ESCUELA COLOMBIANA DE INGENIERÍA
JULIO GARAVITO

PROYECTO DE GRADO

Development of a system based on Artificial Intelligence able to interact with suspects of cybercrimes

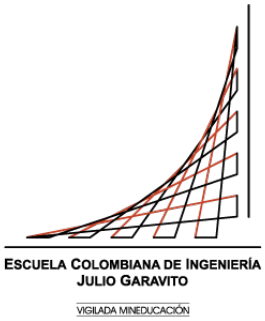
---

Autores:
John David IBÁÑEZ RODRÍGUEZ y
Santiago ROCHA DURÁN

Director:
Daniel Orlando DÍAZ LÓPEZ, *Ph.D.*

Programa de Ingeniería de Sistemas

BOGOTÁ, COLOMBIA



24 de julio de 2020

Índice general

Índice de figuras

Índice de cuadros

Escuela Colombiana de Ingeniería
Julio Garavito

Resumen

Programa de Ingeniería de Sistemas

Estudiante de Ingeniería de Sistemas

Development of a system based on Artificial Intelligence able to interact with suspects of cybercrimes

por John David IBÁÑEZ RODRÍGUEZ y
Santiago ROCHA DURÁN

Este libro presenta el desarrollo del proyecto de grado "Development of a system based on Artificial Intelligence able to interact withsuspects of cybercrimes". El libro contiene el desarrollo de la teoría investigada y la implementación realizada durante la asignatura Proyecto de Grado. Se compone de la investigación, implementación y mejora del chatbot C3-Sex, enfocandose en los dos modelos utilizados para interactuar con el sospechos (modelo basado en reglas y modelos retrieval), en la introducción de Snapchat como nueva plataforma para interactuar con el sospechos y en la mejora al clasificar el perfil del sospechoso. Estas mejoras permiten que la interacción con el sospechoso sea mucho mas eficaz permitiendo que el intercambio de contenido se de con muchas más frecuencia y a su vez que los datos recolectados de cada conversación sean analizados de mejor manera.

Capítulo 1

Introducción

El quinto dominio de la guerra (el ciberespacio) trae grandes retos para las agencias de seguridad del Estado, dentro de los que se cuenta la capacidad de anticiparse a los ciberdelitos para evitar impactos negativos a poblaciones vulnerables o incluso a infraestructura crítica nacional. Por otra parte, el advenimiento de la inteligencia artificial plantea múltiples alternativas para la automatización y mejora de los procesos, fundamentándose en aspectos como la predicción y prescripción, entre otros. Es por esto que se propone el presente proyecto que busca construir una solución basada en inteligencia artificial que permita apoyar las labores de ciberinteligencia que realizan actualmente las agencias de seguridad del estado, permitiendo consolidar información útil que de indicios sobre la ejecución de cibercrímenes y permita prevenir los delitos. Aunque existen trabajos relacionados, existen grandes retos de investigación que aun deben resolverse asociados principalmente a la capacidad de lograr una interacción efectiva con un sospechoso.

Línea de investigación a la que pertenece en CTG-Informática:

- Arquitectura tecnológica y seguridad
- Ingeniería de software

Proyectos de grado relacionados:

- Profiling criminals through Emotional Machine Learning
- Inteligencia de Fuentes Abiertas aplicadas al contexto Colombiano

Se prevee que los potenciales usuarios del presente proyecto pueden ser las unidades cibernéticas de la Policía Nacional de Colombia o de cada una de las Fuerzas Militares.

Capítulo 2

Cronograma

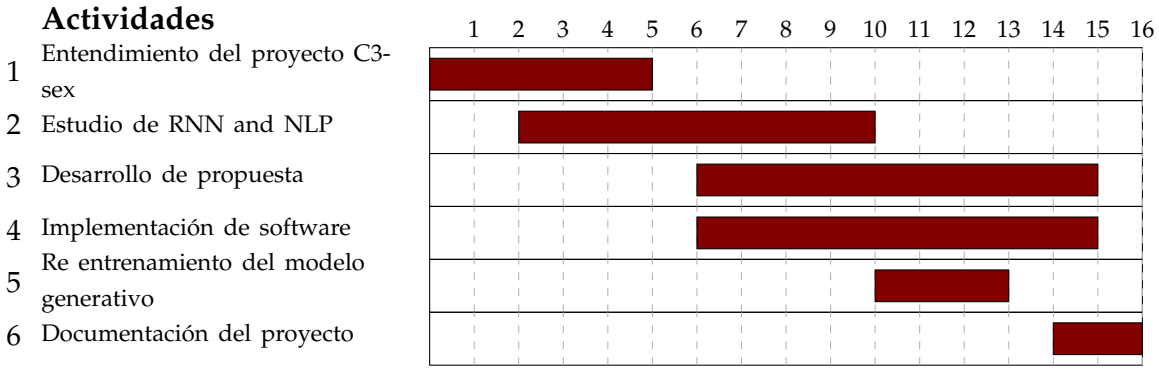


FIGURA 2.1: Diagrama Gantt de actividades de 1er periodo.

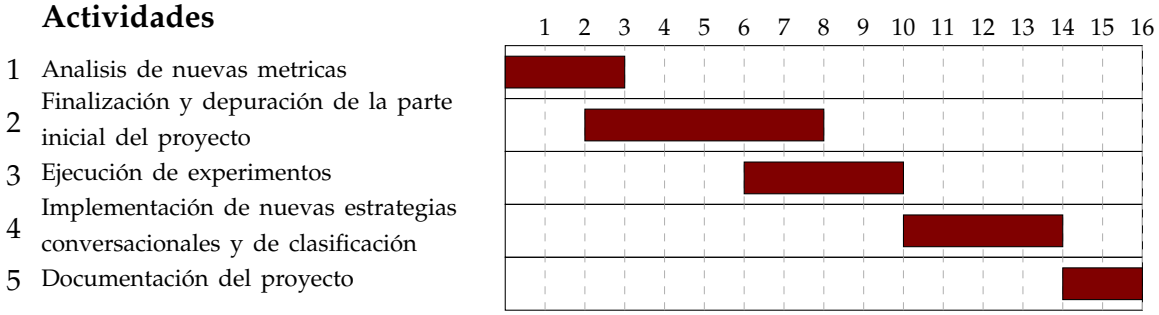


FIGURA 2.2: Diagrama Gantt de actividades de 2do periodo.

Para tener el conocimiento necesario sobre el tema, se realizaron la siguente actividades con el proposito de lograr la manipulación de android, mejorar el modelo generativo y mejorar las metricas ya existentes de la primera version de C^3-Sex

#	**Descripción**
1.1	Lectura del artículo "C3-Sex: a Chatbot to Chase Cyber perverts" Lectura del libro del proyecto de grado.
1.2	Descargar y ejecutar el proyecto disponible en github, Para lo anterior apoyarse en la documentación disponible en la página de github y en los anexos del proyecto disponibles en el repositorio.
1.3	Estudio del clasificador de polaridad de las frases (''Opinion Classification Model´´) a partir del siguiente vídeo: https://www.udemy.com/natural-language-processing-made-easy-using-python/learn/v4/t/lecture/11591238?start=15 .
1.4	Entendimiento del siguiente paper y el video sobre el clasificador de emociones: paper25.pdf , el video sobre el clasificador .
1.5	Ejecutar el modelo de clasificación de emociones (''Emotional Classification Model´´) y entender el código en R de la implementación. .
1.6	Realizar el siguiente ejercicio de redes neuronales para entender como funciona el modelo seq-to-seq (''Generative Model´´): Modelo seq to seq.
1.7	Revisar los videos del modulo ''Encoder-decoder-attention architecture´´ de la semana 4 del curso de NLP al igual que el ejercicio propuesto en ese modulo: https://www.coursera.org/learn/language-processing.
1.8	Estudiar y hacer los ejercicios de los 3 Módulos: Get Started with TensorFlow, Learn and use ML, sequences del tutorial de Tensorflow del siguiente link: https://www.tensorflow.org/tutorials Ejecutar el modulo seq-to-seq disponible en el repisitorio del proyecto C3-Sex.
1.9	Ejecutar todos los módulos al tiempo (''Generative Model´´, ''Retrieval based model´´, ''Opinion Classification Model´´, ''Emotional Classification Model´´) y validar el funcionamiento con Omegle..
2.1	Terminación y mejora de software desarrollado en el primer periodo.
2.2	Escritura de la primera versión paper.

CUADRO 2.1: Descripción del cronograma de actividades.

Capítulo 3

Definición del proyecto

3.1. Objetivo general

Construir una solución basada en inteligencia artificial capaz de interactuar con sospechosos de ciber crímenes con el fin de identificar actitudes sospechosas que conduzcan a la generación de indicios que pueda ser utilizado por agencias de seguridad del estado en investigaciones enfocadas a la prevención del delito.

3.2. Objetivos específicos

- Identificar un escenario de cibercrimen con características propicias (tipo de interlocutor, tipo de interacción, contenido intercambiado, mecanismo de transferencia, etc.) para ser abordado con una solución basada en inteligencia artificial.
- Utilizar modelos de inteligencia artificial para detectar e interpretar los inputs provenientes de un sospechoso
- Generar respuestas acordes a los inputs recibidos que permitan mantener la interacción y adquirir suficiente material para constituir indicios asociados a crímenes pasados o tentativos
- Analizar la interacción mantenida con el sospechoso para aportar a la categorización y perfilamiento por parte de personal vinculado a agencias de seguridad del estado que trabajen en la atención a ciber crímenes

3.3. Resultados esperados (entregables)

- Prototipo mejrado del agente conversacional C3-Sex basado en inteligencia artificial capaz de interactuar con sospechosos de cibercrimenes
- Libro de Proyecto de Grado con la documentación de la investigación realizada que incluya: introducción, marco teórico, estado del arte, propuesta, experimentos y conclusiones. Adicionalmente, el libro de proyecto de grado contiene como anexos: manual de instalación, manual de configuración y uso, código debidamente documentado.
- Artículo de Investigación que describe las contribuciones y presenta los resultados de la experimentación

3.4. Logros

- Comprensión: Revisión del estado del arte de modelos de machine learning y basados en conocimiento aplicables a chatbots.
- Re-Entrenamiento: Se re entrenó el modelo generativo usando una base de conocimientos que le permitiera responder de forma más dinámica a una conversación casual con un sospechoso
- Integración con Snpachat: Se diseñó y implementó una aplicación que permite la manipulación de Snapchat con el fin de tener un nuevo escenerio de cibercrimen y poder perfilar más cantidad de sospechosos.
- Análisis: Se agregaron 26 nuevas métricas de evaluación enfocadas en los metadatos de la conversación (extension, duración, intercambio de contenido, etc) con el fin de entender las tendencias que suelen tener los sospechosos para realizar un mejor perfilamiento de los ciber-criminales

Capítulo 4

Estado del arte

Hasta ahora se han realizado varios trabajos científicos en el campo de los chatbots. Así, por ejemplo, Gapanyuk et al. [1] proponen un modelo híbrido de chatbot compuesto por un módulo de preguntas y respuestas y un esquema basado en el conocimiento. El módulo de preguntas y respuestas contiene una lista de pares de preguntas y respuestas, de modo que cuando un usuario hace una pregunta que coincide con una de las listas, se le devuelve la respuesta correspondiente. La principal contribución de esta propuesta es la aplicación de un sistema basado en reglas que se encapsula en un metagráfico como agentes múltiples.

La mayoría de los primeros trabajos sobre sistemas de conversación se basan generalmente en el conocimiento y están diseñados para dominios específicos. Estos enfoques basados en el conocimiento no requieren que se construyan datos, sino que requieren mucho más esfuerzo manual y conocimientos expertos para construir el modelo, lo cual suele ser costoso. Así pues, [2] propone un modelo híbrido de chatbot de aprendizaje profundo que está basado en la generación. Esta propuesta se compone de 22 modelos de respuesta, que incluyen redes neuronales basadas en la recuperación, redes neuronales basadas en la generación, sistemas de preguntas y respuestas basados en el conocimiento y sistemas basados en plantillas. Además, desarrolla un módulo de aprendizaje de refuerzo basado en la estimación de un proceso de decisión de Markov.

La integración de un módulo emocional en los robots de charla es una de las formas de atraer a los usuarios, es decir, de dar al sistema de conversación la capacidad de ser amigable y amable en función del estado emocional actual del usuario. Siguiendo este enfoque, [3] construye un complejo sistema de agente conversacional incorporado, capaz de hacer un procesamiento de lenguaje natural de alta calidad, así como una sofisticada manipulación de las emociones basada en el Modelo de Plutchik [4]. Este chatbot analiza y sintetiza el estado emocional actual y la expresión emocional representada en los mensajes del usuario, de modo que se puede generar una respuesta de forma personalizada.

Con la suposición de que el estilo lingüístico puede ser un indicador de temperamento, se propuso un chatbot con una personalidad explícita en [5]. El objetivo de este chatbot es generar respuestas coherentes con una disposición o perfil preestablecido. La propuesta utiliza datos genéricos de conversación de los medios sociales para generar respuestas coherentes con el perfil, que representan un perfil de respuesta específico adecuado para un mensaje de usuario recibido.

Heller et al. [6] describen otro trabajo relacionado, donde un chatbot llamado "Freudbot" fue construido usando la arquitectura de código abierto del Lenguaje de Marcado de Inteligencia Artificial (AIML). El objetivo de este chatbot era mejorar la interacción

estudiante-contenido en la educación a distancia. Explícitamente, esta tecnología de chatbot es prometedora como herramienta de enseñanza y aprendizaje en la educación a distancia y en línea.

A su vez, Sabbagh et al. [7] presento la herramienta HI^{2}P que se centra en el fomento de una cultura de seguridad de la información y la sensibilización entre los usuarios. Esta herramienta incorpora diferentes tipos de métodos de aprendizaje y temas como la respuesta a incidentes y las políticas de seguridad. La interacción con el usuario es basado en el chatbot de ALICE que utiliza AIML, lo que hace que las soluciones sean simples y eficientes.

Otro caso de chatbot utilizado para la formación en seguridad se presenta en [8], donde el chatbot Sally es capaz de interactuar con algunos grupos de empleados de una empresa que tienen una educación o experiencia diferente en seguridad. Sally fue capaz de proporcionar entrenamiento en seguridad, lo que se evidenció en un crecimiento en el conocimiento de los usuarios objetivo.

Además, el trabajo presentado en [9] investiga sobre el comportamiento de las personas cuando son conscientes de que están interactuando con los chatbots. Los resultados muestran que en tal situación, la conversación puede llegar a ser simple y compuesta de mensajes cortos, incluso si puede ser extendida en el tiempo. Por el contrario, las conversaciones con un humano pueden convertirsen complejas y compuestas de mensajes extensos, pero más cortos en el tiempo. Además, la misma investigación encontró que las habilidades lingüísticas como el vocabulario y la expresión se transfieren fácilmente a una máquina.

La Máquina de Charla Emocional (ECM) [10] es una propuesta con un enfoque de aprendizaje automático que considera el estado emocional de la conversación para generar respuestas apropiadas en el contenido (relevante y gramatical) y en la emoción (emocionalmente consistente).

Particularmente relacionado con el tema del acoso sexual y la pornografía infantil, Zambrano et al. [11] presentan BotHook, un chatbot para identificar a los ciberpederastas y atrapar a los ciberdelincuentes. En este trabajo, se desarrolló un módulo de atracción de intereses y caracterización de pedófilos. Asimismo, el trabajo presentado en [12] analiza la eficacia de los métodos actuales de detección de ciberpervertidos y propone algunos métodos futuristas como el análisis de metadatos y datos de contenido de las comunicaciones de VoIP, así como la aplicación de chatbots totalmente automatizados para operaciones encubiertas.

Se propone un sistema para detectar el acicalamiento en [13] utilizando un enfoque BoW para el preprocesamiento y la selección de rasgos difusos (FRFS) para la selección de los rasgos más importantes. Para la clasificación se utilizó la máquina de vector de soporte de gemelos difusos. Se utilizaron dos conjuntos de datos para el entrenamiento, uno proveniente del sitio web de justicia pervertida y el otro de PAN13.

Se propone un clasificador para detectar las conversaciones de Grooming en [14] usando una máquina de vectores de apoyo y un vecino K-nearest (KNN). También se propone un grupo de 17 características asociadas a una etapa de acicalamiento espacial, que se utiliza para la selección de características y el entrenamiento. La comparación de los modelos se realiza mediante el Indicador de Precisión.

[15] propone un clasificador que utiliza un enfoque de aprendizaje profundo (Red

neuronal convolucional) se utiliza en comparación con un modelo de detección de anomalías, es decir, un modelo de clasificación SVM de una clase, logrando una mejor puntuación F1. Se utilizaron dos conjuntos de datos: PAN-2012 y SQ (Surete du Quebec). Disparidad de edades.

En [16] se propone un modelo de detección de ciber bulling que utiliza la información del usuario (víctima o acosador) de diferentes fuentes, como las redes sociales: youtube, facebook y twitter. Considera la información proveniente de diferentes grupos de género y edad. Se desarrolló un conjunto de datos propio que contiene mensajes de acoso.

Este trabajo [17] propone la detección del ciberbulling considerando aspectos como el perfil del usuario, sus características (edad, género), su comportamiento en las diferentes redes sociales después de una experiencia de acoso. Esta perspectiva de sistemas múltiples requeriría la definición de un algoritmo que pueda clasificar los comentarios

Se propone un agente de conversación que se haga pasar por niño en [18] que tiene como objetivo prevenir los cibercrímenes asociados a la pedofilia, la pornografía infantil y la explotación sexual. Este agente de conversación puede conectarse a chats y redes sociales, y su principal componente se basa en la teoría del juego. Para ello, se utilizan 7 chats.

Este sistema (GARS) propuesto en [19] calcula el riesgo de cibersalida considerando algunos valores de riesgo particulares provenientes de controladores de lógica difusa. GARS hace un análisis en tiempo real. La lógica difusa apoya el proceso de gestión de riesgos utilizando funciones de membresía de triángulo, variables (exclusividad de las conversaciones, cambio de los valores de riesgo). Las reglas difusas se generan de acuerdo con la experiencia de las conversaciones de preparación. Cuando se supera un umbral de riesgo, se dispara una alarma.

Un estudio de diferentes enfoques computacionales para predecir la tensión social en los medios sociales (Twitter) se define en [20]. La comparación de los enfoques se basa en la precisión, la memoria, la medida F y la exactitud. Los enfoques considerados son el motor de análisis de tensión que es una propuesta propia basada en el método de análisis de conversación MCA (Membership, categorization, analysis) [21], el enfoque de aprendizaje automático con el clasificador Bayes ingenuo y el análisis de sentimiento con la herramienta SentiStrength [22]. El análisis conversacional y las reglas de minería de texto sintácticas y basadas en léxico mostraron un mejor rendimiento que los enfoques de aprendizaje automático.

Se hace una propuesta para la detección de la manipulación de niños en línea en [23], donde se comparan una máquina de vector de apoyo, un vecindario más cercano y una propuesta propia basada en el número de características de manipulación. La construcción de los clasificadores se hace utilizando un conjunto de 17 características identificadas de la conversación de aseo que se unen a una etapa específica de la conversación de aseo. A partir de los experimentos, el método SVM con un núcleo lineal obtuvo el mejor rendimiento con una precisión del 98,6 %.

Esta propuesta en [24] trata de detectar si un adulto está fingiendo ser un niño como parte de un abuso de aseo en línea. La propuesta identifica a una persona como adulto de un niño basado en el estilo de escritura, luego determina si un niño es un niño falso o no. Este documento sugiere que es difícil diferenciar a los niños de los adultos en textos informales (blogs o registros de chat). La propuesta utiliza un conjunto de 735

características recogidas de la literatura, que se utilizan para construir modelos basados en algoritmos como Adaboost, SVM y Naive Bayes.

Como se ha descrito anteriormente, encontramos diferentes esfuerzos en la literatura w.r.t. el desarrollo de los chatbots utilizando principalmente modelos basados en reglas y en la generación, que también pueden ser mejorados con la adición de módulos emocionales. La utilización de los "chatbots" en la educación y la capacitación en materia de seguridad es un ámbito de aplicación cada vez mayor. Además, están surgiendo también varias obras que aplican los "chatbots" para hacer frente a los delitos sexuales relacionados con el abuso de menores, utilizando el "chatbot" como emulador de la víctima, es decir, del niño abusado. En el presente proyecto, proponemos una mejor version C^3-Sex para hacer frente al abuso infantil con un enfoque diferente que no se ha considerado anteriormente. En esencia, nuestro chatbot emula a un individuo interesado en el tema de la pornografía infantil. Además, nuestra propuesta tiene un importante componente que se centra en el perfilamiento del sospechoso utilizando 6 diferentes métricas que en conjunto contribuyen con información importante para LEA en la caza de pervertidos.

Capítulo 5

Aplicación del chatbot C3-sex en un escenario de pornografía infantil

Este proyecto se basa en C3-Sex [25], dicho proyecto tiene como objetivo diseñar y construir un Bot con inteligencia sentimental interactiva, capaz de descubrir y revelar el perfil de los ciberdelincuentes en chats masivos en línea, mediante la elaboración de respuestas pertinentes teniendo en cuenta el contenido lingüístico y sentimental. A continuación se describe el funcionamiento del bot y las mejoras que se realizaron.

5.1. Modelo Generativo mejorado

La arquitectura usada en este modelo se basa en la estrategia LSTM el cual genera una respuesta a partir de un entrenamiento previo. Esta estrategia es la base del modelo Seq-to-Seq el cual usa redes neuronales recurrentes que permiten analizar los patrones en la oración y mantener una memoria de las palabras. Cuando se ingresa una oración, el modelo procesa y almacena la oración, y responde de acuerdo a la memoria almacenada.

Con el fin de mejorar las respuestas obtenidas sin modificar la aproximacion inicial se opto por utilizar otro tipo de datos para realizar el entrenamiento. La base de conocimiento utilizado anteriormente era PAPAYA [26] el cual contiene datos de diferentes fuentes, como noticias, conversaciones e interacciones en foros. Esto le permitía al bot responder a conversaciones que tuvieran diferentes temáticas. No obstante las interacciones en linea no suelen ser tan elaboradas por eso se decidió utilizar otra base de conocimiento que permitiría mayor fluidez en las respuestas.

Posteriormente el modelo generativo fue entrenado con el conjunto de datos de Cornell Movie Dialogs. Este conjunto de datos contiene conversaciones entre personajes de más de 600 películas. En este sentido, el modelo se configura con un dinamismo y una naturalidad realmente interesantes para que el bot responda en situaciones casuales o joviales, para las que no se espera una respuesta muy elaborada, lo cual es una caracteristica tipica del escenario hacia el cual va dirigido este chatbot. Las ventajas obtenidas usando este nuevo dataset son descritas a continuación:

- **Mayor cantidad de datos:** Al usar este nuevo modelo con 220.579 conversaciones podremos obtener más y mejores respuestas. No obstante, el tiempo de entrenamiento fue superior para alcanzar un nivel de exactitud similar.
- **Más respuestas coherentes:** Como el conjunto de datos se enfoca en conversaciones naturales entre personas este tiene respuestas más coherentes

- **Respuestas más casuales:** Se logro el objetivo que se buscaba inicialmente, es decir que ahora el modelo tiene respuestas mucho mas joviales respecto a temas sin profundidad o sin contexto bien definido

El codigo utilizado para el entrenamiento del modelo generativo se encuentra aquí

5.2. Modelo *Retrieval* mejorado

Este modelo esta montando en un sistema basado en reglas el cual es un sistema que aplica reglas creadas, para almacenar, clasificar y manipular datos. Los sistemas basados en reglas se han convertido en una de las herramientas más eficientes, pero para funcionar bien, estos sistemas requieren que la fuente de datos cumpla con un cierto patron para que las reglas puedan lanzarse. Estas reglas tambien son denominanads "declaraciones de Si", ya que tienden a seguir la línea de **Si** Pasa X,**entonces** haz Y, así mismo estos sistemas basados en reglas contiene una base de conocimiento el cual es un tipo especial de base de datos para la gestión del conocimiento. A su vez existe un componente llamado el motor de inferencia,el motor de inferencia es el capaz de extraer conclusiones aplicando métodos de la lógica clásica sobre esta base. Una regla en este contexto es una proposición lógica que relaciona dos o más objetos del dominio e incluye dos partes, la premisa y la conclusión.

Esta forma que trabaja un motor de inferencia es conocido como el sistema basico de Modus Ponens, el cual es un encadenamiento de reglas hacia adelante.

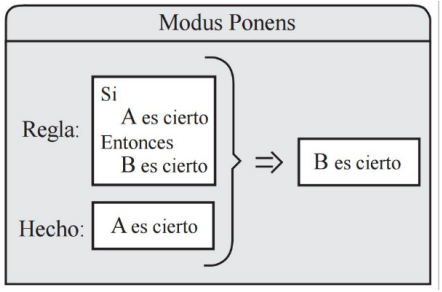


FIGURA 5.1: Representación del sistema logico Modus Ponens

El modelo retrival propuesto en el presente proyecto fue implementado con el lenguaje de modelado de inteligencia artificial o más conocido como AIML , AIML tiene como objetivo construir un comportamiento humando el cual manteniene una implementación simple de programar, fácil de entender , altamente mantenible y es un lenguaje de marcado basado en XML destinado a crear aplicaciones de inteligencia artificial.

Ejemplo de regla AIML

```
<category>
        <pattern>* FEEL *</pattern>
        <template>cool, u know whats more exciting? pics haha</template>
/category>
```

Este modelo es el capaz de interpretar las respuestas del sospechoso y darles una respuesta más acorde al mensaje, así mismo se trata de establecer el rumbo de la conversación objetivo la cual busca el Bot, pero en caso de que la interacción del sospechoso no se le puede dar una respuesta con ninguna regla existentes en la base de conocimiento, entonces el modelo generativo entra en operación y genera una respuesta.

Este modelo se mejoró en dos aspectos, refactorización de reglas existentes y nuevas reglas. La primera versión de este proyecto contenía un total de 113 reglas las cuales 13 se refactorizaron con el fin de que el sospechoso pasara a la segunda fase del proyecto y se agregaron 27 nuevas reglas asociadas a hard content , BSDM (Bondage, Disciplina, Dominación, Sumisión, Sadismo y Masoquismo) y pornografía infantil, logrando un total de 153 reglas para el modelo retrival.

1. **Refactorización** : En la refactorización de las reglas, lo que se cambió fueron las reglas previas que ya se tenían para el cambio de plataforma a telegram, ahora ya que se tiene la parte de snapchat, se le invita al usuario a agregar a al bot a su cuenta de snapchat y así poder continuar con la conversación.

   Ejemplo de regla AIML

   ```
   <category>
         <pattern>* KIK * SNAP *</pattern>
         <template>Trade: i prefer to have the conversation on snap, add me
   p_ramirezxxx</template>

   </category>
   ```

2. **BDSM** : Las reglas nuevas asociadas a BDSM son reglas que incitan al sospechoso a compartir el contenido sobre sumisión y masoquismo en niños. así tratando de lograr una transferencia de contenido asociada a ello en snapchat.

   Ejemplo de regla AIML

   ```
   <category>
         <pattern>* share * submissive * children *</pattern>
         <template>i'd like but prefer in private add me snap , my nickname
   is p_ramirezxxx<template>
   /category>
   ```

3. **Pornografia infantil** : Las nuevas reglas asociadas a la pornografía infantil se crearon con el enfoque de la transferencia del contenido asociado a la pornografía infantil, tales como links, videos o imágenes.

Ejemplo de regla AIML

```
<category>
        <pattern>* link to * child *</pattern>
        <template>i'd like but prefer in private add me snap , my nickname
is p_ramirezxxx<template>
/category>
```

Las reglas descritas anteriomente junto con las nuevas que se agregaron se encuentran aqui

5.3. Modelo de detección de emociones

La evaluación de las emociones detectadas en la conversación son extraídas utilizando una maquina de soporte vectorial (SVM). La SVM esta entrenada con un conjunto de datos de Evaluación semántica como paradigma de clasificación y aprendizaje supervisado. Las respuestas del sospechoso son clasificadas dentro de una de las seis emociones (enojo, disgusto, miedo, alegría, tristeza y sorpresa). En las pruebas realizadas anteriormente se obtuvo aproximadamente un 60 % de accuracy. Este modelo fue re utilizado del proyecto C3-sex [27].

El modelo de detección de emociones se puede encontrar aqui

5.4. Modelo de análisis de sentimientos

Con el fin de distinguir si un frase dada genera una opinión inclinada hacia (o en contra) un tema específico, el proyecto usa un modelo de clasificación de opinión que aprovecha una red bayesiana multinomial con un preprocesamiento simple (vectorización, eliminación de palabras clave, etc.). El conjunto de datos utilizado consta de 2000 muestras de reseñas positivas y negativas de películas, restaurantes y otros productos en base a la propuesta de Kotzias, D., Denil, M., De Freitas, N., y Smyth, P (2015). La representación de las características se basó en una matriz de términos de documentos teniendo en cuenta la frecuencia de las palabras. Después de entrenar con el 90 % de las muestras, probamos nuestro modelo con el 10 % restante, logrando un 80 % de precisión. Este modelo fue re utilizado del proyecto C3-sex. [27].

El modelo de análisis de sentimientos se puede encontrar aqui

5.5. Entendimiento de la manipulación de la aplicación Android

En la primera versión de C^3-Sex , Se propuso como trabajo futuro la manipulación de un dispositivo android con el fin de tener un nuevo escenario de cibercrimen y poder mejorar el perfilamiento del Bot; Este nuevo escenario de cibercrimen nos brindo una nueva métrica importante, la trasferencia de contenido alusivo a la pornografía infantil, esto se logra con la implementación SnapBot[1], SnapBot es un software elaborado con

[1]https://github.com/CrkJohn/Snapchat

el fin de manipular la aplicación móvil de snapchat para interactuar con los posibles sospechosos en una segunda fase.

5.5.1. Manipulación de un dispositivo Android utilizando código

En esta versión de C^3-Sex se agrega el componente de manipulación de la aplicación Andorid, este se implemento con el framework APPIUM[2], APPIUM es un framework open source para la implementacion de test automatizados de iOS, Android, hasta Windows apps el cual usa protocolo WebDriver.

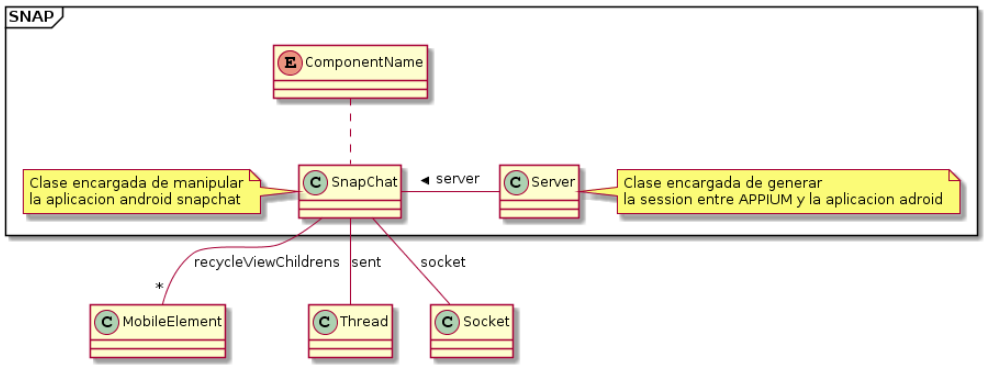


FIGURA 5.2: Representación de comunicacion de SnapBot con C^3-Sex

En la figura 5.2. se refleja la estructura del código SnapBot, en el paquete SNAP, podemos evidenciar dos clases y un enumerable, el enumerable componentName es enumerable el cual nos permite saber los nombres de cada componente que conforma la aplicación móvil Snapchat, es decir, TEXT_VIEW , LINEAR_LAYOUT, VIEW , etc. En las dos clases restantes están la clase $SnapChat$ y $Server$. $Server$ es la clase encargada de crear la conexión entre el servidor Appium y la aplicación de android, que en este caso es Snapchat. $Snapchat$ es la clase encargada de toda la manipulación de la aplicación de android desde su ejecución hasta su terminación, incluso establece una conexión por medio de socket con la aplicación de Python C^3-Sex , el cual tiene entrenado el Bot.

Se tuvieron algunos desafios para que esta fase del proyecto para que se ejecutará totalmente bien, los impedimentos que se tuvieron fueron los siguientes: descarga de archivos multimedia, selenium y automation (librería de la conexión con Android) y problemas con servidor appium. A continuación se describirán los desafios encontrados y la solución implementada.

Descarga de los componentes de una aplicación

Para analizar la información de la aplicación(ID, Clases, Text, Index) se utilizó la herramienta uiautomatorviewer, la cual proporciona una interfaz gráfica los cual permite escanear y analizar la aplicación Android. Se puede utilizar esta herramienta para

[2]http://appium.io/docs/en/about-appium/intro/?lang=es

inspeccionar la jerarquía de diseño y ver las propiedades de los componentes de la interfaz de usuario que son visibles al usuario. Esta inspección sobre los componentes nos permite crear selectores sobre la interfaz. [28]

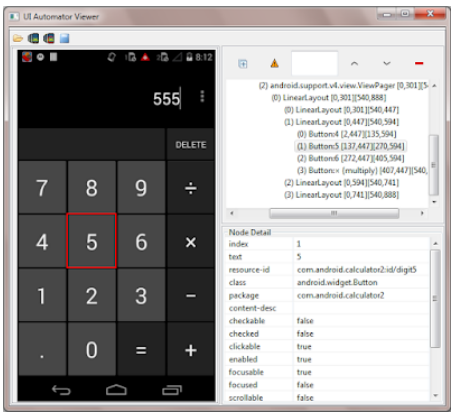


FIGURA 5.3: Herramienta Uiautomatorviewer de Android

En la figura 5.3 se muestra un ejemplo de cómo se ve la aplicación y cómo se visualizan los componentes. El problema de la descarga de los archivos multimedia era que no existe ningún atributo en los componentes para saber sí es un archivo multimedia, por lo cual era difícil saber si sí o no era un archivo, primero se intento estandarizar la jerarquía que podría tener un archivo multimedia y así lograr la descarga de los archivos, pero desafortunadamente cuando se paso a la fase de pruebas esta jerarquía planteada no fue exitosa ya que esta jerarquía dependía de cómo chateaba el sospechoso, es decir, la jerarquía estandarizada que habíamos planteado era que el sospechoso solo mandaba una foto y nada más, por consiguiente tuvimos que buscar otra alternativa de solución y la encontrada fue tomar capturas de pantallas cada vez que hubiera una interacción por el Bot, esta solución no es optima pero nos permite identificar cuando ocurre una tranferencia de contenido multimedia en Snapchat.

Manipulación de los componentes de una aplicación

Selenium y Automation son librerías las cuales nos permiten manipular la aplicación, pero estas librerías dependen de la estructura que contiene la aplicación e incluso de la versión de android que tiene el dispositivo,por esta razón este es un problema a tener en cuenta ya que al realizar las pruebas de manera repetitiva el GUI de la aplicación móvil (Snapchat) cambiaba simultáneamente y así mismo cambiaba la estructura de sus componentes, por lo cual hacia que Snapbot no funcionara bien y no fuera capaz de encontrar los componentes que se tenían establecidos. Nunca encontramos la solución de este problema y lastimosamente perdímos esas conversaciones en nuestras pruebas y analisis.

El servidor Appium

Appium es un servidor que se activa mediante línea de comando, inicialmente se hicieron las pruebas iniciando solo una vez este servidor. Debido a que comenzamos

a tener dificultades con la aplicación Java - Snapbot, encontramos que el servidor selenium se detenia después de iniciar la aplicación más de dos o tres veces, por lo cual tuvimos que agregar una funcionalidad extra al proyecto y era que el proyecto se encargara de inicializar y acabar con el proceso del mismo.

5.5.2. Comunicación entre la aplicación Android y Python.

Debido a que el proyecto C^3-Sex trabaja en dos plataformas(Omegle y Snapchat), se vio la necesidad de realizar dos proyectos por separado, el primero donde esta el Bot entrenado y el control de Omegle, y el segundo proyecto se controla por completo la apliación movil.

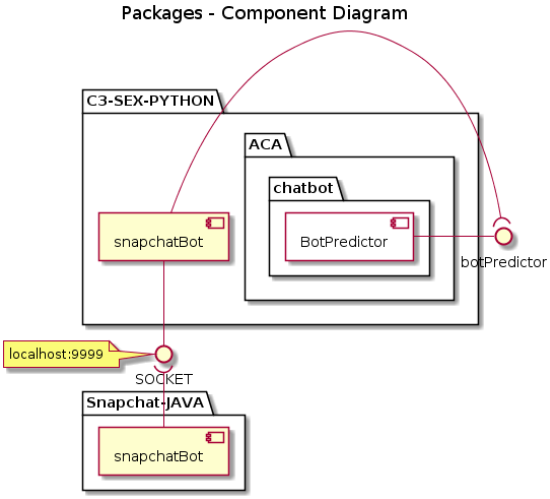


FIGURA 5.4: Diagrama de componentes de la segunda version de C^3-Sex

En la figura 5.4 se refleja el diagrama de componentes de C^3-Sex , el cual refleja la comunicación entre las dos aplicaciones, las aplicaciones se conectan por un socket debido a que debemos establecer una sesión con tensorflow entre las dos aplicaciones, al tener ésta sesión funcionando por medio de mensajes usando el socket, damos respuestas a los mensajes capturados por el segundo proyecto, SnapBot. Así generando la interacción entre el sospechoso y la aplicación móvil.

Problema con el Socket

Inicialmente el proyecto comienza desde el componente de Python(main.py) por consiguiente este es el encargado de iniciar el servidor appium , el proyecto Java y abrir el socket entre Java Y Python. Cuando sucedía el caso que la snapbot no iniciaba el socket entre Java Y Python por lado de Python quedaba abierto y así mismo ese puerto en el computador quedaba abierto por lo que no se podía volver a instanciar el socket ya que el puerto ya estaba en uso, por esto se opto por establecerle un timeout al socket por si la aplicación de Java fallaba al iniciar este puerto no quedara abierto.

Capítulo 6

Evaluación del chatbot C3-sex en un escenario de pornografía infantil

6.1. Atributos generados

El identificador(ID) que le dimos a cada experimento que se realizó en este proyecto fue con la fecha y la hora local actual, lo decidimos así ya que el proyecto es secuencial y por esta razón este valor nunca se va a repetir.

Para poder responder preguntas tales como : cual es el promedio de tiempo de las conversaciones, cuales son las horas donde las conversaciones en Omegle o Snapchat duran más, incluso saber en que horarios es mejor poner en ejecución el Bot. establecimos las siguientes seis métricas.

1. startCoversationSP: Tiempo de inicio de la conversación en Snapchat con el formato con YYYY-MM-DD HH:mm:ss.ff ⇒ Esta métrica es creada cuando se encuentra un nuevo chat después de finalizar la conversación en Omegle.

2. endCoversationSP: Tiempo de fin de la conversación en Snaphat con el formato con YYYY-MM-DD HH:mm:ss.ff ⇒ Esta métrica no se crea hasta que no pase que el sospechoso se retire del chat, se despida o termine el tiempo de conversación estipulado de tres minutos.

3. startCoversationOMG: Tiempo de inicio de la conversación en Omegle con el formato con YYYY-MM-DD HH:mm:ss.ff ⇒ Tan pronto como la interacción del Bot. y el sospechoso comience, esta métrica será creada.

4. endCoversationOMG: Tiempo de fin de la conversación en Omegle con el formato con YYYY-MM-DD HH:mm:ss.ff ⇒ Esta métrica se crea una vez que las reglas AIML alusivas a cambiarse de plataforma hacen match y la aplicación le escribe al sospecho el nickname del snapchat en omegle.

5. Total duration Omegle [Seconds]: Duración total en la pagina web Omegle en segundos ⇒ Esta métrica es generada con la resta de las metricas 3 y 4

6. Total duration Snapchat [Seconds]: Duración total en la aplicacion movil Snapchat ⇒ Esta métrica es generada con la resta de las metricas 1 y 2

7. Time Metric [seconds]: Representa el tiempo entre la generación de una respuesta por parte del bot y la siguiente interacción del sospechoso

Una de las métricas que deben resaltar más en este proyecto es la métrica de transferencia de contenido ya que con esta métrica podemos realmente afirmar que el sospechoso es un pervertido o no.

1. TransferContent: Indica si se ha compartido contenido ilegal (1) o no (0) ⇒ Esta métrica desafortunadamente no se pudo realizar automatizada lo cual implica que se tiene que hacer una inspección manual sobre el contenido tranferido en Snapacht.

Para determinar la eficiencia de los modelos y el uso que se esta dando a cada uno de ellos, se establecieron las siguientes métricas, con estas métricas nosotros podemos determinar el uso que se tiene de cada modelo a lo largo de las conversaciones con el sospechoso.

1. numberRulesMatched [N+]: Número de reglas matcheadas por Bot ⇒ Este contador aumenta uno si una regla AIML concordo con lo escrito por el sospechoso
2. numberRulesNotMatched [N+]: Número de respuesta del modelo generativo ⇒ Una vez que el modelo de reglas no encuentra una regla acorde con lo escrito por el sospechoso, se encarga el modelo generativo de responder. Cuando esto sucede, esta métrica aumenta en uno.
3. numberInteractions [N+]: Número de respuestas que tuvo el Bot ⇒ Es suma de los items 1 y 2.

1. size_bytes_bot_file [N+]: Número total de bytes de texto enviados por el chatbot.
2. Average suspect posts length: Longitud media (en número de caracteres) de los posts de un sospechoso.
3. Length conversation [N+]: Es el número de mensajes que hubo por parte del bot y el sospechoso, con esta métrica podemos analizar si las respuestas del Bot fueron coherentes.

Para determinar la frecuencia con la que aparecen ciertas emociones en los mensajes del sospechoso se definieron las siguientes metricas, con estas podemos calcular la tendencia hacia emociones positivas o negativas que tuvo la conversación

1. anger_EC: # veces que el sentimiento “anger” fue identificado por el modelo de emociones en los mensajes del sospechoso - (Nueva Feature).
2. disgust_EC: # veces que el sentimiento “disgust” fue identificado por el modelo de emociones en los mensajes del sospechoso - (Nueva Feature).
3. fear_EC: # veces que el sentimiento “fear” fue identificado por el modelo de emociones en los mensajes del sospechoso - (Nueva Feature).
4. sadness_EC: # veces que el sentimiento ’sadness’ fue identificado por el modelo de emociones en los mensajes del sospechoso - (Nueva Feature).
5. other_EC: Otro sentimiento identificado por el modelo de emociones en los mensajes del sospechoso - (Nueva Feature).

6. Emotion Metric [0 - 1]: Representá una ponderación de las emociones encontradas en la respuesta del sospechoso, donde 0 es una emoción negativa y 1 una emoción positiva

Para el caso de las opiniones se hace un proceso muy similar que con las emociones, primero hay unas metricas que cuantifican la frecuencia de las opiniones de la conversación y con estas, podemos calcular la tendencia hacia una opinion positiva o negativa

1. Neg_SA: # veces que fue identificado una emoción negativa en la conversación por el modelo de opinión en los mensajes del sospechoso (Nueva Feature).
2. Pos_SA: # veces que fue identificado una emoción positiva en la conversación por el modelo de opinión en los mensajes del sospechoso (Nueva Feature).
3. Neut_SA: # veces que fue identificado una emoción neutra en la conversación por el modelo de opinión en los mensajes del sospechoso (Nueva Feature).
4. Opinion Metric [0 - 1] : Representa una ponderación de las opiniones encontradas en la respuesta del sospechoso, donde 0 es una opinion negativa y 1 una opinion positiva

Finalmente se introdujeron metricas que permitan conocer el numero de enlaces a paginas externas compartidos en la conversación, tanto en Snpachat como en Omegle

1. url_snap [N+]: Número de enlaces enviados por el sospechoso en Snapchat.
2. url_omg [N+]: Número de enlaces enviados por el sospechoso en Omegle
3. url_total [N+]: Numero total de enlaces enviados por el sospechoso.

Capítulo 7

Estrategias de conversación y de clasificación avanzadas

7.1. Aplicación de Teoría de Juegos en un agente conversacional

7.1.1. Qué es teoría de juegos

La teoría de juegos es definida como un estudio formal de la toma de decisiones en el que varios jugadores deben tomar decisiones que pueden afectar a los intereses de otros jugadores, es decir, la forma en que los individuos actúan tiene que ser estratégica, es decir, deben ser conscientes del hecho de que sus acciones afectan a los demás. El hecho de que las acciones tengan un efecto en el resultado no requiere un comportamiento estratégico, si no se es consciente de ese hecho. Por lo tanto, decimos que la teoría del juego estudia la interacción estratégica dentro de un grupo de individuos, siendo la interacción estratégica la capacidad de que los individuos entiendan que sus acciones tendrán un efecto en el resultado y actuarán en consecuencia.

7.1.2. Descrición de una aplicación existente para la prevención y la detección del Grooming

Objetivo

El objetivo principal del sistema de Negobot [29] es reunir, a través de la conversación, la máxima cantidad de información para evaluarla después. Esta evaluación determinará si el sujeto es un pedófilo o no, para comunicar, en este último caso, a las autoridades pertinentes.

Procesamiento inicial

Las entradas de datos son llamadas unidad conversacional (CU). cada vez que entra una nueva CU se remueven las palabras repetidas y los emoticones son remplazados, posteriormente se traduce el CU a ingles utilizando el servicio de google, a todo este proceso lo llamaremos normalización. Luego de que el CU este normalizado se busca este dentro de las conversaciones ya procesadas utilizando Lucence (herramienta de búsquedas optimizadas) con el fin de encontrar uno similar para ubicar este CU dentro de un rango. El resultado final es entonces utilizado para calcular el valor B de la función de evaluación (equation 1)

Aplicación de la teoria de juegos

El sistema Negobot asume que hay dos actores en una conversación: el supuesto pedófilo y el propio sistema. Con el objetivo de aplicar la teoría de juegos se definen una serie de niveles acorde a las acciones que se deben tomar respecto a el mensaje enviado por el sospechoso. Cada vez que se recibe un mensaje el sospechosos es evaluado, el resultado de la evaluación determinará si el sujeto que habla es un presunto pedófilo o no, dependiendo del contenido de la conversación y asi decide cambiar o no el nivel de la misma, a continuación se describe cada nivel.

- **Estado inicial (Nivel 0):** En este nivel la conversación ha iniciado recientemente, la conversación puede permanecer este nivel de forma indefinida. El bot suministra poca información personal como: nombre, edad, genero, ciudad de origen.
- **Posiblemente No (Nivel -1):** En este nivel el sospechosos no quiere continuar la conversación, el bot intentará reactivar la conversación, para ello preguntará acerca de problemas familiares, bullying u otro tipo de problemas típicos de adolescentes
- **Probablemente No (Nivel -2):** En este nivel el sospechoso está cansado de la conversación, su lenguaje es mucho menos educado, la conversación está a punto de perderse. El bot actuará como una víctima que se siente ignorada con tal de recapturar la atención.
- **No es un pedofilo (Nivel -3):** En este nivel el sospechoso ha dejado de hablar con el bot. La estrategia para recuperar la atención en este caso es buscar afecto a cambio de sexo. Esta estrategia se utiliza ya que los pedófilos suelen esconderse para no ser capturados o identificados
- **Posiblemente si (Nivel +1):** En este nivel el sospechosos muestra interés y preguntá por más datos personales. Los temas del bot son peliculas y musica favorita, ropa, consumo de drogas y alcohol y problemas familiares.
- **Probablemente si (Nivel +2):** En este nivel los temas se vuelven más privados como situaciones sexuales y experiencias personales. La información es más detallada y privada. Después de alcanzar este nivel no puede disminuir
- **Pedofilo (Nivel +3):** En este nivel el bot determina que el sospechosos es un pedofilo. La conversación sexual se vuelve muy explicita. Con el objetivo de mantener interesado al sospechosos el bot suministra mucha información personal. Después de alcanzar este nivel no puede disminuir

El sistema NegotBot tiene tres sensores, tres actores y tres acciones a realizar con el fin de alcanzar el objetivo

Sensores

- Conocimiento del nivel actual
- Conocimiento de la conversación actual
- Conversación asimilada por el sistema

Actores

- El nivel de la acusación sube
- El nivel de la acusación baja
- El nivel de la acusación siguel igual

El nivel de acusación se refiere al perfil que se le da al sospechosos en cada nivel de conversación **Acciones**

- **El nivel de la conversación sube:** Esta acción incrementa el nivel de acusación e incrementa el nivel de conversación agregando más información personal a la conversación
- **El nivel de la conversación baja:** Esta acción reduce el nivel de acusación y reduce el nivel de conversación agregando más información tentativa a la conversación
- **El nivel de la conversación siguel igual:** Esta acción mantiene el nivel de acusación manteniendo el nivel de conversación

Estrategia

El objetivo del sistema es alcanzar un estado final mientras se extrae la la mayor cantidad de información. El estado del sistema está determinado tanto por la conversación y el tema de la conversación. El ambiente es estocástico porque las conversaciones se componen sólo de preguntas, respuestas, afirmaciones y negaciones y cuando el usuario es escribiendo una frase, el ambiente no cambia. Además, el sistema utiliza una función de evaluación para analizar las respuestas. Este análisis comienza evaluando el nivel de conversación. Luego, genera la respuesta y, finalmente, se comunica con el sujeto. Este tarea debía realizarse en un tiempo coherente y variable, de modo que se asemejara a la forma de escribir de un niño. Para ello, el sistema calcula un tiempo de espera para la respuesta basada en el número de palabras de la respuesta y una estimación la velocidad de escritura del niño.

Función de evaluación

La función de evaluación determina el comportamiento del bot según la evaluación de las frases que el sospechosos envía, y determina, en tiempo real, su nivel de pedofilia. Esta función también es responsable de evaluar la evolución de la conversación (es decir, si el nivel se mantiene, sube o baja)

$$f(x) = \alpha + \beta + \gamma \tag{1}$$

α Representa los valores históricos de las conversaciones con este sospechoso. Este se define como la suma de los niveles de las conversaciones anteriores, dividida por el número total de mensajes intercambiados con el sospechosos.

$$\sum_{n=1}^{n=x-1} f(x)/n$$

β Representa que tan *slimyness* es el CU actual. Se calcula el promedio del resultado de la multiplicación la puntuación de similitud de cada una de las conversaciones recuperados de los pedófilos

$$\sum_{n=1}^{n=x} puntaje/CA$$

Puntaje se asigna utilizando Lucence.
CA es el número de conversaciones recuperados de los pedófilos

γ Evalua con que frecuencia el sospechosos interactura con el bot. Hay dos posibles valores para γ segun el valor de β

- si $\beta = 0$, γ se define como la resta entre la *slimyness* del CU actual y el total de las CU del sospechosos, este valor se divide por el delta de tiempo entre el primer y el ultimo mensaje
- si $\beta > 0$, γ se define como el numero de CU que superan el umbral de *slimyness* dividido por el delta de tiempo entre el primer y el ultimo mensaje

7.1.3. Uso de la teoria de juegos en C^3-Sex

Con el objetivo de mantener una conversación prolongada que permita llegar a una transferencia de contenido ilegal por parte del sospechoso se definieron cuatro estrategias que permitan utilizar ciertos conjuntos de reglas de acuerdo al comportamiento del sospechoso para responder de forma más acertada y dirigir la conversación a un ambiente de intercambio de contenido. La estartegia principal en la que se va a enfocar es aumentar el nivel de estrategia:

- **Estrategia inicial:** En este nivel el bot presenta información personal muy escasa aunque relevante, como edad, sexo, lugar donde recide.El foco central de esta estrategia es hacer que el sospechoso se interese en el bot y quiera mantener la conversación para que evite saltarla antes de pasar a un proximo nivel.
- **Estrategia Interesado:** En este nivel se comienza a mostrar interes por el contenido ilegal para ver si el sospechoso desea continuar con la conversación y lograr llegar a tener un intercambio.
- **Estrategia Intercambio:** El foco de este nivel es motivar al sospcehoso a continuar la conversación por otra red social(en este caso snapchat) que agregue al bot y que el sospechosos sea el encargado de iniciar la conversación. En este nivel el usuario de la red social sera repetido numerosas veces hasta llegar a desconectarse obligando al sospechoso a continuar en snapchat.
- **Estrategia Snapchat:** En este nivel lo primordial va a ser buscar el intercambio de contenido ilegal. Al llegar a esta estrategia es imposible regresar a las anteriroes.

Para decidir que estrategía es la mas adecuada para aplicar se optó por dos aproximaciones. Para ambas aproximaciones se utilizaron los datos descritos en la sección

anterior, sin embargo solo se utilizaron los que se podian calcular en tiempo de ejecución, como el número total de mensajes, el tiempo total de conversación, el número de reglas disparadas, entre otras.

La primera aproximación que se definió fue delimitar cada estrategia por un rango basado en datos. Para ellos utilizamos el número de reglas disparadas y la longitud total de la conversación de la siguiente manera:

- Si el número de reglas es **menor a 5** y el número de mensajes enviados es **menor a 10** se utiliza la **estrategía inicial**
- Si el número de reglas está **entre 5 y 10** y el número de mensajes enviados es **menor a 20** se utiliza la **estrategía Interesado**
- Si el número de reglas está **entre 11,15** y el número de mensajes enviados es **menor a 30** se utiliza la **estrategía Intercambio**
- Si los resultados no están dentro de las condiciones anteriores se usa la **estrategía Snapchat**

La segunda aproximación utiliza una fórmula para definir la estrategía a utilizar.

Función de decisión

Para llegar a esta función utilizamos algunas de las métricas descritas anteriormente, dichas métricas fueron seleccionadas porque eran calculadas en tiempo de ejecución, lo que permite recalcular la función constantemente. El objetivo de la función es que a mayor cantidad de mensajes y reglas disparadas mayor fuera la estrategia a utilizar (directamente proporcional) y a menor demora entre mensajes y menor cantidad de mensajes respondidos por el modelo generativo mayor fuera la estrategia a utilizar (inversamente proporcional). Teniendo todo esto en cuenta se obtuvo la siguiente función:

$$((Tt/Tu) * Rm)/(Tsm * Mg)$$

- Tt: Tiempo total o delta entre primer y ultima interacción
- Rm: # de reglas disparadas hasta ese momento
- Tu: Tiempo de espera entre la respuesta y el siguiente mensaje
- Mg: # de respuestas del modelo generativo
- Tsm: # número de mensajes totales enviados por el sospechoso

Resultados

El objetivo principal de utilizar esta nueva caracteristica dentro del modelo basado en reglas es que la conversación parezca mucho más natural y que el bot parezca una persona realmente interesada en el contenido ilegal y no solo una maquina programada. Es decir que la conversación va a ir escalando progresivamente y no se va a compartir información sensible hasta que la conversación alcance cierto punto, esto tambien reduce las interacciones con otros bots que lleguen a snapchat.

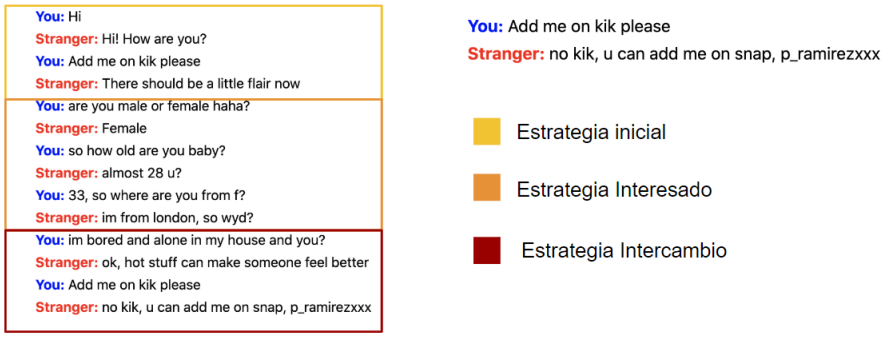


FIGURA 7.1: Respuesta del bot actualmente vs respuesta del bot sin la mejora

Como se puede apreciar cuando utilizamos esta nueva característica (figura de la izquiqerda) dentro del modelo basado en reglas este responde a la misma pregunta de forma distinta dependiendo del nivel al que haya basado la conversación, dando información sensible, en este caso el usuario de Snapchat, cuando se note un interés real en la conversación.

7.2. Aplicación de Lógica difusa en un agente conversacional

7.2.1. Qué es lógica Difusa.

La logica difusa es una extension de la logica tradicional que utiliza conceptos de pertenencia de conjuntos mas parecidos a la manera de pensar humana, asi mismo la lógica difusa no usa valores exactos como 1 o 0 pero usa valores entre 1 y 0 (inclusive) que pueden indican valores intermedios (Ej. 0, 0.1, ...,1.1, ...etc) [30], brevemente en pocas palabras la lógica difusa es una lógica multivaluada que nos permite representar matemáticamente la incertidumbre y la ambigüedad, proporcionando herramientas formales para su tratamiento [31].

Ejemplo:

- se considera a una persona como alta si mide mas de 1.80mts, pero de igual forma se considera a una persona como alta si mide 1.7999mts ← En el anterior ejemplo se planetea una medida donde una persona es alta, pero una apoximacion al valor exacto puede ser también alto o bajo dependediento del interprete.

Qué es un varible difusa

Una variable difusa tiene un valor nítido que adquiere algún número sobre un dominio predefinido (en términos de lógica difusa, llamado universo).

Qué es un conjunto difuso

Un conjunto difuso es aquel donde una variable difusa puede tomar un valor z siendo catalogado dentro un rango en un universo.

Qué es una función de membresia

Sea f una función difusa la que determina que tan bueno es un valor x en un conjunto difuso.

7.2.2. Descripción de una aplicación existente para la prevención y la detección del Grooming

Los problemas sobre la pornografía infantil cada vez son más grandes y debido a la rápida expansión de la Internet en el mundo. Esto ha implicado también un aumento a gran medida los casos de abuso infantil o rondar archivos multimedia asociado a la intimidad de menores en muchas redes sociales con una amplia difusión , es por esto que hay muchos investigadores que se han interesado sobre este tema y tratan de encontrar posibles soluciones que pueden ayudar a solucionar esta problemática. Algunos investigadores interesados en estos temas han sido Philip Anderson, Zheming Zuo, Longzhi Yang, investigadores de "Faculty of Engineering and Environment Northumbria University,United Kingdom "[13] Ellos han realizado una investigación sobre esta y lo exponen en el articulo : An Intelligent Online Grooming Detection System Using AI Technologies , en el cual usa lógica difusa, a continuación se explica las fases del funcionamiento del Framework.

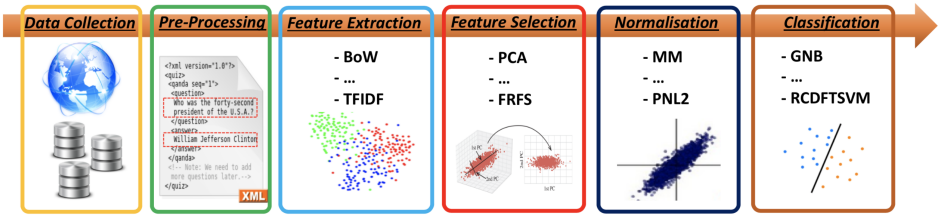


FIGURA 7.2: Representación del Framework que combate el Grooming, Tomado de [32].

1. Data Harvesting : La recolección de datos sobre grooming y non-grooming, en el cual pudieron recolectar más de 600 conversaciones de menores en las que participan los autores $menores$ y voluntarios adultos que actúan como niños para poder capturar sospechosos. Utilizaron dos fuentes donde en la primera fuente pudieron tomar 200 conversaciones funcionales en un formato de *.txt, por otra parte en la segunda fuente solo pudieron tomar 1000 archivos XML, ya que contenían entrenamiento y evaluación de English Courpous estos archivos.

2. Data Pre-processing : Los investigadores tienen que hacer pre-procesamiento de los datos ya que ellos definen un estándar en su Framework y es que todo dato para el entrenamiento serán en archivos XML, lo que implica que los dataset que son encontrados en diferentes redes sociales, no van a estar en su estándar definido, sino en otros tipos de extensiones como HTML o .db, lo que implica que se tienen que convertir estos archivos al formato definido. Stylus Studio es la herramienta que les ayuda convertir archivos *.txt, HTML o .db a XML.

   Esquema:

```
<author uuid="Bpm020" lang="en">
    <conversation id = "11">
        <!-- the TXT file -->
    </conversation>
</author>
```

3. Feature Extraction and Selection : La extraccion y selección de características se basa en seleccionar una pequeña proporción de atributos mediante el mapeo de los datos en el conjunto de datos de entrenamiento. Ellos aplican diferentes tecnicas para la seleccion de dichos atributos, tales como

   a) El análisis de componentes principales (PCA)
   b) El análisis discriminante lineal (LDA),
   c) La proyección de preservación de la localidad (LPP),
   d) La incrustación por proximidad estocástica (SPE)
   e) La incrustación por vecindad estocástica (SNE)
   f) Los enfoques de selección de características difusas (FRFS)

   Asi logrando caracteristicas asociadas con logica difusa y a travez de esta poder desarrollar la propuesta de la deteccion de grooming online. La cantindad escogida para la clasificación binaria fue de 150 y 30 para la clasificación multi-etiqueta.

4. Feature Normalisation : Debido a que los rangos de las respuestas de las características propuestas en el ítem anterior puede variar significativamente, se tienen que normalizar para que no se deteriore el rendimiento de la clasificación en la etapa posterior Framework, no se propone un método en especifico para la normalización ya que resalta la flexibilidad del desarrollo.

5. Classification : Debido a que hay demasiado ruido en la clasificacion de los usuarios, los investigadores utilizaron dos variantes nueva sobre SVM(LCDFTSVM , RCDFTSVM), precisamente, clasificadores no lineales y lineales (i.e. usando el kernel RBF) coordinate descent fuzzy twin support vector machine (CDFTSVM), el cual les ayudaba a eliminar el ruido mediante una función de membresía difusa y reducir la complejidad de los cálculos utilizando una coordenada estrategia de descenso, asi mismo ellos plantean dos ecuaciones diferentes al convecional SVM (i.e. $y = ax + b$) donde utilizan dos diferentes clases los positivos '+' y lo negativos '-', para asi lograr dos cojuntos difusos de posibles soluciones, brevemente, en otras palabras, posibles soluciones de coordenadas (a^s_+, a^s_+) y (a^s_-, a^s_-) de dos planos de decisión no paralelos (i.e $a^T_+ x + b_+ = 0$ y $0 = a^T_- x + b_-$), concluyendo asi la función de membresia para cada muestra de entrenamiento en cada una de las dos clases anterior descritas.

$$y = \arg\min_{\pm} \frac{\left| {a^*_\pm}^T x + b^*_\pm \right|}{||a^*_\pm||}.$$

FIGURA 7.3: Ecuacion de membresia, Tomado de [32].

7.2.3. Uso de logica difusa en C^3-Sex .

En la primera versión de este proyecto se planeto la ecuación (7.1) la cual tenia como objetivo poder clasificar un usuario como *Indifferent* , *Interested* o *Pervert* ; Ahora se plantea un nuevo escenario de clasificación del sospechoso con lógica difusa puesto que se deben plantear las variables difusas, conjuntos difusos y la función de membresía respectiva, debemos darle un nuevo significado a la métricas y una transformación respectiva.

$$\varphi = \frac{1}{1 + \exp\left(-\frac{R \cdot E \cdot O}{\tau \cdot \delta_1} + \delta_2\right)} \tag{7.1}$$

Función triangular de membresia sobre Child pornography matched rules

Para poder aplicar la métrica Child pornography matched rules en el nuevo modelo de clasificación,se cambio el significado de R y además su valor. Ahora R se define como el numero de reglas acertadas en la conversación, por consiguiente , también se define un conjunto difuso respecto a este nuevo significado de R.

En la figura 7.4 se refleja una función triangular de membresia sobre Child pornography matched rules la cual tiene 3 tipos de conjuntos difusos:

- Less : Una funcion de membresia con un conjunto difuso definido entre [0 %-50 %] de la totalidad de conversacion, teniendo un pico en 0 representando un total de zero reglas acertadas en la interaccion del C^3-Sex y el sospechoso.
- Medium : Una funcion de membresia con un conjunto difuso definido entre [0 %-100 %] de la totalidad de conversacion, teniendo un pico cuando la cantidad de reglas acertadas correspoden a la mitad de la cantidad de la conversacion total, asi representando un más interes que el conjunto less anteriormente descrito.
- High : Una funcion de membresia con un conjunto difuso definido entre [50 %-100 %] de la totalidad de conversacion, teniendo un pico cuando la cantidad de reglas acertadas es igual a la cantidad de conversacion total, asi representando un interes alto y un perfil a fin de lo que se espera.

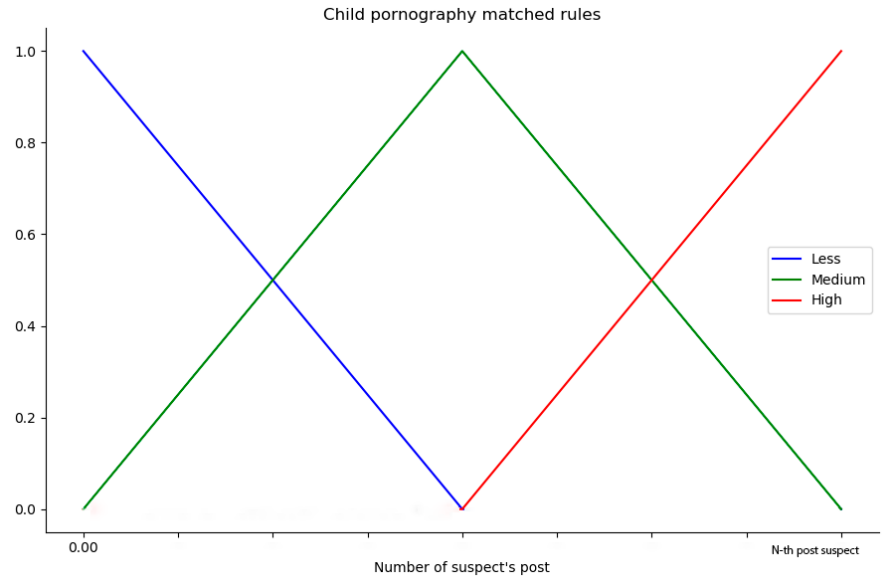


FIGURA 7.4: Función triangular de membresia sobree Child pornography matched rules

Función triangular de membresia sobre *Average response time*

Definir función triangular de membresia sobre *Average response time* fue complejo y sencillo ya que la metrica τ estaba definida como el tiempo transcurrido entre la generación de una respuesta por parte de el ACE y la siguiente interacción del sospechoso, considerada como la respuesta a la respuesta anterior, dado que este significado, se pudo crear una funcion de membresia respecto a esto, ya que en la implementacion tiene un tiempo establecio de 3 mininutos equivalente a 180 sec. por lo tanto ya se tenia definido un universo ya defino para partir de ahi y formar los conjuntos difusos.

En la figura 7.8 se refleja una función triangular de membresia sobre *Average response time* la cual se definieron los siguientes 3 tipos de conjuntos difusos:

- High : Una variable difusa en el conjunto difuso donde se cataloga alto, si esta entre la promedio de respuesta entre el tiempo de 0sec - 90sec , con la condición de que si cada vez el valor τ tienda a 0 será un sospechoso que busca un contenido rapido o esta interesado en la conversacion.

- Medium : Una variable difusa en el conjunto difuso donde se cataloga media, si esta entre la promedio de respuesta entre el tiempo de 0sec - 180sec , con la condición de que si cada vez el valor τ tiende a 90sec será un sospechoso que busca tal vez una conversacion

- Less : Una variable difusa en el conjunto difuso donde se cataloga baja, si esta entre la promedio de respuesta entre el tiempo de 90sec - 180sec , con la condición de que si cada vez el valor τ tiende a 180sec será un sospechoso que busca tal vez no busca una conversacion a fin a nuestra tema o no se reciben mensajes por parte de el.

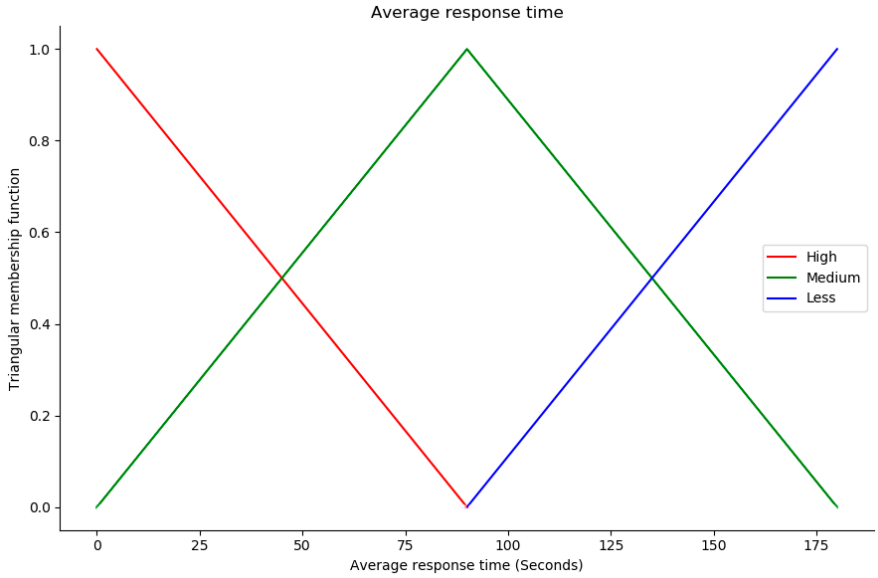


FIGURA 7.5: Función triangular de membresia sobre *Average response time*

Función triangular de membresia sobre *Recognized emotions*

La funcion triangular definida sobre *Recognized emotions*, fue basada en la forma de cómo se calculaba anteriormente para la ecuación (7.1) $E = \frac{\sum_{i=0}^{N} E_i}{N}$, ahora nosotros tomamos la misma definicion de E, mientras que E tiende más a 1 el sospechoso refleja emociones positivas sobre la conversacion a cambio si tiende más a 0, representara representa emociones negativas sobre el tema.

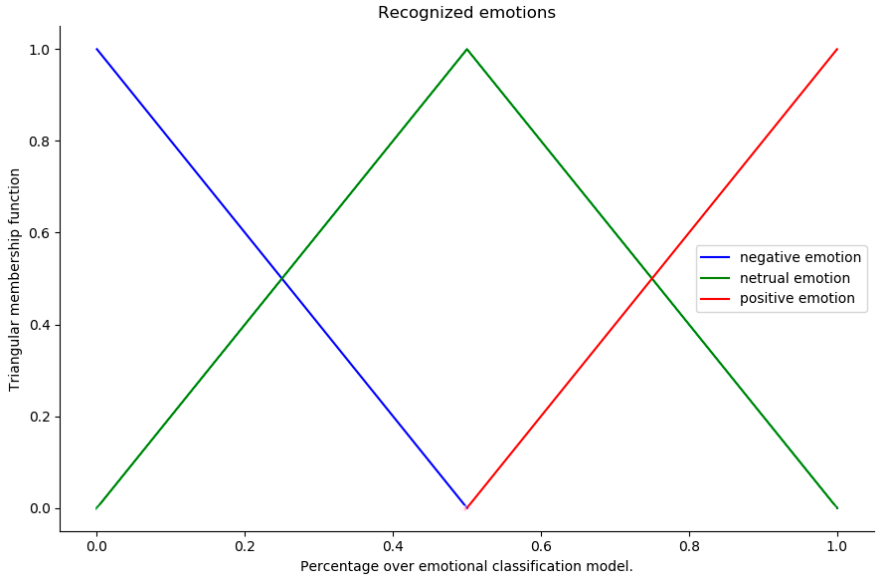


FIGURA 7.6: Función triangular de membresia sobre *Recognized emotions*

Función triangular de membresia sobre *Opinion classification*

La funcion triangular definida sobre *Opinion classification*, fue basada en la forma de cómo se calculaba anteriormente para la ecuación (7.1) $O = \frac{\sum_{i=0}^{N} O_i}{N}$, ahora nosotros tomamos la misma definicion de O, mientras que o tiende más a 1 el sospechoso refleja conformidad cpm la conversacion a cambio si tiende más a 0, representara representa incoformidad sobre el tema.

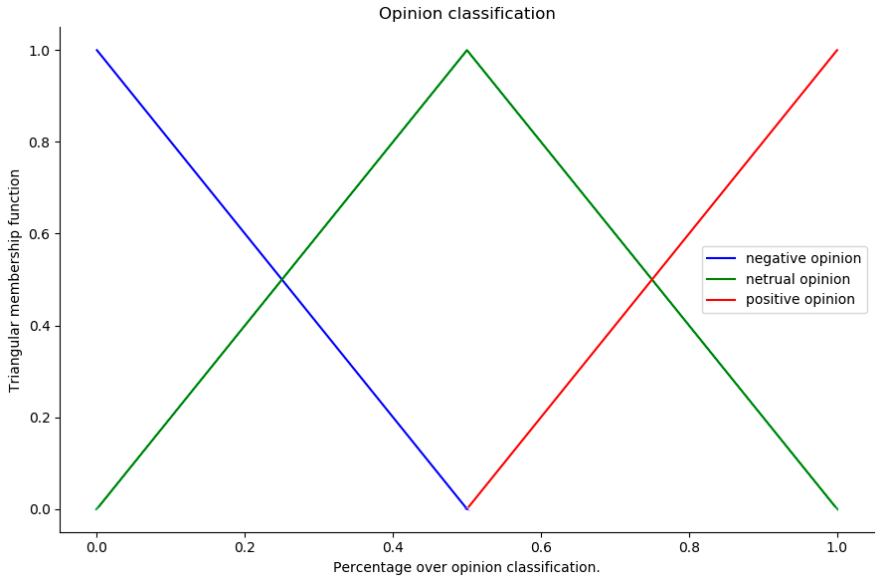


FIGURA 7.7: Función triangular de membresia sobre *Opinion classification*

Cómo clasificar el sospechoso con lógica difusa

Para que las funciones triangulares sean útiles, tenemos que definir la relación difusa entre las variables de entrada y salida, es decir, se necesitan crear reglas de tal forma que ayude para la clasificación del sospechoso.

Algunos ejemplos de reglas implementadas en codigo:

- Si el tiempo de respuesta es alto y emocioniones son positivas y la cantidad de reglas acertdas es alto entonces se puede concluir que puede ser un pervertido.
- Si el tiempo de respuesta es medio y las emociones altas entonces puede ser un interesado.
- Si el tiempo de respuesta es medio y las emociones son neutrales y igualmente su opicion entonces puede ser un interesado.

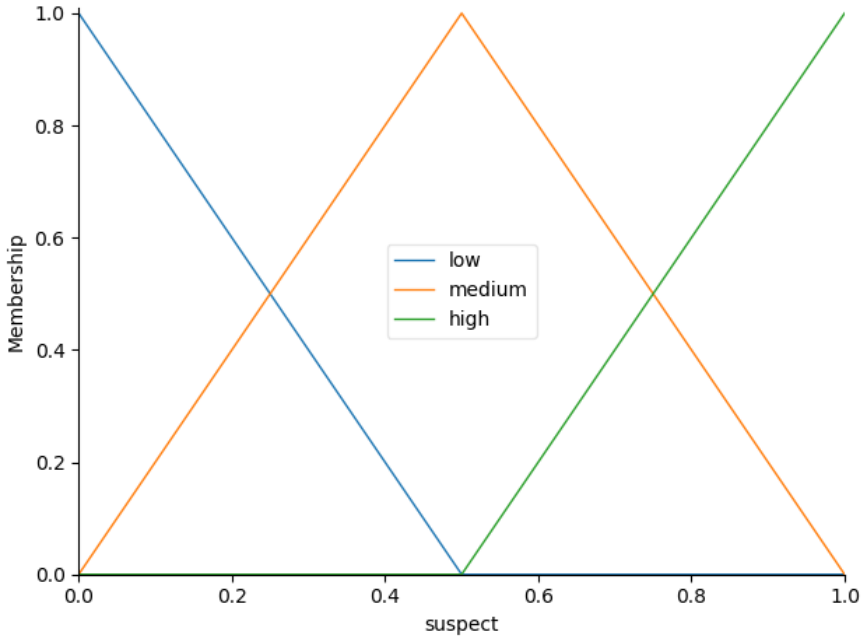


FIGURA 7.8: Función triangular de membresia del clasificador

Recordar que estas reglas pasan por un sistema de control difuso que nos va permitir responder reglas que tal vez no se tengan planteadas, a su vez cuando un valor no se encuentra en dominio de las funciones definidas, se hace una interpolación lineal sobre las funciones.

7.2.4. Lógica Difusa - Experimentos de unidad

A partir de tres tipos de conversaciones diferentes(bajo interes o sin interes,medio interes,alto interes respecto al tema) con al menos 100 interacciones entre el bot y el sospechoso se encontraron los siguientes datos

- En los tres primeros escenarios se evalúo un perfil indiferente respecto al tema de pornografía infantil, como se puede evidenciar en la figura 7.9; La primer gráfica expone un sospechoso indiferente, pero con cualidades como: Un tiempo bajo de respuesta entre un mensaje y otro, desacuerdo respecto al tema y con emociones negativas, en la segunda gráfica se tiene un sospechoso neutro, no refleja agrado respecto al tema y asimismo no demuestra algún sentimiento,y por ultimo, es un caso donde no hubo interacción con el sospechoso.

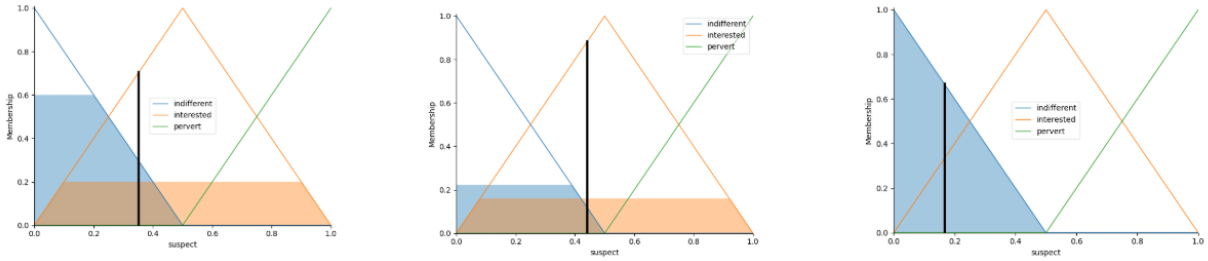


FIGURA 7.9: Experimentos de unidad sobre un sospechoso catalogado como indiferente.

	Conversación 1	Conversación 2	Conversación 3
time	10s	170s	180s
emotion	0.1	0.1	0
opinion	0.08	0.8	0
rules	10	10	0

- En el segundo grupo de experimentos se evalúo un perfil ”interesado´´, como se puede evidenciar en la figura 7.10, la primera gráfica expone un perfil con un nivel de interés alto donde sus tiempos de respuesta son medianos como su cantidad de reglas acertadas y a su vez una métrica de emociones altas, pero su opinión respecto al tema baja; el segundo caso expone un perfil con una cantidad de altas reglas acertadas y igualmente con una opinion respecto al tema muy positiva, pero sus tiempos de respuesta y sus emociones no son tan buenas, ya que el tiempo de respuesta es muy alto y la emociones son muy bajas, por ultimo, se evidencia un perfil "mediano"donde todas las reglas tieden a un valor medio, lo que significa que esta interesado pero no muestra un interes excesivo o un interes muy bajo.

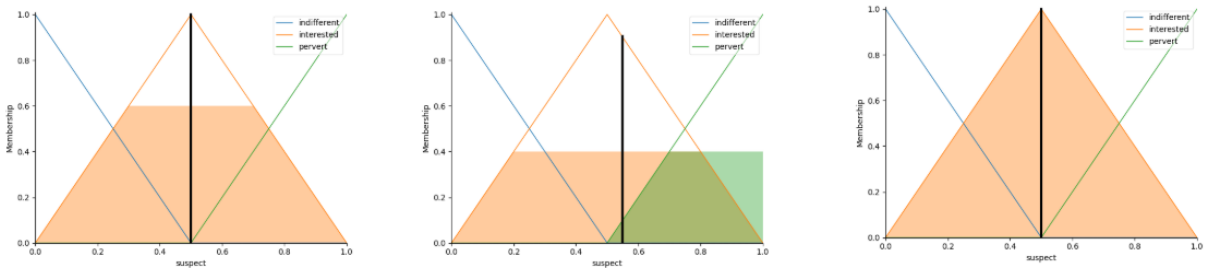


FIGURA 7.10: Experimentos de unidad sobre un sospechoso catalogado como interesado.

	Conversación 4	Conversación 5	Conversación 6
time	90s	160s	90s
emotion	0.7	0.3	0.5
opinion	0.3	0.8	0.5
rules	45	70	50

- Los ultimos experimentos se evalúo un perfil "pervertido´´ ,como se puede evidenciar en la figura 7.11, la primera y la segunda gráfica expone un perfil de sospechoso con respuestas rápidas respecto a cada interacción del bot, asimismo métricas altas en emociones y opiniones, lo único que varia es cuando se evalúa el primer caso debido a que hay una regla difusa que cumple con el mismo formato descrito por la tabla hace que el área coloreada, el tercer caso es un caso de un sospechoso totalmente pervertido donde todas sus métricas son altas.

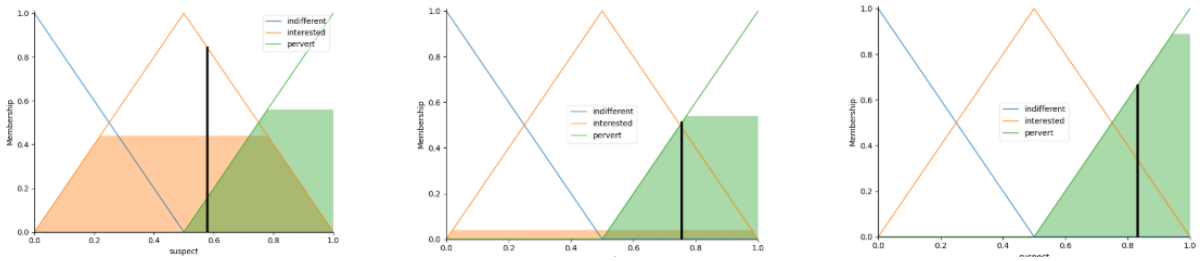


FIGURA 7.11: Experimentos de unidad sobre un sospechoso catalogado como pervertidow .

	Conversación 7	Conversación 8	Conversación 9
time	10s	10s	10s
emotion	0.87	0.77	1
opinion	0.78	0.98	1
rules	98	82	98

Capítulo 8

Conclusiones y trabajos futuros

8.1. Conclusiones

- Se logró mejorar la eficiencia del modelo generativo basado en redes neuronales seq-to-seq por medio del uso de un nuevo conjunto de datos de entrenamiento
- Se logró mejorar el modelo basado en conocimiento por medio de la incorporación de un nuevo conjunto de reglas AIML que permiten conducir la conversación con el sospechoso
- Se logró incorporar un nuevo indicador en el perfilamiento asociado a la transferencia de contenido multimedia
- Se logró las construcción de un artículo para la revista MDPI con el análisis de 6267 conversación sobre el bot.
- Se propuso y se desarrolló un nuevo modelo para la clasificación del sospechoso con ayuda de lógica difusa.

8.2. Trabajos futuros

- Desarrollar y implementar un sistema distribuido con maquinas virtuales para la manipulación del sistema de android con el fin de tener una mejor eficiencia en el perfilamiento de sospechosos.
- Identificar si el sospechoso a que rango de edad pertenece ya que se puede sospechar de alguien cuando realmente puede ser un niño o joven molestando o sorprendido por el tema de la conversación.
- Implementar redes neuronales convulsiónales para la identificar si el contenido transmitido si es alusivo a pornografía infantil.
- Para poder agregar la metrica de contenido ilegal y poder trabajar con logica difusa es necesario hacer un nuevo modelo que sea el encargado de interpretar una imagen con el objetivo de darle un porcentaje de similitud a la imagen en comparacion con un contenido ilegal por esa razón se plantea un nuevo trabajo futuro donde utilizando CNN se puedan categorizar las imagenes y asi darles un porcentaje de similitud.

- Implementar una nueva arquitectura para el bot, ya que al tratar de crear unas nuevas métricas para el análisis, no fue posible por que la arquitectura anteriormente hecha esta muy acoplada sobre las reglas AIML

  - **Uso de Convolutional Neural Networks o CNN**

    Una CNN (i.e. Convolutional Neural Networks) "es un tipo de Red Neuronal Artificial con aprendizaje supervisado que procesa sus capas imitando al cortex visual del ojo humano para identificar distintas características en las entradas que en definitiva hacen que pueda identificar objetos y «ver». Para ello, la CNN contiene varias capas ocultas especializadas y con una jerarquía: esto quiere decir que las primeras capas pueden detectar lineas, curvas y se van especializando hasta llegar a capas más profundas que reconocen formas complejas como un rostro o la silueta de un animal."(Na8 , 2018) [33]

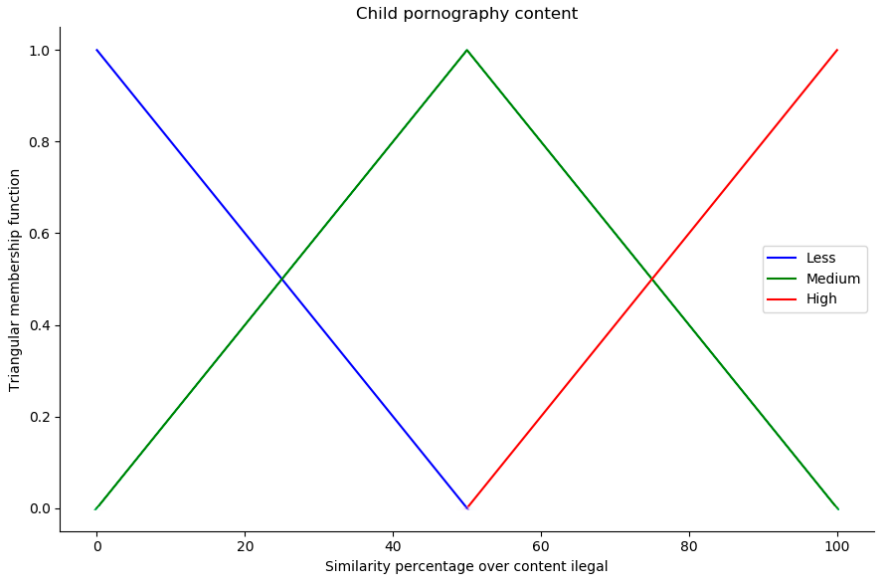


FIGURA 8.1: Función triangular de membresia sobre contenido ilegal

En la figura 8.1 se refleja una función triangular de membresia sobre contenido ilegal la cual tiene 3 tipos de clasificación con sus picos definidos:

- Less : Una funcion de membresia con un conjunto difuso definido entre [0 %-50 %] con un pico en 0 % representando una similitud realmente baja o nula asociada a la pornografia infantil.
- Medium : Una funcion de membresia con un conjunto difuso definido entre [0 %-100 %] con un pico en 50 % representando una similitud realmente mediana asociada a la pornografia infantil.
- High : Una funcion de membresia con un conjunto difuso definido entre [50 %-100 %] con un pico en 100 % representando una similitud alta o identica asociada a la pornografia infantil.

Capítulo 9

Información general del proyecto

9.1. Repositorio del proyecto

El repositorio principal que contiene todos el Software, documentación y este documento se encuentran en GitHub.

Repositorios de GitHub

Hiper-vinculo de los repositorios:
https://github.com/CrkJohn/Snapchat.git

https://github.com/Santiago-Rocha/PGR.got

Para clonar los repositorios es necesario ejecutar el comando en la Terminal (o bien *Git Bash* en Windows):

```
git clone --recurse-submodules
https://github.com/CrkJohn/Snapchat.git

git clone --recurse-submodules
https://github.com/Santiago-Rocha/PGR.git
```

9.2. Guía de Inicio Rápido

9.2.1. C^3-Sex

Objectivos del proyecto

Project Objectives

Este proyecto tiene como objetivo diseñar y construir un Bot con inteligencia sentimental interactiva, capaz de descubrir y revelar el perfil de los ciberdelincuentes en chats masivos en línea, mediante la elaboración de respuestas pertinentes considerando el contenido lingüístico y sentimental.

Además, a continuación se exponen los objetivos específicos que se establecen para describir los resultados previstos del proyecto:

- Diseñar una solución de chatbot basada en la integración de diferentes modelos del dominio de la inteligencia artificial; un modelo basado en la generación y otro en la recuperación.
- Implementar un chatbot capaz de generar respuestas para empatizar con el sospechoso y manipularlo para obtener información útil.
- Hacer un perfil de los sospechosos (mediante el análisis de los sentimientos y la clasificación de las emociones) utilizando los datos obtenidos de la interacción con el Bot con la inteligencia sentimental interactiva.

Cómo instalarlo y configurarlo

Requerimientos del proyecto

	Mínimo	Màximo
CPU:	Cores: 2; Frequency: 1,70GHz; Cache: 3MB	Cores: 4; Frequency: 2,80GHz; Cache: 6MB
Memory:	8GB RAM	12GB RAM
Free Disk Space:	2GB	2GB
Operative System:	Windows 7/8/10	Windows 7/8/10
GPU:	None	Nvidia GeForce GTX 950
Software:	Python 3.6.2; Pip; Google Chrome;	Python 3.6.2; Pip; CUDA Toolkit 10.1; Google Chrome
Network Access:	Yes	Yes

Requirimientos de librerías de python.

Name	Description
selenium	It provides the Selenium WebDriver API, which allows the project to connect to a browser natively.
tensorflow	As a Machine Learning library, the project uses it to build the recurrent neural network (RNN), which simulates the generative conversational agent
pandas	A data analysis library used to manage and read certain data structures, such as csv and dataframes
sklearn	In addition to making use of tensorflow, sklearn is available for the adoption of ML algorithms, such as the Bayesian network in the opinion classification model.
nltk	The natural language tool gives the project the advantage of tokenize words, its used in the different models.
bs4	The Beautiful Soup library allows you to extract data from HTML files, the project use it to extract the knowledge of a GitHub repository, for mapping Slangs in the English language.
python-aiml	This library serves as an inference engine to read xml files as the AIML form, which contain the entire knowledge base of the retrieval-based conversational module

Para descargar cualquiera de los paquetes se recomienda usar el comando pip, reemplazando la parte del nombre del paquete:

Comandos

```
pip install package-name
```

En la shell de python

```
import nltk
nltk.download()
```

Antes de usar el proyecto

Antes de comenzar a utilizar el proyecto y para asegurarse de que cada funcionalidad funciona correctamente, debe tener en cuenta las siguientes instrucciones.

- Para utilizar el módulo ACA donde reside toda la lógica del agente conversacional, debe asegurarse de que está entrenado, para ello, los siguientes archivos deben estar en la carpeta ./ACA/Data/Resultados

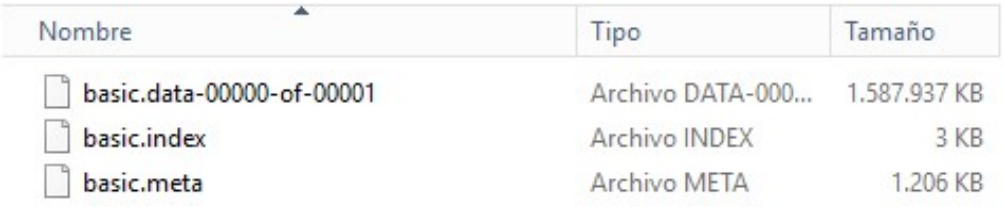


FIGURA 9.1: Representación del entrenamiento de C^3-Sex

Para generarlos hay que entrenar el modelo generativo (lo que requerirá mucho tiempo, dependiendo de los datos de entrenamiento que se quieran utilizar) o utilizar los datos ya entrenados que se pueden encontrar Aqui

Uso de los módulos de software

La siguiente información se proporciona para que el usuario pueda utilizar cada uno de estos servicios de forma individual, aprovechando todo lo que puedan ofrecer.

Artificial Conversational Agent/Entity (ACA)

Este módulo de conversación puede ser usado de diferentes maneras, dependiendo de lo que quieras hacer:

1. **Entrenamiento modelo**: Durante esta etapa, el modelo se entrenará en base al conjunto de datos y el vocabulario proporcionado. Para ello, debe ejecutar el siguiente comando desde la raíz del proyecto.

   Comandos

   ```
   py ACA/chatbot/bottrainer.py
   ```

Ten en cuenta que el entrenamiento puede llevar mucho tiempo (dependiendo de los parámetros con los que quieras entrenar). Estos parámetros pueden ser modificados en ACA/-Data/Corpus/hparams.json. Asimismo, el conjunto de datos que se proporciona para el entrenamiento puede modificarse añadiendo archivos en ACA/Data/Corpus/Argument0 con el formato Pregunta-Respuesta, tal y como se muestra en los datos limpios que hemos utilizado. Una vez que el entrenamiento ha comenzado, debe asegurarse de que los siguientes archivos sean creados en ACA/Datos/Resultado, los cuales son usados para pruebas y predicciones.

- basic.data-00000-of-00001
- basic.index

2. **Interfaz de prueba**: Esta etapa proporciona una interfaz simple y una basada en la web. Tenga en cuenta que los archivos ACA/Data/Corpus/vocab.txt y ACA/Data/Rules/hotstartup.aiml son necesarios para la ejecución. Para ver el rendimiento del modelo, debe ejecutar el siguiente comando y esperar a que aparezca el indicador >:

Comandos

```
py ACA/chatbot/botui.py
```

Analisis de sentimientos

Para el módulo de análisis de sentimientos, se pueden utilizar los dos clasificadores; opinión y emociones, el primero desarrollado en Python y el segundo en R.

1. **Clasificador de opinión:** Puede acceder al archivo SA_Module/sentimentAnalysis.py desde la raíz, y hacer uso de las siguientes funciones, teniendo en cuenta sus entradas.

Comandos

```
predict(data) # data should be a response
predict_proba(data) # Same way, a sentence and the output
↪ differs 'cause returns the probability
sa_measure(replies_file) # A text file with
↪ sentences/responses to be classified with probabilities
```

2. **Clasificador de emociones** En el caso del módulo de emociones implementado en R, se puede utilizar el modelo entrenado por el archivo EC_Module/emotional_classifier.py, el cual ejecuta un script R para generar predicción para aquellas frases que serán clasificadas.

Comandos

```
ec_measure(replies_file) # A text file with sentences to
↪ classify, returns a list with positive (1) emotions or
↪ negative (0)
```

Hay que tener en cuenta que los dos módulos además de devolver lo anterior, generan un archivo con las respectivas opiniones y emociones por frase, estos archivos tienen el mismo nombre que el proporcionado, pero añadiendo un "_SA z ´´ _EC ´´ al final respectivamente.

9.2.2. Uso general del software

Para utilizar el software como un todo (que integra todos los módulos mencionados) Es necesario ejecutar el siguiente comando desde la raíz del proyecto.

Sin embargo, antes de hacerlo, debe asegurarse de las siguientes indicaciones:

- Comprueba que la version de tu navegador Google Chrome y del ChromeDriver sea la misma.
- Debes cambiar en el archivo ./ACA/Data/Rules/illegalcontent/hotmaterial.aiml todos aquellos campos sobre el nombre de usuario de Snapchat (por defecto es @p_ramirezxxx) que vayas a utilizar.
- Para configurar el dispositivo que se va a utilizar en la interacción con snapchat se debe modificar las siguientes partes del codigo con la información del dispositivo

```
1  public class Server{
2      //Version de Android
3      setCapability(MobileCapabilityType.PLATFORM_VERSION, "5.0.2");
4
5      //Nombre del dispositivo
6      setCapability(MobileCapabilityType.DEVICE_NAME, "LG Magna LTE");
7
8      //ID del dispositivo
9      setCapability(MobileCapabilityType.UDID, "4TIBNREMDAMFPB9L");
10 }
```

- La sección de Snapchat tiene unicamente el codigo fuente, es decir que debe compilarse antes de integrarlo con el bot. para ello nos dirigmos a la carpeta raiz del proyecto de Snapchat ../../Snpachat y ejecutamos el siguiente comando

Comandos

```
mvn clean compile
```

Una vez configurado correctamente el ambiente para iniciar el bot debemos ejecutar el siguiente comando desde la carpeta raiz del proyecto ../../C3-sex

Comandos

```
py main.py
```

Bibliografía

[1] Yuriy Gapanyuk y col. «The Hybrid Chatbot System Combining Q&A and Knowledge-base Approaches». En: *7th International Conference on Analysis of Images, Social Networks and Texts (AIST 2018)*. 2018, págs. 42-53.

[2] Iulian V Serban y col. «A deep reinforcement learning chatbot». En: *arXiv preprint arXiv:1709.02349* (2017).

[3] Gábor Tatai y col. «The chatbot who loved me». En: *proceedings of the ECA workshop of AAMAS*. 2003.

[4] Robert Plutchik. «The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice». En: *American scientist* 89.4 (2001), págs. 344-350.

[5] Qiao Qian y col. «Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation.» En: *IJCAI*. 2018, págs. 4279-4285.

[6] Bob Heller y col. «Freudbot: An investigation of chatbot technology in distance education». En: *EdMedia: World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education (AACE). 2005, págs. 3913-3918.

[7] B. A. Sabbagh y col. «A prototype For HI2Ping information security culture and awareness training». En: *2012 International Conference on E-Learning and E-Technologies in Education (ICEEE)*. 2012, págs. 32-36. DOI: 10.1109/ICeLeTE.2012.6333397.

[8] Stewart Kowalski, Katarina Pavlovska y Mikael Goldstein. «Two Case Studies in Using Chatbots for Security Training». En: *Information Assurance and Security Education and Training*. Ed. por Ronald C. Dodge y Lynn Futcher. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, págs. 265-272. ISBN: 978-3-642-39377-8.

[9] Jennifer Hill, W Randolph Ford e Ingrid G Farreras. «Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations». En: *Computers in Human Behavior* 49 (2015), págs. 245-250.

[10] Hao Zhou y col. «Emotional chatting machine: Emotional conversation generation with internal and external memory». En: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[11] P. Zambrano y col. «BotHook: An option against Cyberpedophilia». En: *2017 1st Cyber Security in Networking Conference (CSNet)*. 2017, págs. 1-3. DOI: 10.1109/CSNET.2017.8241994.

[12] Kemal Veli Açar. «Webcam Child Prostitution: An Exploration of Current and Futuristic Methods of Detection». En: *International Journal of Cyber Criminology* 11.1 (2017), págs. 98-109. DOI: 10.5281/zenodo.495775.

[13] P. Anderson y col. «An Intelligent Online Grooming Detection System Using AI Technologies». En: *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2019, págs. 1-6. DOI: 10.1109/FUZZ-IEEE.2019.8858973.

[14] Fergyanto E Gunawan, Livia Ashianti y Nobumasa Sekishita. «A simple classifier for detecting online child grooming conversation». En: *Telkomnika Telecommunication, Computing, Electronics and Control* 16.3 (2018), págs. 1239-1248. ISSN: 1693-6930.

[15] Mohammadreza Ebrahimi. «Automatic Identification of Online Predators in Chat Logs by Anomaly Detection and Deep Learning». Tesis de mtría. 1515 Ste-Catherine St. W. Montreal, Canada: Concordia University, abr. de 2016.

[16] Maral Dadvar y Franciska de Jong. «Cyberbullying Detection: A Step toward a Safer Internet Yard». En: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12 Companion. Lyon, France: Association for Computing Machinery, 2012, 121–126. ISBN: 9781450312301. DOI: 10.1145/2187980.2187995.

[17] Maral Dadvar y col. «Towards User Modelling in the Combat against Cyberbullying». En: *Natural Language Processing and Information Systems*. Ed. por Gosse Bouma y col. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, págs. 277-283. ISBN: 978-3-642-31178-9.

[18] Carlos Laorden y col. «Negobot: A Conversational Agent Based on Game Theory for the Detection of Paedophile Behaviour». En: *International Joint Conference CISIS'12-ICEUTE´12-SOCO´12 Special Sessions*. Ed. por Álvaro Herrero y col. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, págs. 261-270. ISBN: 978-3-642-33018-6.

[19] Dimitrios Michalopoulos, Ioannis Mavridis y Marija Jankovic. «GARS: Real-time system for identification, assessment and control of cyber grooming attacks». En: *Computers & Security* 42 (2014), págs. 177 -190. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2013.12.004.

[20] Pete Burnap y col. «Detecting tension in online communities with computational Twitter analysis». En: *Technological Forecasting and Social Change* 95 (2015), págs. 96 -108. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2013.04.013.

[21] Mike Thelwall y col. «Sentiment strength detection in short informal text». En: *Journal of the American Society for Information Science and Technology* 61.12 (2010), págs. 2544-2558. DOI: 10.1002/asi.21416.

[22] Harvey Sacks, Gail Jefferson y Emanuel Schegloff. *Lectures on Conversation*. Vol. I - II. The address: Wiley-Blackwell, jun. de 2010. ISBN: 978-1-444-32830-1.

[23] F. E. Gunawan y col. «Detecting online child grooming conversation». En: *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*. 2016, págs. 1-6. DOI: 10.1109/KICSS.2016.7951413.

[24] Maxime Meyer. «Machine learning to detect online grooming». Tesis de mtría. Uppsala University, Department of Information Technology, 2015, pág. 64.

[25] Moreno Rodríguez Sebastián Murcia Triviño Jossie Esteban. *Profiling Criminals Through Emotional Machine Learning*. 2019. URL: https://repositorio.escuelaing.edu.co/handle/001/972.

[26] bshao001. *ChatLearner*. URL: https://github.com/bshao001/ChatLearner.

[27] Jossie Murcia Triviño y col. «C3-Sex: a Chatbot to Chase Cyber perverts». En: *4th IEEE Cyber Science and Technology Congress*. Fukuoka, Japan, 2019, págs. 32-33. ISBN: 978-1-7281-3024-8. DOI: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00024. URL: https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00024.

[28] Android company. *UI Automator*. URL: https://developer.android.com/training/testing/ui-automator#ui-automator-viewer.

[29] Longzhi Yang Philip Anderson Zheming Zuo. «Negobot: A conversational agent based on game theory for the detection of paedophile behaviour». En: (2013).

[30] Carlos González Morcillo. *Lógica Difusa Una introducción práctica*, pág. 4.

[31] Tomás Arredondo Vidal. *Logica Difusa*. 2014. URL: http://profesores.elo.utfsm.cl/~tarredondo/info/soft-comp/Introduccion-a-la-Logica-Difusa.pdf.

[32] Longzhi Yang Philip Anderson Zheming Zuo. «An Intelligent Online Grooming Detection System Using AI Technologies». En: (2019).

[33] Na8. *¿Cómo funcionan las Convolutional Neural Networks?* 2018. URL: https://www.aprendemachinelearning.com/como-funcionan-las-convolutional-neural-networks-vision-por-ordenador/.

[34] Jossie Murcia Triviño y col. «C3-Sex: a Chatbot to Chase Cyber perverts». En: *4th IEEE Cyber Science and Technology Congress*. Fukuoka, Japan, 2019, págs. 50-57. ISBN: 978-1-7281-3024-8. DOI: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00024. URL: https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00024.

[35] Gary Ericson y col. *Team Data Science Process Documentation*. Inf. téc. Microsoft Azure, 2017, pág. 456.

[36] P. Zambrano y col. «BotHook: An option against Cyberpedophilia». En: *2017 1st Cyber Security in Networking Conference (CSNet)*. 2017, págs. 1-3. DOI: 10.1109/CSNET.2017.8241994.

[37] Carlo Strapparava y Rada Mihalcea. «Semeval-2007 task 14: Affective text». En: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. 2007, págs. 70-74.

[38] Sepp Hochreiter y Jürgen Schmidhuber. «Long Short-Term Memory». En: *Neural Comput.* 9.8 (1997), págs. 1735-1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

[39] Mike Walker. «Hype Cycle for Emerging Technologies, 2018». En: *2018 Hype Cycles: Riding the Innovation Wave* (2018).

[40] Edward Loper y Steven Bird. «NLTK: The Natural Language Toolkit». En: *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*. 2002.

[41] Simon Keizer y Harry Bunt. «Multidimensional Dialogue Management». En: *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Sydney, Australia: Association for Computational Linguistics, 2006, págs. 37-45. URL: http://aclweb.org/anthology/W06-1306.

[42] Bobby Filar, Richard Seymour y Matthew Park. «Ask Me Anything: A Conversational Interface to Augment Information Security Workers.» En: *SOUPS*. 2017.

[43] Geoff Dean. «Framing the Challenges of Online Violent Extremism:"Policing-Public-Policies-Politics" Framework». En: *Violent Extremism: Breakthroughs in Research and Practice*. IGI Global, 2019, págs. 302-335.

[44] Al Rahman, Abdullah Al Mamun y Alma Islam. «Programming challenges of chatbot: Current and future prospective». En: *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. Dic. de 2017, págs. 75-78. DOI: 10.1109/R10-HTC.2017.8288910.

[45] Pan Juin Yang Jonathan, Chun Che Fung y Kok Wai Wong. «Devious Chatbots - Interactive Malware with a Plot». En: *Progress in Robotics*. Ed. por Jong-Hwan Kim y col. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, págs. 110-118. ISBN: 978-3-642-03986-7.

[46] Dimitrios Kotzias y col. «From group to individual labels using deep features». En: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, págs. 597-606.

[47] Max Taylor y Ethel Quayle. *Child Pornography: An Internet Crime*. Feb. de 2003. ISBN: 978-1583912447.

[48] Emily D. Gottfried, Emily Knight Shier y Abby L. Mulay. «Child Pornography and Online Sexual Solicitation». En: *Current Psychiatry Reports* 22.3 (2020), pág. 10. ISSN: 1535-1645.

[49] William Adams, Abigail Flynn y col. *Federal prosecution of commercial sexual exploitation of children cases, 2004-2013*. US Department of Justice, Office of Justice Programs, Bureau of Justice …, 2017.

[50] Janis Wolak, Marc Liberatore y Brian Neil Levine. «Measuring a year of child pornography trafficking by U.S. computers on a peer-to-peer network». En: *Child Abuse Neglect* 38.2 (2014), págs. 347 -356. ISSN: 0145-2134. DOI: https://doi.org/10.1016/j.chiabu.2013.10.018. URL: http://www.sciencedirect.com/science/article/pii/S0145213413003232.

[51] S. Raaijmakers. «Artificial Intelligence for Law Enforcement: Challenges and Opportunities». En: *IEEE Security Privacy* 17.5 (2019), págs. 74-77. ISSN: 1558-4046. DOI: 10.1109/MSEC.2019.2925649.

[52] T. J. McIntyre. «Blocking child pornography on the Internet: European Union developments». En: *International Review of Law, Computers & Technology* 24.3 (2010), págs. 209-221. DOI: 10.1080/13600869.2010.522321.

[53] Tony Krone. *International Police Operations Against Online Child Pornography*. English. Vol. 2005. Trends and Issues in Crime and Criminal Justice 296. Australia: Australian Institute of Criminology, 2005. ISBN: 064253876X.

[54] A. Ulges y A. Stahl. «Automatic detection of child pornography using color visual words». En: *2011 IEEE International Conference on Multimedia and Expo*. 2011, págs. 1-6. DOI: 10.1109/ICME.2011.6011977.

[55] Jianfeng Gao, Michel Galley y Lihong Li. «Neural Approaches to Conversational AI». En: *Foundations and Trends® in Information Retrieval* 13.2-3 (2019), págs. 127-298. ISSN: 1554-0669. DOI: 10.1561/1500000074. URL: http://dx.doi.org/10.1561/1500000074.

[56] Amit Gupte y col. «Comparative study of classification algorithms used in sentiment analysis». En: *International Journal of Computer Science and Information Technologies* 5.5 (2014), págs. 6261-6264.

[57] Mike Walker. «Hype Cycle for Emerging Technologies, 2018». En: *2018 Hype Cycles: Riding the Innovation Wave* (2018).