el sitio web se publican conversaciones de depredadores sexuales con sus "víctimas", que en realidad son personas que se hacen pasar por adolescentes y/o niños. Además, se añade al conjunto el *training* de la competencia "PAN 2012", obtenidas del sitio web de dicha competencia (*http://pan.webis.de/*). El *test* de dicha competencia se utiliza también como *test* de los experimentos realizados. El comité organizador proporciona el *gold* de los usuarios etiquetados como depredadores.

En base a los conjuntos que datos con los que se cuentan, se propone una metodología como se muestra en la figura 1. En general, la propuesta se divide en dos fases principales:

1. Clasificación de las conversaciones en dos posibles tipos: *"Predator"* y *"No Predator"*.
2. Clasificación de los diálogos que conforman las conversaciones recuperadas en la fase 1 para detectar si un usuario es una víctima y un depredador.

3.1. Preprocesamiento de corpus

Una vez determinados los conjuntos de conversaciones para el *training* y *test*, se obtuvieron algunos promedios y medidas para su análisis, los cuales se

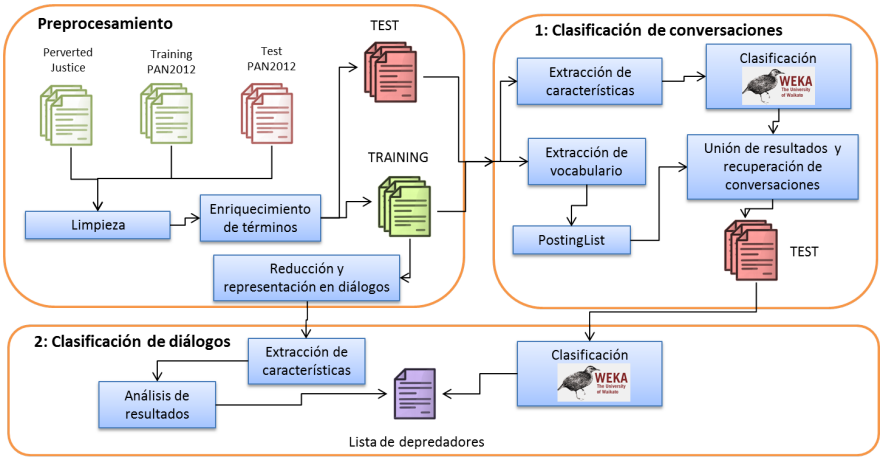

Fig. 1. Metodología propuesta para la investigación.

muestran en la tabla 1[4]. Aquí se observa que el corpus de entrenamiento tiene más riqueza en cuanto a la extensión de las conversaciones, aunque cabe señalar que los valores son muy extremos, por ejemplo en el *training* la conversación con mayor número de usuarios es de 30, mientras que en el *test* es de 115. Para el caso de los usuarios, el que participa en mas conversaciones tiene 30 en el *training*, mientras que en el *test* la mayor participación es de 128 conversaciones.

Tabla 1. Descripción del corpus usado.

DATO	*Training*	*Test*
Vocabulario	317,450	624,755
Conversaciones de Depredadores	2,353	3,715
Conversaciones de No Depredadores	64,884	151,413
Depredadores	480	250
No depredadores	97,807	218,431
Usuarios por conversación	2.28	2.29
Conversaciones por usuario	1.62	1.56
Líneas	17.24	13.23
Longitud	568.30	532.24
Palabras	108.85	96.16

Algo notable es que el vocabulario del *test* es considerablemente mayor al vocabulario del *training*, pero dada la naturaleza de los textos, mas que palabras diferentes, pueden existir distintos modos de escritura para una sola palabra, por ejemplo, en salas de chat es muy común poner el término "lol", pero si un usuario escribe "lol" y "lool", se tomarán como dos palabras distintas.

Las conversaciones contienen un exceso de términos no reconocidos por un diccionario, además de que abundan los emoticones, cadenas de símbolos no

[4] Los últimos cinco datos están expresados en promedios

reconocibles, que pueden ser URLs, imágenes, entre otros, ésto en gran medida por el tipo de vocabulario utilizado en estas conversaciones, además, algunas abordan temas de programación, por lo que el número de símbolos extraños aumenta. Por lo tanto, se construyeron 3 recursos léxicos (diccionarios) para ayudar a enriquecer los textos. Estos recursos son:

1. **Emoticones**[5]: Se obtuvo una lista con los emoticones mas comunes, dicha lista fue enriquecida con los emoticones predefinidos de *Facebook* y *Gmail*. La lista recopilada cuenta con 344 elementos.
2. **Contracciones**[6]: Esta lista contiene alrededor de 65 contracciones más usadas en los Estados Unidos.
3. **Vocabulario SMS**: Es una lista obtenida de [10], la cual contiene 820 abreviaciones o simplificaciones más usadas 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 *training* como en el *test*, todas las ocurrencias de alguno de estos recursos fueron sustituidas por su correspondiente significado, además de documentar el número de ocurrencias de cada tipo por conversación para futuros experimentos. Con los cambios hechos al corpus, el vocabulario disminuyó significativamente (de 317,455 a 181,291 palabras para el caso del *training* y de 624,755 a 360,880 palabras para el *test*).

3.2. Conjuntos de Características

En experimentos preliminares se trabajó con varios conjuntos de características, pero en algunos los resultados eran muy bajos o simplemente inviables de procesar dada la magnitud del corpus, como el caso del uso del vocabulario. Por lo tanto, se escogieron las características que se considera pueden diferenciar los textos de una víctima, de los textos del depredador. Los conjuntos seleccionados se mencionan a continuación.

SUBDUE: Se analiza el *training* con la herramienta *SUBDUE*[7], a fin de extraer las palabras mas representativas del texto utilizando la extracción de sub-estructuras mas comunes con una representación de grafos. Se utiliza una representación basada en la notación *gSpan*[13], en la cual un grafo es representado como una 4-tupla $G = (V, E, L, l)$ donde V es un conjunto (no vacio) de vértices, E es un conjunto de aristas de la forma $E \subseteq V \times V$, L es un conjunto de etiquetas y l es una función que asigna una etiqueta a un par de vértices asociados. En la figura 2 se muestra un ejemplo de esta representación con el texto "of course i need help with the machine".

Cabe mencionar que el tiempo de ejecución para extraer esta representación depende de la cantidad de palabras pertenecientes a cada conjunto, por lo que es inviable aplicarlo en el *training* de la fase 1.

[5] *http://www.netlingo.com/smileys.php*
[6] "Corrupciones fonéticas", en *http://es.wikibooks.org*

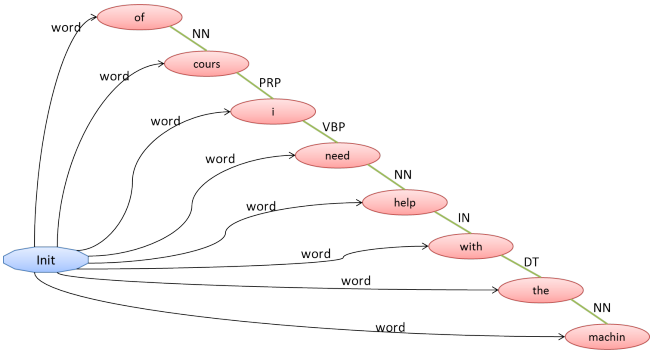

Fig. 2. Ejemplo de representación basada en secuencias de palabras.

Características Léxicas: Se realiza un conteo de algunos símbolos y palabras escritas de una manera singular. En la tabla 2 se listan dichas características, las cuales son obtenidas en la fase de preprocesamiento. Cabe mencionar que en este conjunto están contemplados los conteos de cada elemento de la lista, por lo que todos los atributos son numéricos, excepto el de la clase.

Tabla 2. Características léxicas.

Palabras que inician con mayúscula	URLs	"?"
Emoticones	Palabras escritas en mayúsculas	"."
Palabras truncadas	Contracciones	"!"
Números	";"	Total de signos utilizados
","	":"	

Categorías Gramaticales: Se utilizó la herramienta *POS-tagger* de la Universidad de Stanford[11] para obtener las partes del discurso y el lema de cada término dentro de las conversaciones. Al igual que en el conjunto anterior, fueron considerados como características los conteos de las apariciones por conversación.

Sufijos: Se tomaron como atributos los sufijos existentes para el idioma inglés donde cada sufijo representa un atributo y las veces que aparece en cada conversación es el valor para dicho atributo. Un sufijo se define como "Un afijo que va pospuesto y, en particular, los pronombres que se juntan al verbo y forman con él una sola palabra"[7].

Signos: Se realiza el conteo de todos los signos existentes en *string.punctuation*[8], la cual es una constante definida en el módulo *String* de *python*. Esta constante contiene una cadena con los signos de puntuación en *ASCII* (figura 3).

[7] `http://www.rae.es/`
[8] https://docs.python.org/2/library/string.html

```
! `` # $ % & / ( ) * + . ,
- : ; < = > ? @ [ \ ] ^ _
          ` { : } ~
```

Fig. 3. Cadena de símbolos contenida en *string.punctuation*.

3.3. *Posting List*

Partiendo de la idea de un sistema básico de consultas, se creó un algoritmo que clasifica las conversaciones del *test* en base a un *Posting List*[5] construido con las conversaciones del *training* (Algoritmo 1).

Algoritmo 1 Algoritmo para consultas con *Posting List* basado en trigramas de lemas.

Entrada: $CONVER$: Conjunto de conversaciones del *test* sin palabras cerradas.
Entrada: $INDICE$: Elementos del índice.
Entrada: $LIST$: Lista de conversaciones del *training* con su respectiva categoría.
Entrada: $TOPE$: Porcentaje de documentos tomados en cuenta para la asignación de la categoría.
Salida: $CATEGORIA$: Lista de conversaciones con una categoría asignada.

para $a \in Conver$ **hacer**
  $lemas \leftarrow Extrae\ Lemas\ Conversacion(a, Conver)$
  para $b \in lemas$ **hacer**
    $post \leftarrow Extrae\ Claves\ Conversacion(b, INDICE)$
    para $c \in post$ **hacer**
      $Valor \leftarrow Extrae\ Valor\ Trigrama(c, INDICE)$
      $peso_c \leftarrow peso_c + Valor$
      $conteo_c++$
    fin para
  fin para
  para $d \in conteo$ **hacer**
    si $conteo_d >= TOPE$ **entonces**
      $etiqueta \leftarrow Extrae\ Categoria(d, LIST)$
      $Categoria_{etiqueta} \leftarrow Categoria_{etiqueta} + peso_d$
    fin si
  fin para
  si $Categoria_{\text{“}depredador\text{”}} <= 0$ **and** $Categoria_{\text{“}no\ predator\text{”}} <= 0$ **entonces**
    $CATEGORIA_a \leftarrow$ “*sin palabras*”
  si no, si $Categoria_{\text{“}depredador\text{”}} == Categoria_{\text{“}no\ predator\text{”}}$ **entonces**
    $CATEGORIA_a \leftarrow$ “*sin categoria*”
  si no, si $Categoria_{\text{“}depredador\text{”}} > Categoria_{\text{“}no\ predator\text{”}}$ **entonces**
    $CATEGORIA_a \leftarrow$ “*predator*”
  si no
    $CATEGORIA_a \leftarrow$ “*no predator*”
  fin si
fin para
retornar $CATEGORIA$

Para la creación de dicho recurso, se extraen los lemas de cada palabra y se eliminan las palabras cerradas, posteriormente se crean trigramas de lemas para anexarlos a un índice, en donde se incluye el trigrama y la lista de conversaciones donde este aparece. Además, para cada trigrama se calcula su *tf-idf*, el cual le asigna un peso para cada trigrama dentro de la colección de conversaciones. Posteriormente se tomaron en cuenta las siguientes consideraciones:

- Algunos términos de las conversaciones del *test* no se encuentran en el índice creado, por lo que estos no se consideran en la consulta.
- En la asignación de la categoría, se utiliza una variable llamada TOPE, la cual determina el porcentaje de documentos necesarios para la asignación de la categoría final. Por lo tanto, una conversación del *test* es tomada en cuenta si el número de términos en los que aparece es mayor o igual a dicha variable.
- Se presentan casos de conversaciones en donde ningún término supera el TOPE, o en su defecto, existen el mismo número de documentos positivos y negativos para asignar la categoría, por lo tanto, se toman en cuenta las siguientes categorías para asignar a una conversación:
  - **"Predator"**: Cuando hay más conversaciones positivas que negativas que contengan los términos de la consulta.
  - **"No-predator"**: Cuando hay más conversaciones negativas que positivas que contengan los términos de la consulta.
  - **"Sin palabras"**: La conversación no tiene palabras que estén incluidas en el índice, o la frecuencia de las conversaciones retornadas no superan el TOPE.
  - **"Sin categoría"**: Se tiene la misma frecuencia de conversaciones en las dos categorías.

3.4. Métricas de evaluación

Los resultados de los diferentes experimentos fueron evaluados con las siguientes métricas utilizadas en la recuperación de información.

1. **Precisión (P):** Fracción de los datos recuperados que son relevantes.

$$P = \frac{items\ relevantes\ recuperados}{items\ recuperados} \tag{1}$$

2. **Recuerdo (R):** Fracción de los datos relevantes que son recuperados.

$$R = \frac{items\ relevantes\ recuperados}{items\ relevantes} \tag{2}$$

3. ***F-score*:** Media armónica entre la precisión y el recuerdo.

$$F - score = 2 * \frac{P * R}{P + R} \tag{3}$$

4. Resultados obtenidos

Para realizar los experimentos de clasificación supervisada se utilizaron varios algoritmos implementados en la herramienta *WEKA*[3], después de experimentos preliminares y dado el tipo de conjuntos de características utilizados, se decidió utilizar árbol de decisión[9], redes neuronales[6] y bosque aleatorio[2]. Además, se implementa un sistema de voto, el cual consiste en tomar como positiva una

conversación si por lo menos uno de los tres clasificadores utilizados la clasificó como tal. Se utilizaron los conjuntos individuales y las uniones de tales conjuntos. En la tabla 3 se muestran los resultados obtenidos.

En la tabla se muestran las conversaciones de depredadores que fueron recuperadas (Rec Pos), precisión y recuerdo, además, para homogeneizar los datos se agrega la variable *F-score* y el número de depredadores incluidos en las conversaciones recuperadas (No Dep). Se observa que los mejores resultados en todas las métricas utilizadas se presentan en el sistema de voto, utilizando diferentes combinaciones de características. La precisión mas alta fue alcanzada utilizando solamente el conjunto de categorías gramaticales utilizando el algoritmo de redes neuronales, sin embargo, dado que ésta se considera una primera fase, se prefiere que el número de depredadores encontrados sea alto para poder proseguir con el análisis de clasificación de la segunda etapa de la propuesta inicial. Por otro lado, aunque el mejor *F-score* fue obtenido utilizando los conjuntos de signos y categorías, se obtienen mas conversaciones con depredadores utilizando los conjuntos de características léxicas y categorías gramaticales, esto se debe a que en el *test* existen usuarios que participan en mas de una conversación, por lo que un mayor *f-score* no implica una mayor cantidad de depredadores recuperados

Tabla 3. Resultados para clasificación con los conjuntos de características.

Conjunto	**Clasificador**	**Rec Pos**	**Precisión**	**Recuerdo**	***F-score***	**No Dep**
Léxicas	Árbol de decisión	1,057	0.306	0.285	0.295	210
	Bosque aleatorio	1,247	0.444	0.336	0.382	214
	Redes neuronales	1,182	0.288	0.318	0.302	211
	Voto	1,431	0.217	0.385	0.277	221
Categorías	Árbol de decisión	923	0.568	0.248	0.346	195
	Bosque aleatorio	1,012	0.627	0.272	0.380	207
	Redes neuronales	851	**0.719**	0.230	0.348	190
	Voto	1,249	0.481	0.336	0.396	216
Sufijos	Árbol de decisión	494	0.554	0.133	0.214	155
	Bosque aleatorio	476	0.563	0.128	0.209	160
	Redes neuronales	277	0.632	0.075	0.133	113
	Voto	709	0.459	0.191	0.270	186
Signos	Árbol de decisión	476	0.452	0.128	0.200	139
	Bosque aleatorio	508	0.342	0.137	0.195	153
	Redes neuronales	156	0.378	0.042	0.076	80
	Voto	714	0.304	0.192	0.236	172
Léxicas categorías	Árbol de decisión	1,148	0.421	0.309	0.356	213
	Bosque aleatorio	1,204	0.602	0.324	0.421	216
	Redes neuronales	1,192	0.488	0.321	0.387	213
	Voto	1,588	0.337	**0.427**	0.377	**226**
Signos categorías	Árbol de decisión	1,056	0.494	0.284	0.361	202
	Bosque aleatorio	1,015	0.698	0.273	0.393	207
	Redes neuronales	980	0.758	0.264	0.391	199
	Voto	**1,492**	0.483	0.402	**0.439**	225
Sufijos signos	Árbol de decisión	831	0.493	0.224	0.308	181
	Bosque aleatorio	818	0.590	0.220	0.322	190
	Redes neuronales	43	0.589	0.012	0.023	35
	Voto	1,104	0.452	0.297	0.359	202
Léxicas sufijos	Árbol de decisión	1,058	0.463	0.285	0.353	210
	Bosque aleatorio	1,063	0.584	0.286	0.384	201
	Redes neuronales	1,183	0.486	0.318	0.385	215
	Voto	1,454	0.342	0.391	0.365	221

En cuanto al sistema de recuperación de información, se realizaron varios experimentos con distintos valores de la variable TOPE, los resultados se muestran en la tabla 4. Se puede observar que a medida que la variable TOPE se reduce, se incrementan las conversaciones recuperadas y por lo tanto, las métricas evaluadas. Además, se recuperan más conversaciones que en los experimentos de clasificación supervisada, aunque con un *F-score* menor.

Tabla 4. Resultados para la clasificación de conversaciones utilizando el sistema de consultas.

TOPE	Sin palabras	Sin categoría	Rec Pos	*Precision*	*Recall*	*F-score*	No Dep
100	112,054	239	709	0.274	0.191	0.225	192
90	110,983	239	833	0.306	0.224	0.259	201
80	103,868	241	852	0.303	0.230	0.261	202
70	97,220	242	878	0.300	0.236	0.264	204
60	85,707	232	953	0.280	0.256	0.267	211
50	71,454	333	1,142	0.265	0.307	0.285	224
40	64,311	334	1,350	0.266	0.363	0.307	228
30	52,874	336	1,717	0.269	0.462	0.340	236
20	42,463	332	2,079	0.277	0.560	0.371	239
10	38,634	330	2,107	**0.319**	**0.567**	**0.408**	**240**

4.1. Clasificación de usuarios

Para la generación del *training* de la segunda etapa se utiliza el mismo conjunto inicial de conversaciones, pero eliminando aquellas que son muy cortas, que contienen muchos caracteres no imprimibles y en donde sólo exista un usuario.

Se obtuvo un conjunto de 15,374 conversaciones, las cuales se dividieron en dos categorías de diálogos: "predator" y "no-predator". Además, se aplicó el preprocesamiento y la expansión de términos utilizados en el conjunto inicial de conversaciones obteniendo un *training* de 26,472 diálogos.

Para la creación del *test*, se crea un conjunto según las conversaciones recuperadas en la fase uno. Se fusionaron los 2 conjuntos de conversaciones que contienen mas depredadores sexuales (240 y 226 depredadores), es decir las conversaciones recuperadas con el sistema de voto utilizando el conjunto de características "Léxicas Categorías" y las conversaciones recuperadas del sistema de consultas con un "TOPE" igual a 10.

Como se puede observar, ninguno de los conjuntos citados contiene los 250 depredadores del *test* original, ya que con la unión de los conjuntos sólo se llegó a 246 depredadores, por lo que se analizaron de manera manual todas las conversaciones donde participan los 6 depredadores faltantes. De manera general, se observó que éstos depredadores participan en mas de una conversación, sin embargo, son los únicos usuarios dentro de cada conversación y los diálogos que escriben son muy cortos y con demasiados emoticones y URLs. Dadas estas características, las conversaciones de estos usuarios no se asemejan a las conversaciones de depredadores incluidas en el *training*, donde generalmente hay mas diálogos y por lo menos son dos usuarios por conversación.

Para esta etapa, los resultados del sistema de consulta fueron demasiado bajos, por lo que se omiten en el análisis, sin embargo, el tamaño del *training* es adecuado para extraer las subestructuras utilizando la herramienta de *SUBDUE*. Se utilizaron las mismas combinaciones de características con los mismos clasificadores de la fase 1, pero ahora para los nuevos conjuntos resultantes de dicha fase. En la gráfica 4 se muestra solamente el *F-score* de todos los experimentos realizados, se omiten las otras métricas por que al ser la última etapa, se necesita recuperar a los usuarios que realmente sean depredadorese sexuales, pero sin que tengan falsos positivos, esto se logra analizando los resultados de esta métrica.

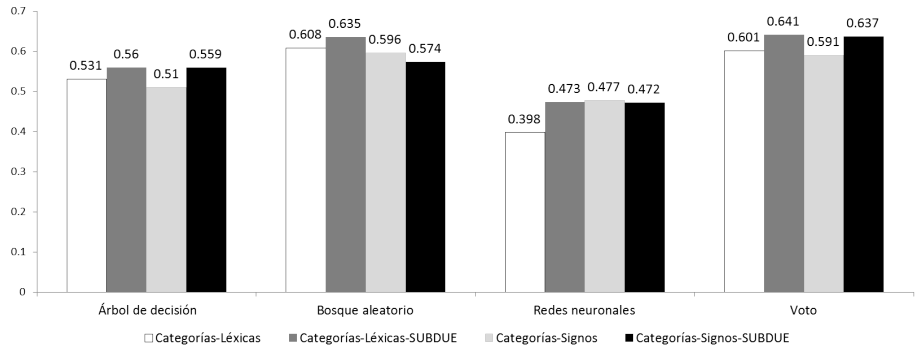

Fig. 4. *F-Score* para los conjuntos de características utilizados en la fase uno.

El *F-score* mas alto se obtiene con el algoritmo de bosque aleatorio utilizando los conjuntos de características léxicas, categorías gramaticales y la herramienta *SUBDUE*. Además, se observa en la gráfica que el clasificador de redes neuronales generó el *F-score* mas bajo para todos los conjuntos de características, incluso al combinarlas con *SUBDUE*. Algo interesante es que los resultados no siguen el mismo patrón que en la fase anterior, donde uno de los mejores *F-score* es reportado con el uso de redes neuronales, en este caso, se puede afirmar que en esta fase dicho algoritmo entorpece los resultados finales..

El sistema de voto demuestra que tiene mejor exactitud a la hora de detectar a los depredadores sexuales de las víctimas, incluso con el ruido que puede tener el uso de redes neuronales como parte del sistema. Al final de los experimentos, se logra un *F-score* de 64.1 % en la detección de los depredadores sexuales utilizando los conjuntos de características léxicas, conteos de categorías gramaticales y las palabras extraídas con la herramienta *SUBDUE*.

5. Conclusiones y trabajo futuro

De los experimentos reportados en esta investigación, se puede concluir lo siguiente:

- El *training* y especialmente el *test* tienen demasiados elementos propios de chats. Con la utilización de los tres diccionarios creados se logra extender los textos y agregarle significado a algunas palabras que no lo tienen, esto hace que el vocabulario se reduzca considerablemente y que los etiquetadores utilizados en fases posteriores tengan un mejor comportamiento. Sin embargo, actualmente los chats utilizan un conjunto mucho mayor de símbolos, palabras e incluso imágenes, las cuales se insertan en la conversación tecleando un código numérico (esto ocurre especialmente en *facebook*), por lo que a pesar de tener una metodología para el preprocesamiento de los datos, esta no logra enriquecer todos los términos utilizados en las conversaciones actuales.
- En la primera fase se experimenta con varios conjuntos de características y técnicas de recuperación de información. Los resultados sirven como un filtro para las conversaciones, de tal manera que en la fase de clasificación de usuarios exista un balance entre las conversaciones de depredadores y las conversaciones de otras personas. Las complicaciones en esta fase radican en la magnitud del corpus, así como en el desbalanceo de las clases. Además, se observa que los conjuntos de características utilizados en esta etapa de la investigación no fueron suficientes para obtener buenos resultados en la segunda etapa.
- En la fase de clasificación de usuarios, el *F-score* obtenido significa un incremento notable respecto a los resultados obtenidos anteriormente para esta misma tarea, sin embargo, aún no supera los resultados obtenidos a nivel mundial.
- A pesar de que los resultados no superan el estado del arte mundial, se considera que la metodología propuesta aporta una nueva visión para tratar esta problemática y otras similares como la detección del género (donde ya se han logrado resultados aceptables con los mismos conjuntos de características) y tareas que involucren textos de redes sociales.
- El sistema de consulta ofrece la oportunidad de clasificar las conversaciones tomando en cuenta el vocabulario usado, además, necesita menos tiempo y recurso computacional para ejecutarse, aunque sus resultados con muy bajos, logra reducir drásticamente el conjunto de conversaciones, manteniendo las que son positivas, es decir, en las que participa por lo menos un depredador sexual.

Como trabajo futuro se pretende incluir los siguientes aspectos:

- Incrementar los diccionarios creados a fin de reforzar el significado de los textos. Se considera agregar mas elementos como emoticonos compuestos, tomar en cuenta la repetición de las letras (por ejemplo que las palabras "hello" y "hellloo" sean consideradas iguales), analizar errores gramaticales, sinónimos y búsqueda de analogías en los diálogos. Esto ayudará también a la reducción de vocabulario y por consecuencia, a la reducción en los tiempos de creación de modelos y clasificación.
- La extracción de nuevas características semánticas que no hayan sido consideradas en este trabajo tomando en cuenta otros aspectos de la conversación

como la hora, tiempo de respuesta, relación entre los conceptos utilizados y las fechas en que se dieron dichas conversaciones.
- Expandir el análisis de los textos mediante un estudio utilizando el cambio de parámetros de los clasificadores utilizados a fin de encontrar la combinación óptima para el incremento del *f-score*.

Además se está trabajando en un estudio comparativo utilizando la misma metodología aquí presentada en otros corpus a fin de estudiar el comportamiento de los conjuntos de características utilizados en cuanto a las métricas de precisión y *F-score*.

Referencias

1. Ayala, D.V., Castillo, E., Pinto, D., Olmos, I., León, S.: Information retrieval and classification based approaches for the sexual predator identification. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF (Online Working Notes/Labs/Workshop) (2012)
2. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (Oct 2001)
3. 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), 10–18 (Nov 2009)
4. Harms, C.M.: Grooming: An operational definition and coding scheme 8(1), 1–6 (2007)
5. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA (1999)
6. Mitchell, T.M.: Machine Learning. McGraw-Hill, Inc., New York, NY, USA, 1 edn. (1997)
7. Olmos, I., Gonzalez, J.A., Osorio, M.: Subgraph isomorphism detection using a code based representation. In: Russell, I., Markov, Z. (eds.) FLAIRS Conference. pp. 474–479. AAAI Press (2005)
8. Pendar, N.: Toward spotting the pedophile telling victim from predator in text chats. In: Proceedings of the International Conference on Semantic Computing. pp. 235–241. ICSC '07, IEEE Computer Society, Washington, DC, USA (2007)
9. Quinlan, J.R.: C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning). Morgan Kaufmann, 1 edn. (Oct 1992)
10. Symens, B.: Acronyms Dictionary for Texting Chatting E-mail. Rebecca J Symens
11. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Human Language Technology Conference (HLT-NAACL 2003) (2003)
12. Villatoro-Tello, E., Juárez-González, A., Escalante, H.J., Montes-y-Gómez, M., Pineda, L.V.: A two-step approach for effective detection of misbehaving users in chats. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
13. Yan, X., Han, J.: gspan: Graph-based substructure pattern mining. In: ICDM. pp. 721–724 (2002)