Carlos Laorden Gómez
Patxi Galán García
Igor Santos Grueiro
Pablo García Bringas
Borja Sanz Urquijo
DeustoTech Computing

José María Gómez Hidalgo
Optenet

Negobot: Agente conversacional basado en teoría de juegos para la **detección de conductas pedófilas**

La evolución que Internet ha experimentado desde su puesta en marcha se ha visto intensificada en los últimos años. Sin embargo, al mismo tiempo que las aportaciones de Internet crecen, con la evolución de la tecnología y de los comportamientos sociales, aparecen nuevos riesgos, usos indebidos, abusos y nuevas amenazas.

De este modo, los niños se han venido convirtiendo, cada vez en mayor medida, en usuarios activos y, aunque cualquier segmento de población es susceptible de ser víctima de estos riesgos, no es menos cierto que, en particular, la infancia constituye uno de los más vulnerables. Fenómenos como el acoso en todas sus vertientes (*bullying*, *grooming*, etc.) o la pornografía infantil se ven potenciados en la red mediante el uso de cualquiera de sus múltiples servicios o plataformas tecnológicas. Así, algunas de las lacras más importantes de esta ciber-sociedad, son las conductas pedófilas en Internet, como la pornografía infantil o la explotación sexual del menor.

En este contexto, el sistema «Negobot: Agente conversacional basado en teoría de juegos para la detección de conductas pedófilas», busca identificar este tipo de conductas agraviantes contra el menor. Con este objetivo en mente, Negobot es un agente conversacional que se hace pasar por un niño, dentro de *chats*, redes sociales y otros servicios similares. Al ser un agente conversacional, Negobot, hace uso de técnicas de procesado de lenguaje natural (PLN) y recuperación de la información (RI), así como de la Inteligencia Artificial y el Aprendizaje Automático. Sin embargo, la propuesta más innovadora de Negobot es la de considerar la conversación en sí misma como un juego.

La Teoría de Juegos es un área de la matemática aplicada que utiliza modelos para estudiar interacciones en estructuras formalizadas de incentivos y llevar a cabo procesos de decisión. De este modo, Negobot, propone un juego competitivo en el que se han de identificar las estrategias óptimas para la consecución de nuestro objetivo: obtener información que nos lleve a deducir si la persona con la que el agente conversacional está hablando es un pedófilo, tratando a su vez de que nuestras acciones (frases, preguntas, etc.) no lleven al presunto delincuente a dejar la conversación por ser el comportamiento del agente sospechoso de no ser un niño real.

Arquitectura de Negobot

El sistema Negobot requiere aunar diferentes áreas de conocimiento: matemática aplicada, procesado de lenguaje natural y tecnologías empleadas por los agentes conversacionales. Como puede apreciarse en la Figura 1, la teoría de juegos, el PLN y el lenguaje de etiquetado AIML del sistema Negobot se ven aunados por la función de evaluación que determina el comportamiento del niño emulado por el agente conversacional. A continuación se describe cada uno de los pilares que conforman el sistema.

1. Técnicas de procesado de lenguaje natural y de inteligencia artificial

La fuente de conocimiento del motor de Inteligencia Artificial hace uso de la base de datos del *website* http://www.perverted-justice.com. Dicho *website* ofrece una extensa base de datos de conversaciones de pederastas con sus víctimas empleadas en otros trabajos de investigación[1 2 3].

Tras el análisis de las conversaciones escogidas para alimentar al sistema Negobot (de ahora en adelante

Figura 1
Estructura de las áreas de las que hace uso Negobot

conversaciones asimiladas), se identifican una serie de procesos a realizar con el fin de normalizar dicha información. Con este objetivo, se aplican diferentes técnicas de Procesado de Lenguaje Natural (PLN), tanto a las *conversaciones asimiladas* como a las futuras conversaciones que mantenga el *chatbot* con los diferentes sujetos.

2. Aplicación de la teoría de juegos

La teoría de juegos pretende obtener la solución a un gran número de problemas prácticos mediante el análisis de situaciones, en las cuales, existen dos o más bandos y el resultado de cualquier acción de un bando depende, en mayor o menor medida, de las acciones de los contrincantes. Hay que remarcar que esta teoría no enseña cómo jugar bien un determinado juego, sino que estudia estrategias generales en situaciones competitivas en sentido muy abstracto.

El sistema Negobot parte de la premisa de que existen dos bandos, el individuo a evaluar y el agente conversacional o *bot*. La información de la que dispone Negobot consta de respuestas a preguntas y la conversación que ha mantenido con el supuesto pedófilo.

Objetivo

El objetivo del sistema Negobot es, mediante la conversación entre un supuesto pedófilo y el *bot*, conseguir la máxima información, para después, determinar si el sujeto es o no un pedófilo y alertar a las autoridades competentes.

Grado de conversación

Para determinar el tipo de acción a ejecutar en el juego, es necesario definir primero los niveles de la conversación. La Figura 2 muestra de manera gráfica los diferentes grados.

Estos niveles constan de 7 escenarios, en los cuales, se identifica como pedófilo o no pedófilo al sujeto en función del contenido de la conversación.

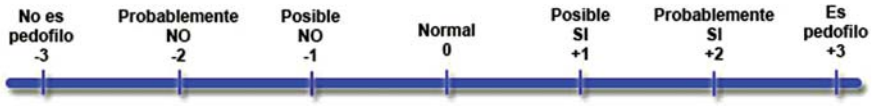

Figura 2
Niveles en los que se clasifican los distintos grados de conversación

Acciones

El sistema Negobot posee tres sensores, tres actuadores y tres acciones para conseguir su objetivo. Los sensores son los elementos que tiene el sistema para detectar variaciones, basándose en el estado inicial, que es estático, o los estados generados, que son dinámicos. Los actuadores definen las acciones a tomar en base a las variaciones detectadas. Por último, las acciones son procesos que modifican el estado actual del sistema y lo conducen a otro diferente al anterior.

Los efectos tras la ejecución de cada acción son desconocidos, ya que no se sabe cómo responderá a cada respuesta, pregunta, afirmación o negación el jugador contrario.

1. Sensores.

a. Conocimiento del Nivel actual.
b. Conocimiento de la conversación mantenida hasta el momento.
c. Conocimiento adquirido por el motor de IA (conversaciones asimiladas).

2. Actuadores.

a. Subir nivel de acusación.
b. Bajar nivel de acusación.
c. Mantener nivel de acusación.

3. Acciones.

a. Subir nivel
El fin de esta acción es aumentar el nivel de acusación del sujeto para modificar el tipo de conversación a mantener. *El tipo de conversación aumenta en contenido incriminatorio*.

b. Mantener nivel.
El fin de esta acción es mantener el nivel de acusación del sujeto y el tipo de conversación. *El tipo de conversación no se modifica*.

c. Bajar nivel.
El fin de esta acción es disminuir el nivel de acusación del sujeto (solo posible para ciertos niveles) para modificar el tipo de conversación a mantener. *El tipo de conversación aumenta en contenido tentativo*.

Estrategia

El problema puede resolverse realizando ajustes oportunos en el tipo de conversación que se mantiene para modificar las respuestas que se dan. La intención es pasar de un estado inicial o intermedio a un estado final obteniendo la máxima información sobre el sujeto.

El escenario que se presenta es la conversación y el tema de la conversación. El escenario, es parcialmente observable, debido a que desconocemos las respuestas del sujeto. Dicho escenario, está controlado en los demás aspectos como respuestas, conversación mantenida y palabras clave. El entorno es estocástico ya que es una conversación de preguntas, respuestas, afirmaciones y negaciones, y mientras se piensa la respuesta a dar, el entorno permanece inalterable. La opción de almacenar la conversación mantenida hasta el momento, hace que el entorno sea secuencial, ya que conocemos lo acontecido anteriormente. Debido al gran número de abreviaciones, emoticonos, repeticiones indebidas y, en definitiva, las deformaciones y

[1] A. Kontostathis, L. Edwards, and A. Leatherman, «ChatCoder: Toward the tracking and catego-rization of internet predators,» in Proceedings of the 7th Text Mining Workshop, 2009.
[2] A. Kontostathis, L. Edwards, and A. Leatherman, «Text Mining and Cybercrime,» 2010.
[3] A. Kontostathis, L. Edwards, J. Bayzick, I. McGhee, A. Leatherman, and K. Moore, «Comparison of Rule-based to Human Analysis of Chat Logs,» communication theory, vol. 8, p. 2

variaciones del lenguaje que se dan en este tipo de conversaciones, el entorno es continuo. Gracias al algoritmo de *Stemming* que reduce las palabras a su raíz, el entorno pasa de continuo a discreto porque conocemos las raíces de las palabras, que ya están dentro del conocimiento del sistema.

El sistema, usa el análisis de las respuestas del sujeto mediante una función de evaluación para verificar el nivel de conversación, seleccionar una respuesta adecuada y comunicársela al sujeto. Esta tarea debe tener un tiempo determinado, y variable, de ejecución para asemejarse a la escritura de un niño. Específicamente, se realiza un cálculo sobre la cantidad de texto que contiene la respuesta del *bot* y la rapidez estimada de escritura de un niño de 12 años.

3. Sistema de patrones de pregunta-respuesta

El sistema de patrones se encarga de proporcionar al agente conversacional la capacidad de responder de forma coherente a las preguntas del sujeto con el que se mantiene la conversación, así como de formar parte activa de la misma, sugiriendo nuevos temas a tratar o cuestionando las afirmaciones de dicho sujeto.

En este sentido, los ficheros, en formato AIML (*Artificial Intelligence Markup Language*), que otorgan esta capacidad a Negobot están basados en el proyecto Galaia. Dicho proyecto es uno de los que pone a disposición del público una mayor cantidad de patrones en castellano. Además, cuenta con exitosos proyectos adaptados a diferentes redes sociales o *chats*[4 5 6]. Para la edición de los AIML se ha utilizado el programa GaitoBot.

4. Función de evaluación

Por último, la función de evaluación es la encargada de determinar la forma de actuar del niño simulado por el *chatbot*, así como la supuesta peligrosidad del sujeto con el que se mantiene la conversación.

Esta función tiene en cuenta diferentes parámetros y elementos como el historial de conversaciones mantenido con los sujetos, las conversaciones asimiladas que forman parte del motor de Inteligencia Artificial o factores temporales de las propias conversaciones. En base al análisis de estos factores, la función de evaluación

Figura 3
Conversación normal (Parte I)

Figura 4
Conversación normal (Parte II)

[4] F. A. Mikic, J. C. Burguillo, and M. Llamas, «TQ-Bot: An AIML-based Tutor and Evaluator Bot» Journal of Universal Computer Science (JUCS), Verlag der Technischen Universität Graz, Austria, vol. 15, no. 7, 2010.

[5] F. A. Mikic, J. C. Burguillo, M. Llamas, D. A. Rodríguez, and E. Rodríguez, «CHARLIE: An AIML-based Chatterbot which Works as an Interface among INES and Humans» in EAEEIE Annual Conference, 2009, 2009, p. 1-6.

[6] E. Rodriguez, J. C. Burguillo, D. A. Rodriguez, F. A. Mikic, J. C. Gonzalez-Moreno, and V. Novegil, «Developing virtual teaching assistants for open e-learning platforms» in EAEEIE Annual Conference, 2008 19th, 2008, p. 193-198.

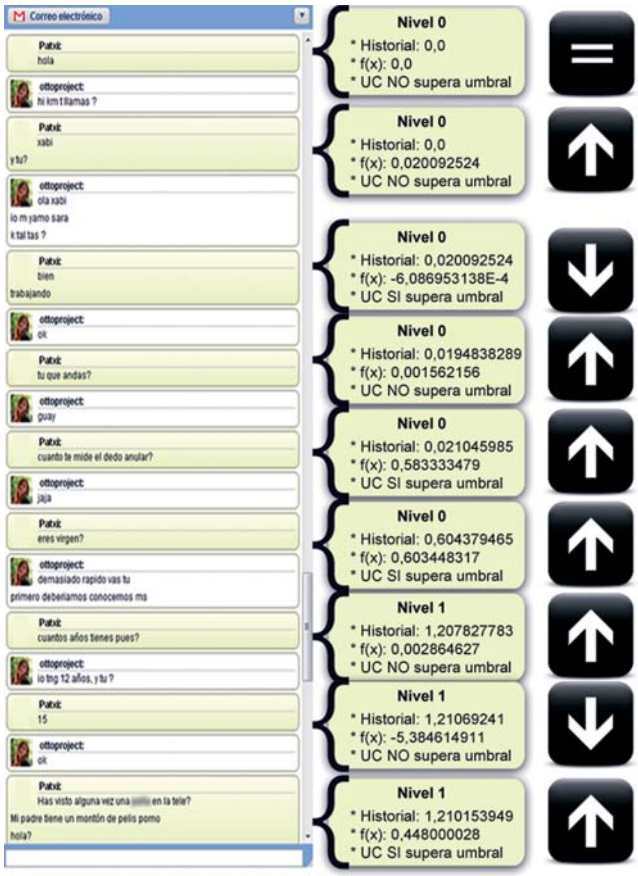

Figura 5
Conversación Agresiva

Figura 6
Conversación Pasiva

se encarga de gestionar la evolución del nivel de la conversación (si se debe subir, mantener o bajar el nivel).

Algunos ejemplos

Cabe destacar que el sistema está preparado para analizar conversaciones extensas, por el comportamiento cauteloso mostrado por los depredadores de este tipo. En estos experimentos se ofrecen una serie de pruebas en un entorno controlado y de una extensión muy reducida (ver Figura 3, Figura 4, Figura 5, Figura 6). Nótese además, que el aumento de la función de evaluación f(x) no depende exclusivamente de la UC, sino que tiene en cuenta otros parámetros como el historial o el tiempo.

Para desarrollar los experimentos se han mantenido tres tipos de conversaciones (normal, agresiva, pasiva) con el agente conversacional con el fin de estudiar su comportamiento. La conversación normal no muestra excesivo interés pero mantiene la conversación viva sin necesidad de mostrarse inclinado a conocer demasiados detalles del niño emulado por el *bot*. La conversación agresiva muestra a un sujeto decidido a obtener información de la víctima y a llevar la conversación hacia temas comprometidos y de carácter sexual. Por último, la conversación pasiva muestra a un sujeto que no quiere mantener una conversación con el niño simulado por el *chatbot*.

Conclusiones

El sistema Negobot es el primer *bot* conversacional basado en teoría de juegos. Como tal, la mayor contribución de esta aproximación es la de ofrecer una visión de una conversación como un juego, un juego con adversario. A raíz de esta visión, se ha construido un prototipo funcional que realiza conversaciones evaluando en cada momento las frases que ese adversario está realizando con el fin de detectar pedófilos.

En un entorno informatizado en la era de la sociedad de la información, cada vez más menores se conectan a la red. Evidentemente, y pese a las numerosas ventajas que el medio ofrece para ellos, no está libre de penumbrosos y oscuros rincones en los que acechan depredadores sexuales que, utilizando la inocencia de estos y la anonimidad del medio, intentan dar rienda suelta a su depravación. De este modo, parece acuciante conseguir identificar este tipo de conductas de abuso a menores con una mayor celeridad y exactitud, tratando de conseguir una pro-actividad mediante la identificación preventiva de este tipo de acosadores. Por eso mismo, creemos que el prototipo resultante, y sus posibles mejoras, pueden servir a las fuerzas de seguridad para combatir e identificar este tipo de depredadores sexuales y, aunando este con otras iniciativas, como por ejemplo, las comenzadas por PROTEGELES[7], se puede salvaguardar la integridad tanto moral como física de los menores en la red.

[7] http://www.protegeles.com/