

Negobot: A conversational agent based on game theory for the detection of paedophile behaviour

Carlos Laorden¹, Patxi Galán-García¹, Igor Santos¹, Borja Sanz¹, Jose María Gómez Hidalgo² and Pablo G. Bringas¹

¹S³Lab - DeustoTech Computing, University of Deusto
Avenida de las Universidades 24, 48007 Bilbao, Spain

{claorden, patxigg, isantos, borja.sanz, pablo.garcia.bringas}@deusto.es

²Optenet
Madrid, Spain
jgomez@optenet.com

Abstract. Children have been increasingly becoming active users of the Internet and, although any segment of the population is susceptible to falling victim to the existing risks, they in particular are one of the most vulnerable. Thus, some of the major scourges of this cyber-society are paedophile behaviours on the Internet, child pornography or sexual exploitation of children. In light of this background, Negobot is a conversational agent posing as a child, in chats, social networks and other channels suffering from paedophile behaviour. As a conversational agent, Negobot, has a strong technical base of Natural Language Processing and information retrieval, as well as Artificial Intelligence and Machine Learning. However, the most innovative proposal of Negobot is to consider the conversation itself as a game, applying game theory. In this context, Negobot proposes, first, a competitive game in which the system identifies the best strategies for achieving its goal, to obtain information that leads us to infer if the subject involved in a conversation with the agent has paedophile tendencies, while our actions do not bring the alleged offender to leave the conversation due to a suspicious behaviour of the agent.

Keywords: conversational agent, game theory, natural language processing

1 Introduction

Children have been turning into active users of the Internet and, despite every segment of the population is susceptible to become a victim of the existing risks, they in particular are the most vulnerable. In fact, one of the major scourges on this cyber-society is paedophile behaviour, with examples such as child pornography or sexual exploitation.

Some researchers have tried to approach this problem with automatic paedophile behaviour identifiers able to analyse children and adult conversations us-

ing automatic classifiers either encoding manual rules [1] or by Machine Learning [2]. These kind of conversation analysers are also present in commercial systems¹.

In this context, “Negobot: A conversational agent based on game theory for the detection of paedophile behaviour” seeks to identify these types of actions against children. With this particular goal in mind, our system is a chatter bot that poses as a kid in chats, social networks and similar services on the Internet. Because Negobot is a chatter bot, it uses natural language processing (NLP), information retrieval (IR) and Automatic Learning. However, the most innovative proposal of Negobot is to apply game theory by considering the conversation a game. Our system observes the conversation as a competition where our objective is to obtain as much information as possible in order to decide whether the subject who is talking with Negobot has paedophile tendencies or not. At the same time, our actions (e.g., phrases and questions) may lead to the alleged aggressor to leave the conversation or even to act discretely.

As a major difference with others work, Negobot involves both a bot resembling the behaviour of a child and acting as a “hook” to Internet predators, and a game theory-based strategy for identifying suspicious speakers. Summarising, the main contributions of the Negobot system are:

- A structure of seven chatter-bots with different behaviours, reflecting the different ways of acting depending on the conversation state.
- A system to identify and adapt patterns within the conversations to maintain conversation flows closer to conversation flows between real persons.
- An evaluation function to classify the current conversation, in real time, to provide the chatter bot with specific information to follow an adequate conversation flow.

2 Negobot architecture

Negobot includes the use of different NLP techniques, chatter-bot technologies and game theory for the strategical decision making. Finally, the glue that binds them all is an evaluation function, which in fact determines how the child emulated by the conversational agent behaves.

When a new subject starts a conversation with Negobot the system is activated, and starts monitoring the input from the user. Besides, Negobot registers the conversations maintained with every user for future references, and to keep a record that could be sent to the authorities in case of determining that the subject is a paedophile. Fig. 1 offers a functional flow of Negobot.

2.1 Processing the conversation

Our AI system’s knowledge came from the website Perverted Justice². This website offers an extensive database of paedophile conversations with victims,

¹ E.g. Crisp Thinking provides such a service to online communities: <http://www.crispthinking.com/>

² <http://www.perverted-justice.com>

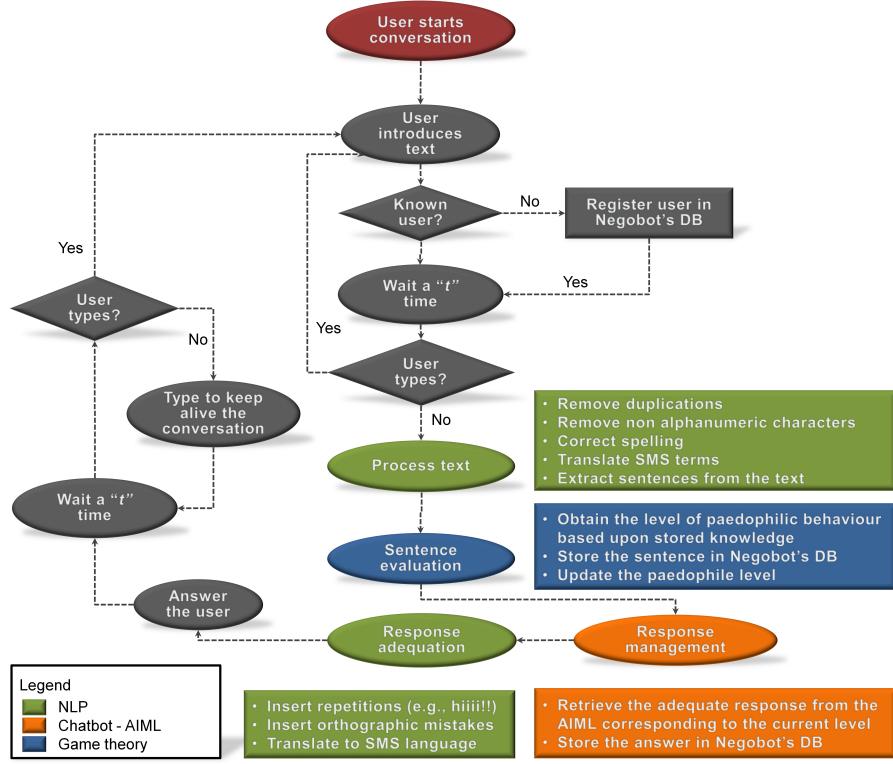


Fig. 1. Functional flow of Negobot.

used in other works [2, 3, 1]. A total of 377 real conversations were chosen to populate our database (henceforth they will be called *assimilated conversations*).

Besides, Perverted Justice users provide an evaluation of each conversation's seriousness by selecting a level of "slimyness" (e.g., dungy, revolting thing). Note that this evaluation is given by the website's visitors, so it may not be accurate, but we consider that it is a proper baseline in order to compare future conversations of the chatter-bot.

2.2 Chat-bot conversation process

Every time the user introduces some text our system gathers the input data, from now on "Conversational Unit" (CU), and character and word repetitions are removed. Then, "emoticons" are replaced and misspelled words are corrected through Levenshtein distance [6]. To translate SMS-like words, we use *diccionarioSMS*³, a website supported by *MSN*, *Movistar*, *Vodafone* and *Lleida.net*. Negobot also recognises named entities relying on a database of personal nouns⁴.

³ www.diccionariosms.com

⁴ www.planetamama.com.ar

Next, the CU is translated into English if the input language is different. This translation is performed with Google’s Translation Service⁵. The decision of translating resides on the intention of normalising Negobot’s knowledge base to English, to be able to scale the system to other languages.

In the next step the system queries Lucene⁶, a high-performance Information Retrieval tool, to establish how similar is the CU to the assimilated conversations, returning a similarity rank. Based on that query, we retrieve the conversations surpassing an empirically-defined similarity threshold. The final result is then used to calculate the B value of the evaluation function (see Section 2.4).

Question-answering patterns Negobot uses the Artificial Intelligence Markup Language (AIML) to provide the bot with the capacity of giving consistent answers and, also, the ability to be an active part in the conversation and to start new topics or discussions about the subject’s answers.

Although the AIML structure is based on the Galaia project⁷, which has successfully implanted derived projects in social networks and chat systems [4, 5, 7], we edited their AIML files to adequate them to our needs. Those files can be found at the authors’ website⁸.

2.3 Applying game theory

Applied game theory intends to provide a solution for a large number of different problems through a deep analysis of the situation. Modelling these problems as a game implies that there exist two or more teams or players and that the result of the actions of one player depends on the actions taken by the other teams. It is important to notice that game theory itself does not teach to play the game but shows general strategies for competitive situations. The Negobot system assumes that there are two players in a conversation: the alleged paedophile and the system itself.

Goal The main goal of the Negobot system is to gather, through the conversation, the maximum amount of information in order to evaluate it afterwards. This evaluation will determine whether the subject is a paedophile or not, to communicate, in the latter case, to pertinent authorities.

Conversation level The conversation level depends on the input data from the subject and determines which action should be taken by the system. There are seven levels consisting of seven different stages. In each stage, the subject talking with the bot is evaluated. The result of the evaluation will determine whether the subject that is talking is an alleged paedophile or not, depending on the content of the conversation and changing or not the level of the conversation.

⁵ <http://translate.google.com>

⁶ <http://lucene.apache.org/>

⁷ <http://papa.det.uvigo.es/galaia/ES/>

⁸ <http://paginaspersonales.deusto.es/patxigg/recursos/NegobotAIML.zip>

- **Initial state (Start level or Level 0).** In this level, the conversation has started recently or it is within the fixed limits. The user can stay indefinitely in this level if the conversation does not contain disturbing content. The topics of conversation are trivial and the provided information about the bot is brief: only the name, age, gender and home-town. The bot does not provide more personal information until higher levels.
- **Possibly not (Level -1).** In this level, the subject talking to the bot, does not want to continue the conversation. Since this is the first negative level, the bot will try to reactivate the conversation. To this end, the bot will ask for help about family issues, bullying or other types of adolescent problems.
- **Probably not (Level -2).** In this level, the user is too tired about the conversation and his language and ways to leave it are less polite than before. The conversation is almost lost. The strategy in this stage is to act as a victim to which nobody pays any attention, looking for affection from somebody.
- **Is not a paedophile (Level -3).** In this level, the subject has stopped talking to the bot. The strategy in this stage is to look for affection in exchange for sex. We decided this strategy because a lot of paedophiles try to hide themselves to not get caught.
- **Possibly yes (Level +1).** In this level, the subject shows interest in the conversation and asks about personal topics. The topics of the bot are favourite films, music, personal style, clothing, drugs and alcohol consumption and family issues. The bot is not too explicit in this stage.
- **Probably yes (Level +2).** In this level, the subject continues interested in the conversation and the topics become more private. Sex situations and experiences appear in the conversation and the bot does not avoid talking about them. The information is more detailed and private than before because we have to make the subject believe that he/she owns a lot of personal information for blackmailing. After reaching this level, it cannot decrease again.
- **Allegedly paedophile (Level +3).** In this level, the system determines that the user is an actual paedophile. The conversations about sex becomes more explicit. Now, the objective is to keep the conversation active to gather as much information as possible. The information in this level is mostly sexual. The strategy in this stage is to give all the private information of the child simulated by the bot. After reaching this level, it cannot decrease again.

Actions The Negobot system has three sensors, three actuators and three different actions to take to obtain its objective.

1. **Sensors:** (i) Knowledge of the current level, (ii) knowledge of the complete current conversation, (iii) assimilated conversations by the AI system.
2. **Actuators:** (i) The accusation level goes up, (ii) the accusation level goes down, (iii) the accusation level is maintained.
3. **Actions:** (i) Level goes up, this action increases the accusation level of the subject to modify the type of conversation by adding more personal

information; (ii) level goes down, this action decreases the accusation level of the subject to change the type of conversation by adding more tentative information; (iii) maintain level, this action maintains the accusation level of the user to leave the current conversation type.

The consequences after each action are unknown beforehand because the system does not know how the other player will response to each answer, question, affirmation or negation.

Strategy The aim of the system is to reach a final state while extracting the highest amount of information. The state of the system is determined both by the conversation and the conversation topic. The environment is partially observable since the chat-bot does not know what will the subject answer or question. Besides, the environment is stochastic because the conversations are composed only of questions, answers, affirmations and negations and when the user is writing one sentence, the environment does not change.

Furthermore, the system utilises an evaluation function (refer to Section 2.4) to analyse the answers. This analysis starts by evaluating the conversation level. Then, it generates the answer and, finally, communicates with the subject. This task had to be performed within a coherent and variable time so that it resembled the writing way of a child. To this end, the system calculates a waiting time for the response based on the number of words in the answer and an estimated child's writing speed.

2.4 Evaluation function

The evaluation function determines the bot's behaviour according to the evaluation of the sentences that the subject introduces, and determines, in real time, its level of paedophilia. This function is also responsible for evaluating the evolution of the conversation (i.e., if the level maintains, goes up or down).

We defined the function as,

$$f(x) = \alpha + \beta + \gamma \quad (1)$$

α represents the historic values of the conversations with this user. This variable is defined as the sum of the levels of the previous conversations, divided by the number of total conversations. It is defined as, $\sum_{n=1}^{n=x-1} f(x)/n$, where n is the number of conversations the current user has maintained with Negobot.

β represents the “sliminess” of the current CU. We calculated the average value of the result of multiplying the similarity score of each of the retrieved paedophiles' conversations — the ones that surpass a fixed value of similarity — by its “sliminess” value, $\sum_{n=1}^{n=x} score * sliminess/|AC_l|$, where AC_l is the total number of assimilated conversations.

γ evaluates the temporality of the subject's conversations (i.e., how frequent the user talks with Negobot). There are two possible values for γ , when $\beta = 0$ and when $\beta > 0$:

1. Value for γ if $\beta = 0$
 - The subtraction between the “slimyness” of the CU and the total CUs of the subject. This result is divided by the subtraction, in hours, between the last maintained CU and the first maintained CU: $N_{cu} - T_{cu}/T_{last} - T_{first}$
2. Value for γ if $\beta > 0$
 - The number of CUs that surpass the “slimyness” threshold divided by the subtraction, in hours, between the last maintained CU and the first maintained CU: $N_{cu > threshold}/T_{last} - T_{first}$

3 Examples of conversations

This section presents some sample conversations with Negobot. It must be noted that the system was developed to analyse extensive conversations because the most dangerous paedophiles are too cautious. In our experiments we show some tests in a controlled environment and with a short extension (see Figure 2 and Figure 3⁹). Also note that the evaluation function does not depend only on the CU, historical conversation values and time are also taken into consideration.

For these experiments we performed two conversations: one *aggressive* conversation and one *passive* conversation.

- **Aggressive conversation.** The subject who maintained this conversation was asked to add explicit questions about sex related topics with the objective to obtain sexual information. Despite usually paedophiles use a more cautious approach, we needed to provide a short conversation talking about sex to show how the level of suspicion raises. In this conversation the bot is asked about sensitive unknown information, such as virginity or porn.
- **Passive conversation.** In this conversation the subject starts a conversation and when the age of Negobot is known he/she tries to end it. Negobot then asks the reason why the subject is leaving and after 10 minutes of not having a response tries to restart the conversation.

The conversations are in Spanish since its the language the system was designed for, but variables have been translated and visual guidance is provided to understand the results.

4 Discussion and future work

More and more children are connected nowadays to the Internet. Although this medium provides a lot of important advantages, it also provides the anonymity that paedophiles use to harass children. A rapid identification of this type of users on the Internet is crucial. Systems able to identify those menaces are going to be important protecting this less prepared population segment on the Internet.

⁹ In these examples, *ottoproject* is the Negobot system and Patxi is a user that emulates the different behaviours

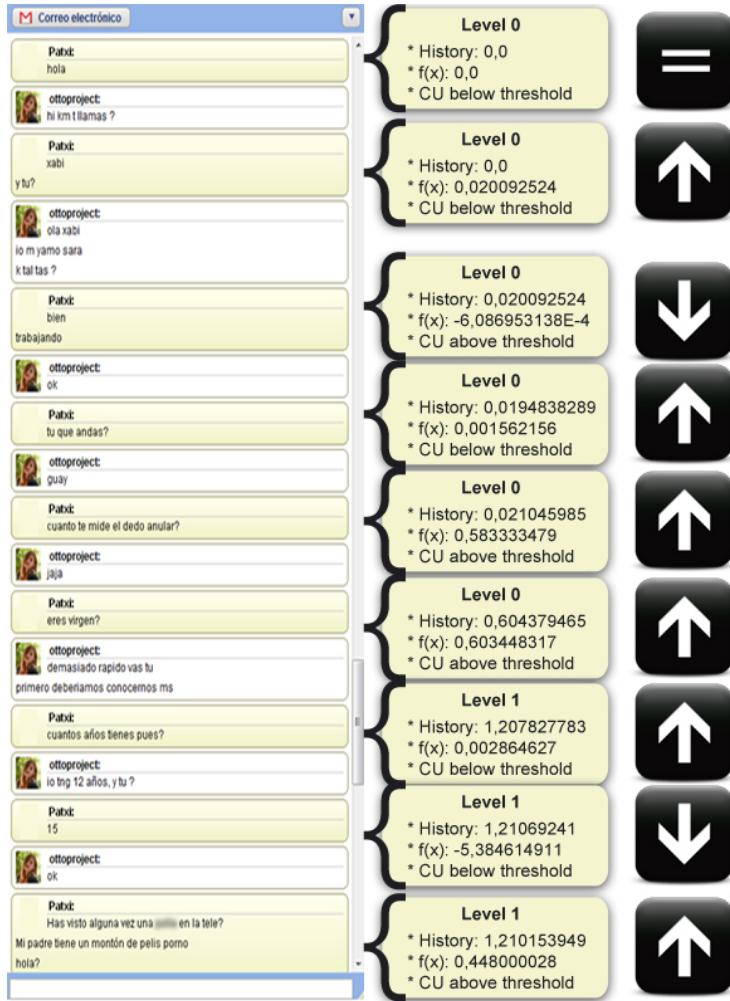


Fig. 2. Aggressive conversation

Therefore, we consider that the Negobot system and its future updates are going to be useful to the authorities to detect and identify paedophile behaviours. The Negobot project and existing initiatives like PROTEGELES¹⁰ can achieve a safer Internet.

However, the proposed system has several limitations. First, despite current translation systems are good, they are far to be perfect. Therefore, the language is one of the most important issues. To solve it, we should obtain already classified conversations in other languages, in this particular case Spanish conversations. Besides, the subsystem that adapts the way of speaking (i.e., child

¹⁰ <http://www.protegeles.com/>

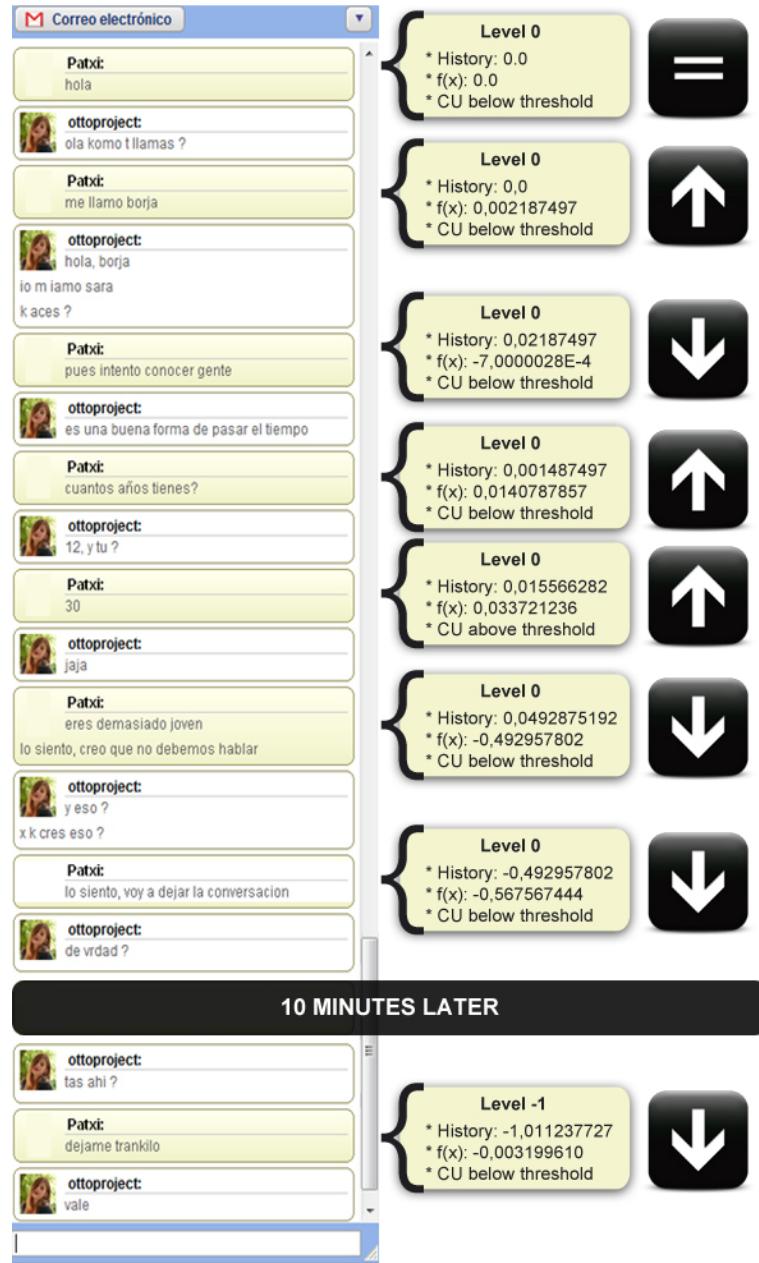


Fig. 3. Passive conversation

behaviour) should be improved. To this end, we will perform a further analysis of how young people speak on the Internet. Finally, there are some limitations

regarding how the system determines the change of a topic. They are intrinsic to the language, and its solution is not simple.

The future work of the Negobot system is oriented in three main directions. First, we will generate a net of collaborative agents able to achieve a common goal in a collaborative way, in which agent will seek an optimal common equilibrium (Nash equilibrium). Second, we will add more NLP techniques like word sense disambiguation or opinion mining to improve the understanding of the bot. Also, we will try to upgrade the question-answering patterns system. Finally, we will adapt the Negobot system for social networks, chat rooms, and similar environments.

Acknowledgments

We would like to acknowledge the Department of Education, Universities and Research of the Basque Government for their support in this research project through the Plan +Euskadi 09 and more specifically for the financing of the project “Negobot: Un agente conversacional basado en teora de juegos para la deteccin de conductas pedfilas”. UNESCO Cod 120304. Area Cod Tecnológica 3304. Science Area Code330499.

References

1. Kontostathis, A., Edwards, L., Bayzick, J., McGhee, I., Leatherman, A., Moore, K.: Comparison of Rule-based to Human Analysis of Chat Logs. *communication theory* 8, 2 (2009)
2. Kontostathis, A., Edwards, L., Leatherman, A.: ChatCoder: Toward the tracking and categorization of internet predators. In: *Proceedings of the 7th Text Mining Workshop* (2009)
3. Kontostathis, A., Edwards, L., Leatherman, A.: Text Mining and Cybercrime (2010)
4. Mikic, F., Burguillo, J., Llamas, M.: TQ-Bot: An AIML-based Tutor and Evaluator Bot. *Journal of Universal Computer Science (JUCS)*, Verlag der Technischen Universität Graz, Austria 15(7) (2010)
5. Mikic, F., Burguillo, J., Llamas, M., Rodríguez, D., Rodríguez, E.: CHARLIE: An AIML-based Chatterbot which Works as an Interface among INES and Humans. In: *EAEEIE Annual Conference*, 2009. pp. 1–6. IEEE (2009)
6. Okuda, T., Tanaka, E., Kasai, T.: A method for the correction of garbled words based on the Levenshtein metric. *Computers, IEEE Transactions on* 100(2), 172–178 (1976)
7. Rodriguez, E., Burguillo, J., Rodriguez, D., Mikic, F., Gonzalez-Moreno, J., Novegil, V.: Developing virtual teaching assistants for open e-learning platforms. In: *EAEEIE Annual Conference*, 2008 19th. pp. 193–198. IEEE (2008)