ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

PROFILING CRIMINALS THROUGH EMOTIONAL MACHINE LEARNING

Realizado por

Jossie Esteban Murcia Triviño y Sebastián Moreno Rodríguez

Director

Daniel Orlando Díaz López

Programa de Ingeniería de Sistemas

22 de mayo de 2019

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

Resumen

Programa de Ingeniería de Sistemas

Realizado por Jossie Esteban Murcia Triviño y Sebastián Moreno Rodríguez

Este libro presenta el desarrollo del proyecto de grado *"Profiling criminals through Emotional Machine Learning"*. El libro contiene el desarrollo de la teoría investigada y la implementación realizada durante la asignatura Proyecto de Grado. Se compone de la investigación en implementación de un modelo conversacional generativo, uno de recuperación basado en conocimiento, con el propósito de generar un chatbot capaz de interactuar con sospechosos para perfilar sus intereses respecto a un tema de interés, en nuestro caso la venta o distribucción de pronografia infantil, un modelo de análisis de opinión y el uso de uno basado en la clasificación de emociones, con el objetivo de realizar el análisis de los sospechosos.

Índice general

Lista de Tablas

Lista de Figuras

Abreviaciones

ECI Escuela Colombiana de Ingeniería Julio Garavito

NLP Natural Language Processing

LSTM Long Short-Term Memory

AIML Artificial Intelligence Markup Language

SOC Security Operation Center

AI Artificial Intelligence

KBS Knowledge Based Systems

ML Machine Learning

Introducción

Un chatbot es la idealización y representación a la idea de un agente de conversación computacional capaz de interactuar con usuarios, con el fin de mantener una interacción humano-máquina de forma aguda, ya sea en un dominio específico como la mayoría de las implementaciones de chatbots a través de los años o en un dominio más abierto como se ha investigado actualmente. Usualmente, su funcionamiento está basado en la noción de dar respuesta a una pregunta o comentario que formule el usuario. Visto de una manera más simple, el bot de conversación habitual lo único que busca es responder a temas determinados, considerados por su inventor, sin dar paso a la incertidumbre de dar respuesta a lo desconocido. Afortunadamente, el panorama y los retos que actualmente presenta la inteligencia artificial en el campo del Procesamiento del Lenguaje Natural (NLP) traen consigo el desarrollo de investigaciones que ayudan a indagar más sobre la capacidad de replicar el lenguaje humano a partir de interpretar los estímulos asociados, y consecuentemente, generar respuestas más empáticas similares a las del ser humano, de manera que se extienda el ámbito de la interacción hacia un dominio abierto y adaptativo. Tal es el caso de trabajos recientes con redes neuronales basadas en memoria profunda a corto plazo (LSTM), y los estudios recién originados para el campo de la inteligencia artificial general.

Actualmente los trabajos de chatbots aplicados a CiberSeguridad están llevados bajo el campo del dominio específico. Entre los ejemplos de tareas particulares que busca solucionar el agente conversacional están, por un lado, dar soporte a Analistas de Incidentes en Centros de Operaciones de Seguridad (SOC) integrando funciones de detección de ataques en progreso, alertas al usuario y dar sugerencias a soluciones sobre ataques (e.g Artemis Endgame), y por otro lado,

existen otros chatbots con un enfoque en otorgar asistencia y asesoría a una víctima de Ciberacoso al respecto de sus derechos (e.g. Akancha). Sin embargo, gracias a las nuevas investigaciones respecto a un ambiente de conversación de dominio abierto, consideramos que es posible generar una estrategia para la obtención de información de individuos o posibles cibercriminales por medio de un bot entrenado en un ambiente de conversación abierto (imitando en la mejor medida el comportamiento humano) y a su vez con patrones de ingeniería social, para darle la capacidad de perfilar sospechosos.

Según el Hype Cycle para tecnologías emergentes de Gartner, las plataformas de inteligencias artificiales conversacionales y los robots inteligentes son tecnologías innovadoras originadas recientemente. Lo anterior, se puede ver reflejado en los diferentes concursos que han existido desde hace varios años y han permitido la evolución de los chatbots hasta una instancia donde sea más difícil diferenciarlos o distinguirlos de un usuario humano (de manera mas tecnica, que cumplan el criterio del Test de Turing). Algunos de los concursos más importantes relacionados con pruebas de desempeño de chatbots son; Loebner Prize, Chatterbox challenge, Alexa Prize y Atos IT Challenge.

Este libro tiene como propósito mostrar el paso a paso del trabajo realizado en el proyecto de grado *Profiling criminals through Emotional Machine Learning*, el cual tuvo origen en la idea de presentar una herramienta de perfilamiento de sospechosos que apoye la labor de las agencias de la ley en la lucha contra diferentes amenazas criminales.

El proyecto *Profiling criminals through Emotional Machine Learning* contempla la realización de 2 fases:

En la primera fase del proyecto se realizó la investigación del estado del arte y el desarrollo del marco teórico. Se revisaron temas relacionados a 2 corrientes

principales de la Inteligencia Artificial: Machine Learning y Knowledge-Based Systems, sobre las cuales se trataron los modelos sequence-to-sequence y basados en reglas. Se analizó la implementación de los modelos y se exploraron distintas aproximaciones de implementación.

En la segunda fase del proyecto se realizó la implementación tanto del modelo de Machine Learning como del basado en reglas, para este último se planteó una definición básica de reglas para el caso de aplicación contra distribución de pornografía infantil.

1. Definición del proyecto.

1.1 Objetivo General

Construir un Bot capaz de descubrir el perfil de cibercriminales en chats online masivos, a través de la elaboración de respuestas pertinentes al contenido lingüístico y sentimental integrando el uso de métricas basadas en el aspecto emocional del sospechoso.

1.2 Objetivos específicos

1. Detectar, por medio de inteligencia artificial bajo el campo del Procesamiento del Lenguaje Natural, el contenido, las opiniones y emociones de lo que escriba el sospechoso.
2. Analizar el contenido, para generar una respuesta adecuada influenciada al mismo tiempo, por un conocimiento en Ingeniería Social proporcionado por expertos, con el fin de empatizar con el posible criminal y manipularlo hasta el punto de conocer sus intenciones o vínculos con algún tipo de ilegalidad.
3. Perfilar sospechosos (análisis de sentimientos y clasificación de emociones) utilizando los datos de la interacción realizada con el Bot con inteligencia sentimental interactiva.

1.3 Logros

1. Comprensión: Revisión del estado del arte de modelos de machine learning y basados en conocimiento aplicables a chatbots.

2. Estructuración: Construcción de la arquitectura de solución de chatbot, bajo la integración de diferentes modelos.
3. Puesta en Marcha: Implementación del prototipo de chatbot capaz de interactuar con un sospechosos en el escenario o dominio específico de distribución de pornografía infantil.
4. Análisis: Se propusieron métricas de evaluación de conversaciones de forma subjetiva de acuerdo a lo que se consideró relevante evaluar, sin embargo, la métrica logró obtener buenos resultados en su medición.

2. Estado del Arte

Basado en nuestro estudio de la literatura e investigaciones actuales sobre aproximaciones que ayuden a alcanzar la construcción de un chatbot en el estado de dominio abierto, existe una serie de estudios pioneros para la conclusión en un sistema capaz de dar respuestas sintácticamente correctas a un comentario o pregunta, se trata del trabajo inicial de Cho et al. [2014] con el objetivo principal de integrar una red neuronal recurrente en un sistema de traducción automática estadística a través de una arquitectura codificador-decodificador. Basados en el trabajo anterior y otros, un equipo de trabajo de Google presentó en Sutskever, Vinyals, and Le [2014] una arquitectura similar pero desarrollada por una red neuronal recurrente basadas en el método LSTM (Long Short Term Memory) multicapas, que permite plasmar las palabras en un espacio temporal donde actúan las diferentes unidades LSTM en su capacidad conjunta de mantener una memoria de la oración y generar una respuesta cada vez más próxima a la

esperada, desarrollada y publicada como código abierto dentro de las librerías de Tensor Flow.

Aunque la aplicación o fin principal del anterior algoritmo fue la traducción entre diferentes lenguas, también se ha estado usando para el tema que nos concierne, esto es soluciones *post-reply* o mensaje-respuesta, como se puede ver en propuestas de sistemas de conversación a gran escala (Vinyals and Le 2015), o máquinas de respuesta neural (Shang, Lu, and Li 2015) y muchos otros. Sin embargo, el gran problema de estos sistemas de respuestas sintácticamente correctas radica en dos perspectivas, en primer lugar, como modelo de aprendizaje automático; la dependencia de los datos se vuelve crucial, y en caso de que los datos sean pocos y/o cerrados a un dominio específico, sus respuestas a una pregunta que desconoce o para la cual no fue entrenado son incoherentes. Y en segundo lugar, las respuestas que genera no son empáticas con el usuario o no son capaces de generar una fluidez en la conversación.

Como aproximación para resolver el problema de la generación de respuestas empáticas, la máquina de generación de conversación emocional con memoria interna y externa en Zhou, Huang, Zhang, Zhu, and Liu [2017] propone una arquitectura basada en el modelo sequence-to-sequence (seq2seq) de Sutskever, Vinyals, and Le [2014] que puede generar respuesta apropiadas no solo en el contenido (Sintácticas), sino también en la emoción (emocionalmente consistentes). Sin embargo, ningún trabajo ha podido resolver el inconveniente de las respuestas coherentes bajo un contexto de incertidumbre o cercano a un dominio abierto y se espera que el campo de inteligencia artificial general pueda resolverlo en el futuro. Para el propósito de éste trabajo, el agente de conversación debería generar respuestas coherentes a cualquier pregunta, incluso cuando desconoce el tema y además, tener conocimiento para encaminar

al sospechoso a mantener una conversación de la cual se pueda lograr obtener información relevante que pueda ser usada en algún proceso legal. Para dar solución a lo anterior, en este libro de proyecto de grado se propone el uso de un modelo nativo de los chatbots - un modelo basado en reglas desarrollado en el habitual Lenguaje de Marcado de Inteligencia Artificial (AIML) - desarrollado bajo el proyecto ALICE en Wallace, Richard S. [2009] e inspirado en el programa ELIZA en Weizenbaum, Joseph [1976].

El chatbot desarrollado en este proyecto estará enfocado a personas con la intención de distribuir o adquirir contenido ilegal de tipo pornográfico, más en específico, pornografía infantil. Uno de los principales hechos que han motivado la investigación en este campo, ha sido la alarmante incorporación de adolescentes y niños en sitios relacionados a sexting y sitios con temáticas sexuales Karaian, L. (2014). Por medio de este tipo de sitios, diferentes transacciones de contenido pornográfico son llevadas a cabo, tanto legales como ilegales. Considerando que una de las formas de comunicación más empleadas en estos sitios son los chats, también se han hecho investigaciones referentes al riesgo que supone el uso indebido de las redes sociales Ybarra, M. L., & Mitchell, K. J. (2008), en donde se estudia el hecho de que depredadores sexuales se mantienen al acecho en busca de posibles víctimas, procurando adicionalmente intercambiar contenidos de diversos tipos como los que estamos planteando identificar. De esta forma, nuestro enfoque durante el trabajo será la construcción de un agente con las características mencionadas, que sea capaz de interactuar con un individuo, buscando un intercambio de contenido, el cual posteriormente, tendría que ser revisado para validar si se trata de contenido ilegal.

3. Hacia la implementación

Basados en el estado del arte, inicialmente se exploró la línea de investigación basada en Machine Learning bajo un modelo generativo de análisis de lenguaje natural textual (LSTM), sin embargo dadas las limitaciones del modelo se optó por proponer un modelo compuesto integrando un modelo generativo y un modelo basado en conocimiento (Reglas). De esta forma, basados en una inteligencia sentimental interactiva, la cual definimos nosotros como el análisis de los resultados basados en un componente sentimental, de esta manera realizamos el perfilmento del sospechoso.

3.1 Líneas de implementación

El procesamiento del lenguaje natural, actualmente ha sido un campo bastante investigado, en el cual se han propuesto diferentes modelos con el fin de resolver el problema de lograr comprender la forma en la cual se realiza la interacción social y brindar así una respuesta a una conversación. Bajo este contexto, el problema se ha enmarcado en 2 tendencias, conversaciones de dominio abierto y cerrado, las cuales se tratan de una conversación general de diversos temas y una conversación sobre un tema específico, respectivamente. Estas tendencias surgieron durante la investigación de los modelos que presentaban una posible solución. Los modelos basados en conocimiento (Retrieval based) presentaron una limitación debido a que la base de conocimiento puede ser limitada, por lo que la interacción del agente conversacional también se vería limitada, sin embargo, se logró con este modelo basado en conocimiento un comportamiento realmente bueno bajo un contexto de dominio cerrado.

Se incursionó también en una solución basada en Machine Learning, en el cual un modelo generativo usa una base de conocimiento para entrenarse, en este caso es más dinámico que el basado en conocimiento, permitiendo una interacción un poco más general, sin embargo no se consigue completamente el objetivo de una solución de dominio abierto.

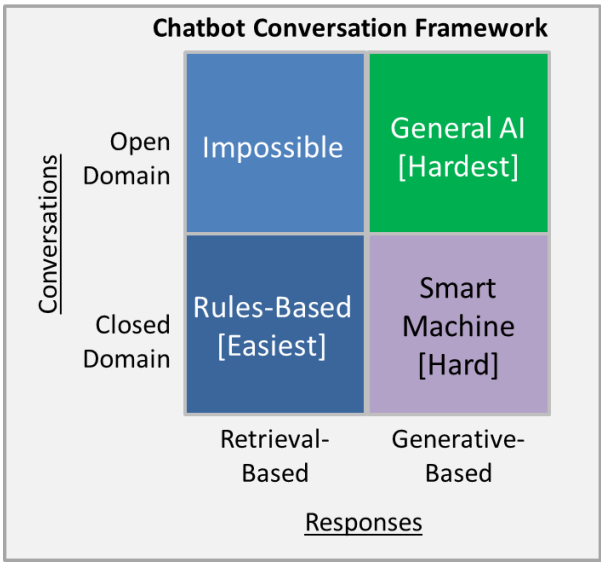

Figura 1. Framework de Teorías en el Campo

[https://chatbotslife.com/ultimate-guide-to-leveraging-nlp-machine-learning-for-you-chatbot-531ff2dd870c]

3.2 Modelo Generativo LSTM

Siendo uno de los modelos más prometedores en el campo, este modelo de Machine Learning se basa en generar su propia respuesta a partir de un entrenamiento previo. Principalmente, bajo esta estrategia se basa el modelo Seq-to-Seq. Se basa en redes neuronales recurrentes, estas permiten analizar los patrones en la oración y mantener un recuerdo de las palabras. Al momento de el usuario ingresar la oración, el modelo procesa y almacena la oración, y responde de acuerdo a la memoria almacenada.

Se escogió el modelo Long Short-Term Memory debido a su capacidad de considerar la memoria de lo que se dice individualmente, es decir, el modelo

puede responder de acuerdo a lo que se dijo por el emisor del mensaje de forma adecuada, cosa que no ocurría con los modelos iniciales como vanilla RNN.

3.2.1 Arquitectura del LSTM

La arquitectura del modelo está basada en la de una red neuronal recurrente. Previamente se realiza una abstracción de las palabras (Word Embedding) hacia un contexto vectorial con el fin de que la red pueda entender las palabras y pueda generar una respuesta. De esa forma, dado un estímulo (Post) nuestro modelo logra generar una respuesta (Reply).

Debido a la capacidad del modelo de tener en cuenta su memoria, su arquitectura interna es más compleja que una vanilla RNN básica, en cuanto al procesamiento y retención de información, esto se puede ver en la figura 2

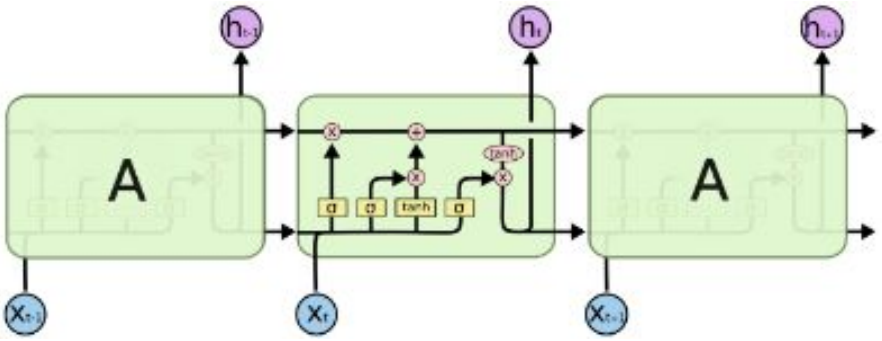

Figura 2. Arquitectura interna de un LSTM [https://colah.github.io/posts/2015-08-Understanding-LSTMs/]

Se puede ver que bajo esta arquitectura hay 3 partes importantes: adicionar información a la memoria (de esta forma aprende), olvidar información (de esta forma se hace más dinámico) y la evaluación de la información (para construir la respuesta).

Adicionalmente, la red tiene una arquitectura externa como se puede ver en la figura 3:

- **Encoder:** Toma el Post y lo analiza completamente, de forma que con la memoria previa obtenida por el entrenamiento con la base de conocimiento, logre identificar los patrones adecuados de respuesta.
- **Decoder:** Toma los patrones que identificó el Encoder y genera el Reply adecuado palabra por palabra como se puede ver en la figura 3.
- **Optimización:** El modelo hace uso del algoritmo de optimización Adam con el fin de obtener la mejor aproximación de respuesta.

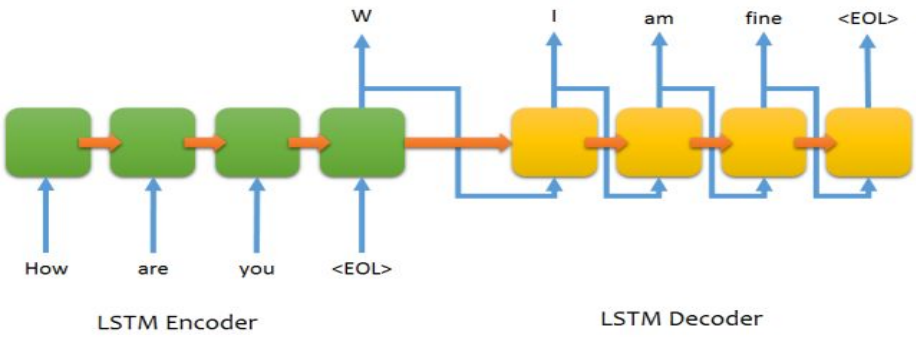

Figura 3. Arquitectura externa de un SeqtoSeq (LSTM)

[https://www.analyticsvidhya.com/blog/2018/03/essentials-of-deep-learning-sequence-to-sequence-modelling-with-attention-part-i/]

3.2.1.1 Algortimo de Optimización

El modelo LSTM necesita de un algoritmo de optimización, que en este caso será Adam, el cual busca en el mejor de los casos un mínimo global o en el peor uno local (Como se puede ver en la Figura 4, derecha), con lo cual logra aprender patrones complejos sobre la base de conocimiento provista. Este algoritmo es basado en un gradiente descendente estocástico con la particularidad de que cada peso en la red tiene un Learning Rate distinto, de esta forma y de acuerdo con la Figura 4 izquierda, converge mucho más rápido que un gradiente normal,

sin embargo tiene la tendencia a tener Overfitting por lo que es necesario tener en cuenta la regularización del modelo.

Como se puede ver en la Figura 4 derecha, el algoritmo lo que hace es bajar por el espacio vectorial modificando los pesos de la red, que este caso sería la memoria de la red neuronal en donde se almacenarán los patrones identificados.

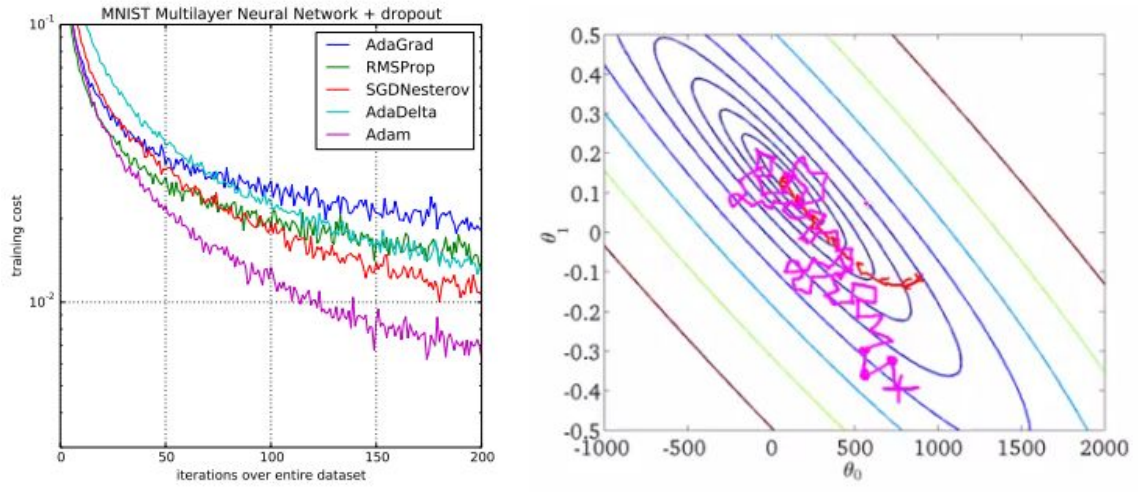

Figura 4. Análisis del algoritmo de optimización ADAM [A Method for Stochastic Optimization, 2015]

3.3 TensorBoard

La implementación del modelo LSTM se realizó sobre python basados en el conjunto de librerías para Machine Learning de Google. Con el fin de conocer el flujo de datos que se seguía en el modelo, se realizó el análisis sobre Tensorboard, sobre el cual se logró identificar cada aspecto del modelo en específico.

3.3.1 Conceptos Iniciales

¿Qué es un tensor?

En TensorFlow, un tensor es la unidad de datos central. Consiste en una generalización de vectores, matrices o estructuras n-dimensionales. Estos tienen 2 propiedades principales, el tipo de dato y la dimensión que se está manejando. Luego de definido el tensor, este servirá de insumo para realizar las computaciones correspondientes al modelo, ya que bajo esta representación vamos a manejar los datos.

- Tipo de dato aceptado: float32, int32, string.
- Dimensiones: Puede ser n-dimensional.

Adicionalmente TensorFlow contiene 2 tipos de datos adicionales, Variables y Constantes, en las cuales hay una similitud con las de un lenguaje de programación convencional.

Grafos y Sesiones

TensorFlow maneja un sistema de grafos y dependencias para mapear todo el flujo de datos de manera que puede hacer cada computación individualmente, y luego pasar los resultados al siguiente elemento en el flujo de datos. De esta forma es mucho más sencillo realizar la computación paralela.

Luego de definir el flujo de datos, y de haber definido las variables, tensores, constantes y operaciones que se van a necesitar, se crea una sesión que corre partes del grafo.

3.3.2 Requerimientos

Para la instalación se necesita Python3, Pip 3, para el caso de linux también es recomendable usar un entorno virtual bajo virtualenv, y las siguientes librerías:

- NLTK: Natural Language Toolkit, es un kit de herramientas para el procesamiento de Lenguaje Natural y el uso de estadísticos,

adicionalmente, cuenta con algunos ejemplos útiles para la realización de visualización de datos.

- Numpy: Es una librería basada en la realización de operaciones complejas sobre vectores y matrices.
- Scipy: Un conjunto de librerías matemáticas bastante completo, incluyendo algoritmos desde álgebra lineal hasta interpolación y procesamiento de señales.

3.3.3 Instalación

La instalación se realizó en Windows y en Linux, sin embargo por facilidad de implementación del entorno de desarrollo se optó por trabajar en Windows.

Instalación Tensorflow - Windows:

- Se corre el siguiente comando en consola para instalar la última versión de Tensorflow, en caso de necesitar otra se agrega la versión como parámetro del comando: `pip install --upgrade tensorflow`
- El comando anterior nos instalará la versión normal, para la versión con GPU es necesario contar con una tarjeta gráfica y con CUDA y se instala con el siguiente comando: `pip install --upgrade tensorflow-gpu`

3.3.4 Flujo de datos y Variables

La instalación de Tensorboard no requiere de librerías adicionales a Tensorflow. Para realizar su integración es necesario tener en cuenta qué métricas se van a monitorear.

- Se definen las variables que se van a monitorear: En este caso estamos analizando la salida de la red al momento de pasar por la función de evaluación *Softmax* y se integra a las variables a monitorear en un histograma: `tf.summary.histogram("softmax_function", y)`

```
net, net_rnn = model(encode_seqs2, decode_seqs2, is_train=False, reuse=True)
y = tf.nn.softmax(net.outputs)
summary1 = tf.summary.histogram("softmax_function", y)
```

- Se guarda el modelo para poder desplegarlo en el servidor: Adicionalmente, luego de inicializar los diagramas, los histogramas se agregan al monitor (merged) que va a estar revisando constantemente los cambios en la variable.

```
# Init Session
sess = tf.Session(config=tf.ConfigProto(allow_soft_placement=True, log_device_placement=False))
train_writer = tf.summary.FileWriter( './logs/1/train.txt', sess.graph)
sess.run(tf.global_variables_initializer())

merged = tf.summary.merge([summary2])
```

- Se agrega el monitor a la sesión: Por último, se agrega el monitor a la sesión en la cual se va a realizar tanto el entrenamiento como el test.

```
    summary, _, err = sess.run([merged, train_op, loss], {encode_seqs1: X, decode_seqs1: _decode_seqs,
                    target_seqs: _target_seqs,
                    target_mask: _target_mask})
```

3.3.5 Despliegue y Análisis

Para el despliegue del servidor se necesita que el modelo esté corriendo, de tal forma que ya haya generado los archivos necesarios para visualizar el flujo.

- Corremos el servidor con: `tensorboard --logdir logs/1`
- Accedemos por medio de la dirección http://localhost:6006/

El Dashboard que nos muestra es el siguiente:

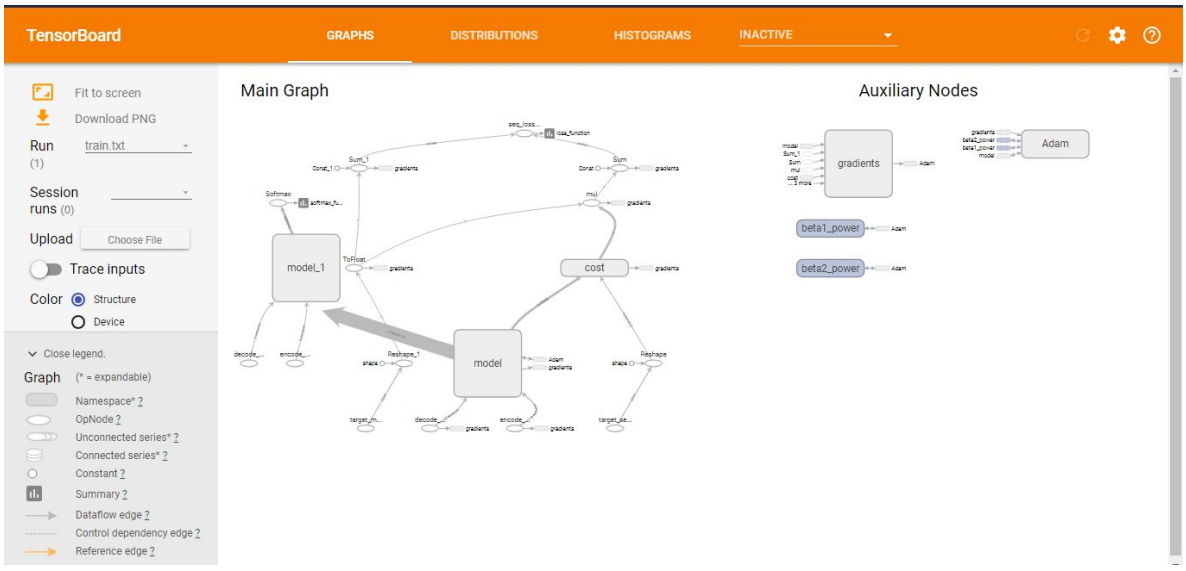

Figura 5. Dashboard inicial con el flujo de datos.

En el dashboard se puede ver el flujo de datos que va entre cada nodo del grafo en los cuales hay ciertas funciones que son usadas por el Sequence-to-Sequence durante el proceso.

Adicionalmente tenemos en la Figura 6 la representación de los modelos usados en el desarrollo (Cuadros en rojo), por una parte, tenemos un modelo que usamos para realizar el entrenamiento, el cual como se mencionó anteriormente, tiene un Encoder, un Decoder y un algoritmo de optimización, adicionalmente el flujo de datos se ve representado por cada flecha que indica un flujo de tensores en el diagrama.

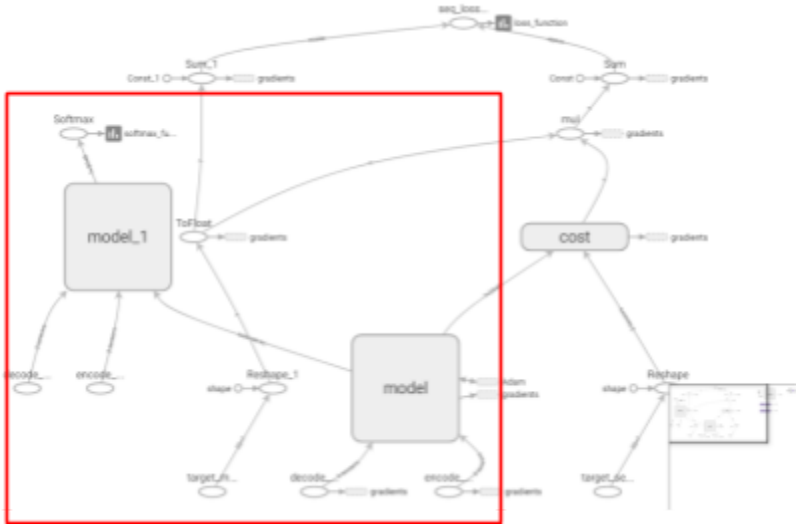

Figura 6. Representación de los modelos usados en el algoritmo

El modelo Sequence-to-Sequence se compone internamente de 2 secciones principales (Figura 7): el preprocesamiento y el entrenamiento. Durante la etapa de preprocesamiento se realiza el word-embedding, en donde se busca pasar de una representación textual a vectorial. Luego pasamos al entrenamiento y se realiza la optimización por medio del algoritmo de Adam el cual calcula sus gradientes para moverse vectorialmente.

Figura 7. Representación interna del SeqtoSeq

Adicionalmente, se realizó una aproximación de visualización basado en el dataset PAPAYA, para el Word Embedding con el fin de comprender más su concepto, esto se puede realizar de la misma forma que realizamos el monitoreo de variables.

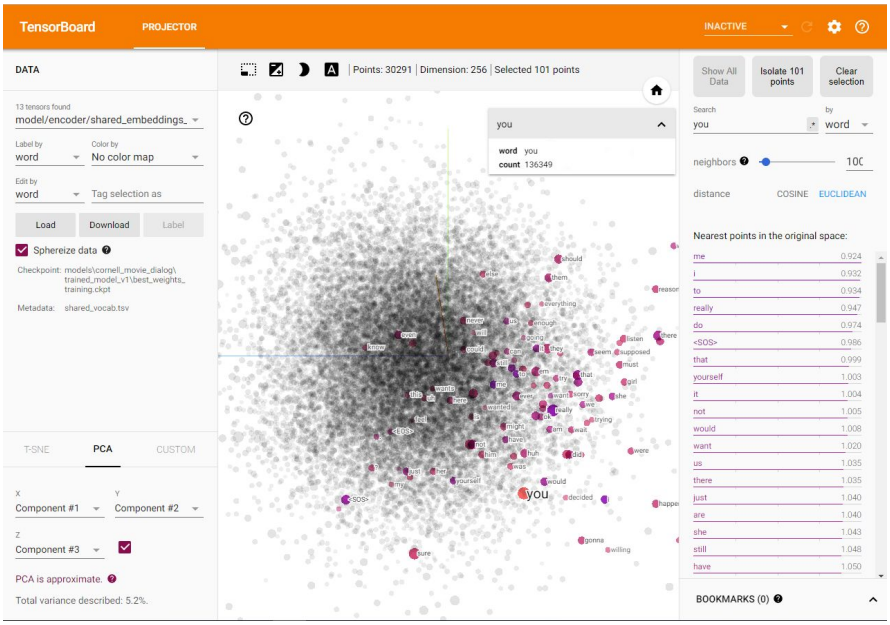

Figura 8. Word Embedding de características

3.4 Modelo Basado en Conocimiento (AIML)

Durante el desarrollo del proyecto se planteó la creación de un modelo basado en conocimiento con el fin de integrar conocimiento más específico que nuestro modelo LSTM (explicado en la sección 3.2) no lograra cubrir. En este caso el modelo fue construido en 3 fases tomando como base ALICE [*https://github.com/datenhahn/python-aiml-chatbot*], un modelo de chat construido para brindar respuestas lo más adecuadas a un contexto. En la primera fase el modelo ALICE fue depurado con el fin de tomar las reglas más relevantes para

nuestro caso y donde identificamos que podrían haber puntos de mejora respecto de las respuestas generadas por el modelo generativo. La segunda fase, consistió en formalizar el conocimiento específico de los chats online tratados, en nuestro caso OMEGLE, de esta forma, le brindamos a nuestro bot la capacidad de responder adecuadamente en el entorno y contexto de las preguntas asociadas típicas de la plataforma. Como última fase de la construcción del modelo, planteamos la formalización de las reglas relacionadas a delitos sexuales y específicamente pornografía infantil, para lo cual se tomaron preguntas y respuestas bajo un contexto de delitos y tráfico de contenido ilegal. De esta forma, al integrar las tres fases se tiene la primer aproximación de nuestra base de conocimiento respecto a varios factores como la interacción efectiva y la obtención de contenido usable (conversaciones y material transferido) para la identificación de un delito sexual. Las reglas del modelo basado en conocimiento se pueden ver como sigue:

```
<category>
    <pattern>* WHERE * FROM *</pattern>
    <template>im from london, so wyd?</template>
</category>

<category>
    <pattern>* BEING *</pattern>
    <template><star index = '2'/>?interesting haha, im just seeing youtube videos</template>
</category>
```

Los anteriores son algunos ejemplos de las reglas usadas en la última fase de la construcción. El modelo construido es apenas una primera aproximación de una base de conocimiento, bajo una arquitectura compuesta internamente por las 3 fases anteriores, sin embargo, con el fin de lograr una mejor aproximación, lo más adecuado es complementar la tercer fase con reglas extraídas del conocimiento de expertos en este tipo de delitos sexuales por medio de ingeniería de conocimiento.

4. Propuesta de solución

Debido a que la construcción de nuestro modelo parte de la incorporación de metodologías de *Data Science*, se siguió el ciclo de vida propuesto por Microsoft, en el cual como primera aproximación de realiza un entendimiento del negocio, que en nuestro caso sería el problema. Posteriormente se obtienen los datos a tratar, en donde se realiza un entendimiento inicial, luego se realiza el modelado, en donde se tienen que tener en cuenta el pre procesamiento de los datos. Finalmente, se realiza la implementación y despliegue del modelo planteado, como se ilustra en la Figura 9.

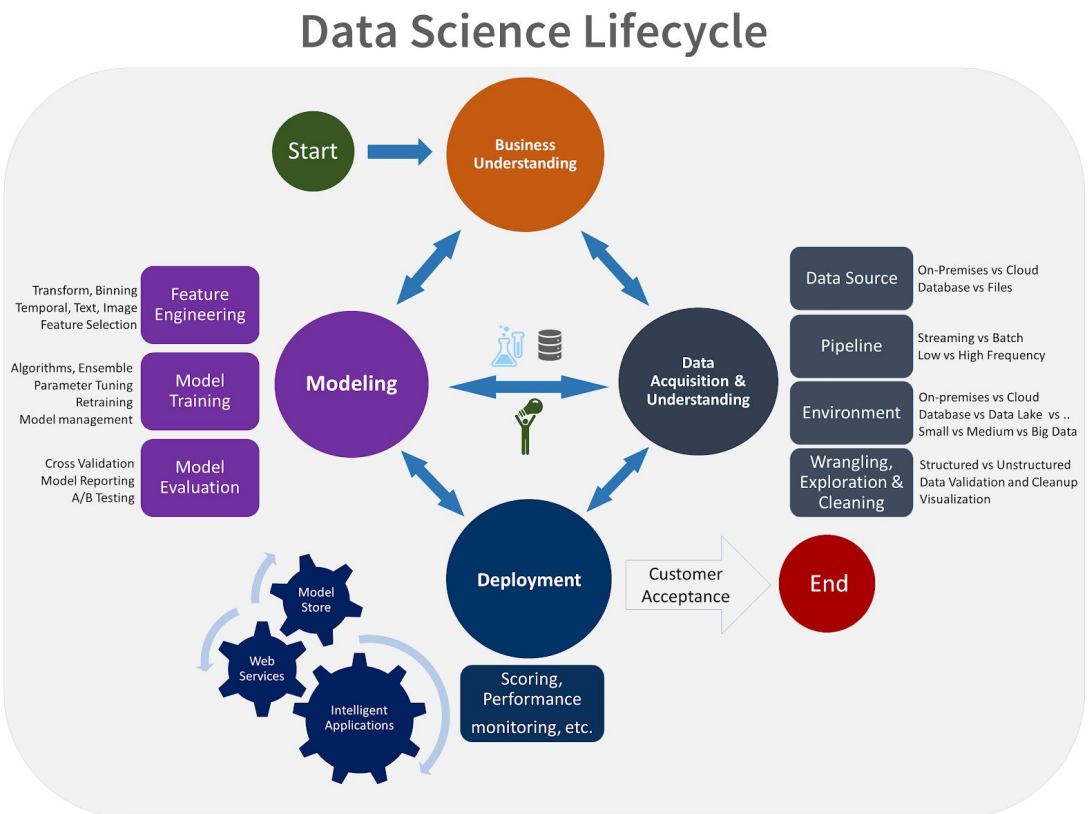

Figura 9. Ciclo de vida de un desarrollo bajo Data Science [TDSP ciencia de datos]

4.1 Entendimiento del problema

Durante el desarrollo de este proyecto se buscó encontrar la forma de realizar una interacción con posibles individuos distribuidores de contenido ilegal, con el fin de perfilarlos de forma que se pudiese alertar a las autoridades respectivas sobre alguna anomalía. Según la investigación realizada, la distribución de contenido ilegal se realizan por medio de un entorno conocido como por ejemplo las redes sociales. Este tipo de comunicaciones se realizan de una forma muy rápida en plataformas como OMEGLE. Esta plataforma, nos permite realizar interacciones con individuos de forma rápida sin perder de vista nuestro objetivo. Adicionalmente, la población con la que se interactúa, se sitúa bajo la premisa de un tema/tópico de interés en común, el cual en nuestro caso será “Sex”. De esta forma, nos aseguramos que nuestro chatbot va a interactuar con un individuo interesado en esta temática, y durante la interacción le daremos un contexto de pornografía infantil.

Durante la construcción de la base de conocimiento, se realizaron pruebas manuales durante más de una semana de forma intermitente, con el fin de identificar la forma como se realiza distribución de contenido y además lograr identificar los puntos en la conversación que son críticos para lograr que el intercambio de realice. Entre los puntos más críticos de la conversación se encuentran:

- Realizar el diálogo de forma directa: Bajo este enfoque, logramos identificar que en una plataforma de chat online como Omegle, una conversación de contexto sexual se beneficia de peticiones directas y empáticas.

- Es más sencillo establecer una comunicación si el sospechoso cree que habla con una mujer: En este caso, la mayoría de las charlas fluyeron cuando el sospechoso creía que estaba hablando con una mujer.
- Finalmente, los individuos de prueba fueron desconfiados al momento de realizar el intercambio de contenido pornográfico en una plataforma en la cual no se conocía nada de la identidad de la otra persona, por lo cual durante la conversación se solicitaba el cambio a telegram.

Telegram es una plataforma para la cual los intercambios pueden mantenerse cifrados entre los dos participantes de la conversación. Adicionalmente, presenta una facilidad de uso respecto a otras plataformas, en el sentido que permite la integración automatización con nuestro bot.

De esta manera, integramos los componentes necesarios para entender el contexto de nuestro problema, la distribución de contenido ilegal como primer factor, el medio de distribución y la manera en la cual se interactúa para lograr un intercambio. Con estos componentes, podríamos decir que tenemos un panorama más claro de nuestro problema.

4.2 Adquisición de datos

Para la etapa de adquisición se buscaron varios datasets en los cuales se mantuvieran una conversación con posibles criminales que distribuyeran contenido ponográfico ilegal, sin embargo, debido a la dificultad por encontrarlo, se optó por un dataset de interacciones generales con la salvedad de que el conocimiento especificó estaría en la base de conocimiento del modelo basado en conocimiento. A pesar de no lograr encontrar este dataset específico, se encontraron aproximaciones de grooming, las cuales se estudiaron y como trabajo futuro se contemplaran para una integración con nuestro modelo. Adicionalmente, el dataset usado para entrenar el LSTM fue PAPAYA

[https://github.com/bshao001/ChatLearner], el cual contiene datos de diferentes fuentes, como noticias, conversaciones e interacciones en foros.

Por otra parte, los datos obtenidos para realizar la construcción de la Base de Conocimiento, fueron tomados a partir de los experimentos manuales de interacción realizados con posibles sospechosos, para de esta forma, crear reglas AIML que permitan acelerar el intercambio de contenido. De igual manera, de acuerdo al estado del arte, se incorporó parte del modelo de ALICE [https://github.com/datenhahn/python-aiml-chatbot], con el fin de que la interacción fuera lo menos plana posible. Se espera que el modelo de conocimiento, en un futuro se integre con conocimiento de agentes especializados en esta área de los ciberdelitos. Adicionalmente, es necesario integrar conocimiento psicológico para mejorar su respuesta, de forma que logremos conducir la conversación con el sospechoso.

4.3 Modelamiento

Nuestra propuesta de solución parte del punto de vista de obtener una interacción adecuada con el sospechoso, por este motivo se emplea una de las configuraciones más generales y frecuentes para chatbots (ver Figura 10), donde se define un canal de interacción (con posibles plataformas como chats de mensajería instantánea, e-mail, páginas web, aplicaciones móviles, entre otros), y se desarrolla la lógica del chatbot teniendo en cuenta que se debe hacer uso de estrategias de procesamiento, entendimiento o generación del lenguaje natural, y una vez se tengan estos analizadores y filtros, se generan modelos que permitan dar respuesta a lo que escribe la gente (ya sean modelos basados en conocimiento o modelos generativos con aproximaciones de Machine Learning) teniendo siempre en cuenta la lógica y las políticas del negocio.

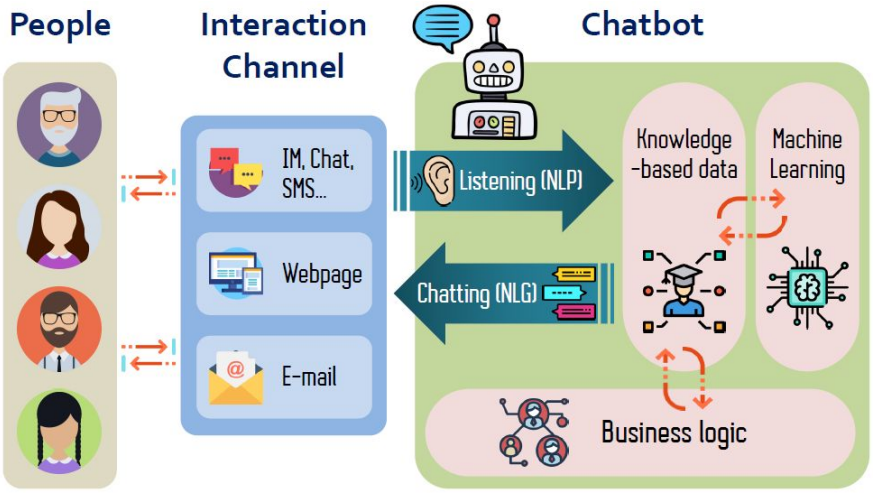

Figura 10. Anatomía de un chatbot

Como se especificó en el marco teórico, nuestra propuesta está enfocada en hacer uso de las dos aproximaciones de solución: un modelo generativo LSTM y un modelo basado en conocimiento AIML. La idea de hacer uso de estos modelos integrados es alimentar al bot con una visión de generalidades en las conversaciones y que también contenga conocimiento sobre el contexto de construcción de una relación de amistad, formación de relaciones de confianza y ofrecimiento y negociación de contenido sexual infantil, a partir de cómo se expresan los usuarios y posibles sospechosos en la red. Sin embargo, para que el chatbot alcance tales estados y logre su objetivo general de perseguir y detectar pervertidos al interactuar con ellos en ciertas salas de chat, la entidad de Conversación Artificial (ACE) correspondiente debe cumplir con los siguientes atributos:

1. **Búsqueda de personas con contenido ilegal:** el chatbot debe exhibir comportamiento de una persona interesada en adquirir pornografía infantil, para identificar sospechosos que poseen contenido ilegal (como imágenes o videos) y están dispuestos a compartirlo con otros.

2. **Búsqueda de licitadores de contenido ilegal:** nuestro chatbot también debe exhibir el comportamiento de un humano interesado en distribuir pornografía infantil, para identificar sospechosos deseosos de obtener y consumir este tipo de contenido ilegal.
3. **Adecuación:** el chatbot también debe poder manejar situaciones en las que la conversación evoluciona hacia temas fuera del contexto principal para el que está destinado el chatbot, es decir, la pornografía infantil.
4. **Perfilamiento sospechoso:** Nuestra solución debe realizar un análisis de la conversación mantenida entre el chatbot y el sospechoso con el fin de perfilar a este último y asignarle alguna categoría dentro de las contempladas en este proyecto.

Nuestra propuesta se compone de 4 módulos principales; el modelo basado en conocimiento, el generativo y los 2 modelos que permiten el perfilamiento (un clasificador de emociones y otro de opinión). El flujo funcional del trabajo de la propuesta se puede ver en la Figura 11. Las siguientes son las descripciones de cada módulo.

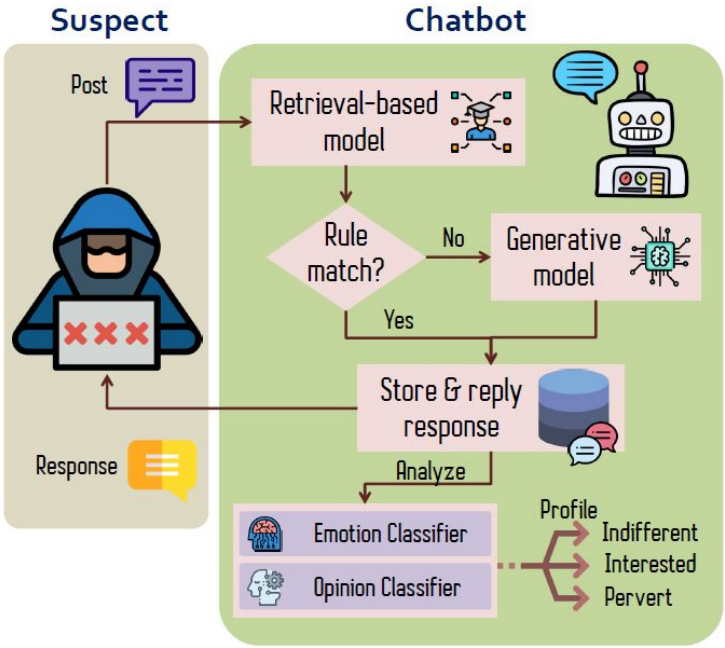

Figura 11. Descripción General del flujo y procesos del chatbot

1. **Modelo de recuperación basado en conocimiento:** Como se mencionó en el marco teórico, es un modelo en el que el conocimiento experto se representa, se guarda y se deduce de un conjunto de reglas expresadas mediante un lenguaje de marca de inteligencia artificial (AIML). Este modelo es el primero que se aplica cada vez que llega un mensaje del interlocutor del chatbot. En caso de que dicha frase coincida con algunas de las reglas del modelo de recuperación, se proporciona la respuesta correspondiente. Sin embargo, en caso de que la publicación no coincida con ninguna regla de las que existen en la base de conocimiento, entonces el modelo de respuesta basado en generativo entra en operación y genera una respuesta. Tales reglas del modelo se esquematizan de la siguiente manera:

```
<aiml version="1.0">
  <category>
    <pattern>* DO * WANNA * WITH ME </pattern>
    <template>
      actually im looking for some hot content, do you have?
    </template>
  </category>
 ...
</aiml>
```

2. **Modelo generativo:** Es un modelo entrenado por medio de una red neuronal (NN) de memoria a largo plazo (LSTM). Las respuestas de un modelo generativo pueden ser más naturales que las de un modelo basado en recuperación (AIML), pero pueden contener fallas gramaticales o, a veces, no tener sentido. Pueden ser observados algunos ejemplos en la Tabla 1.

Enunciado	Respuesta
I need some good content	i will make this one more!
I want weird porn material	i dont think so. Weird?
What are you doing now?	i am chatting with you
I want to offer you something its weird and illegal	Is it weird that i want to be the case?

Tabla 1. Ejemplo de respuestas del modelo generativo

3. **Modelo de clasificaciones de emociones:** Para evaluar las emociones que se encuentran dentro de una conversación dada, este modelo utiliza una Máquina de soporte vectorial (SVM), entrenada con el conjunto de datos de Evaluación semántica (SemEval) 2007 (Strapparava, C., & Mihalcea, R., 2007), como paradigma de clasificación y aprendizaje supervisado. El algoritmo SVM establece los hiperplanos óptimos en un espacio multidimensional que separa claramente las seis emociones etiquetadas en el conjunto de datos SemEval. De esta manera, las respuestas sospechosas son tratadas como nuevas observaciones para el modelo, donde cada respuesta se clasifica con una de las seis emociones SemEval: enojo, disgusto, miedo, alegría, tristeza y sorpresa. La implementación del modelo se realizó en un proyecto pasado, para el cual se obtuvo aproximadamente un 60% de accuracy.

4. **Modelo de clasificación de opinión:** Con el fin de distinguir si un frase dada genera una opinión inclinada hacia (o en contra) un tema específico, hemos desarrollado 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.

4.4 Despliegue del Modelo

Nuestro proyecto busca la integración del chatbot en un contexto real basados en la interacción humano-máquina teniendo en cuenta el el conjunto de modelos mencionados en la sección 4.3. Principalmente, hemos integrado la implementación a plataformas de chat “anónimos” con el fin de conocer los posibles usuarios del sistema. Para llevar a cabo dicho objetivo, aprovechamos las facilidades de Omegle, plataforma Online en la cual se puede interactuar con personas aleatoriamente sin entregar datos personales o que permitan la identificación inicial de los interlocutores.

Hemos de aclarar que la implementación realizada es apenas un prototipo de solución al problema, y las investigaciones actuales en Procesamiento de Lenguaje Natural y Bases de Conocimiento para la interacción Humano-Máquina, siguen en proceso de maduración. De igual forma, nuestra base de conocimiento, puede ser mejorada con el conocimiento experto y nuestro modelo generativo de redes neuronales, puede mejorarse con técnicas más avanzadas de NLP y word embedding.

La construcción del Software, se dió enfocada al punto en el cual un posible sospechoso lograra intercambiar contenido ilegal, y es con esto en mente que se construye un módulo de interacción e intercambio de plataformas con Omegle y Telegram. Luego de algunos experimentos manuales, notamos un patrón usual en aquellos que estaban dispuestos tanto a recibir como a compartir contenidos: No deseaban hacerlo de forma anónima y deseaban mantener una canal para futuras comunicaciones, por este motivo nuestro despliegue se basó en contactar al sospechoso en Omegle, realizar una interacción para establecer confianza con el sospechoso y posteriormente propiciar un cambio a otra plataforma de comunicación, donde se pudiese realizar el intercambio de contenido ilegal.

De esta forma la implementación realizada se compone del módulo del bot en python, adicionando los módulos de análisis realizados: una parte en R (análisis de emociones) y las métricas programadas en python que son descritas en la sección 4.5.

4.5 Experimentación y Resultados

Con el fin de consolidar las métricas en una que nos permita establecer una caracterización se propone la adecuación de la función sigmoid, utilizada comúnmente en el ámbito de Machine Learning. La función sigmoide integra las siguientes métricas:

- *R*, siendo el número de reglas AIML disparadas en una conversación relacionadas al tema de la base de conocimiento (pornografía infantil), según el modelo de recuperación basado en conocimiento, mencionado en la sección 4.3.

- *E*, el número de emociones identificadas (enojo, disgusto, miedo, alegría, tristeza y sorpresa), promediadas, según el modelo de clasificación de emociones mencionado en la sección 4.3.
- *O*, el número de opiniones positivas promediadas, según el modelo de clasificación de opinión mencionado en la sección 4.3.
- $\boldsymbol{\tau}$, el tiempo que duró la conversación integrado con el tiempo necesario para responder entre mensajes por el sospechoso.

La integración de las métricas se puede ver en la siguiente figura:

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

Siguiente a esto, la caracterización del sospechoso, está dada como se puede ver en la Figura 12, teniendo en cuenta que los deltas ($\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$) son los parámetros de sensibilidad de la función φ. Después de ejecutar el chatbot se obtuvieron (usando $\boldsymbol{\delta}_1$= 0.05 - $\boldsymbol{\delta}_2$= 2.0, 35) 320 chats acumulados en 15 horas, los cuales fueron filtrados (quitando chats inconclusos, generados por bots, etc) para obtener un total de 35 chats aprovechables.

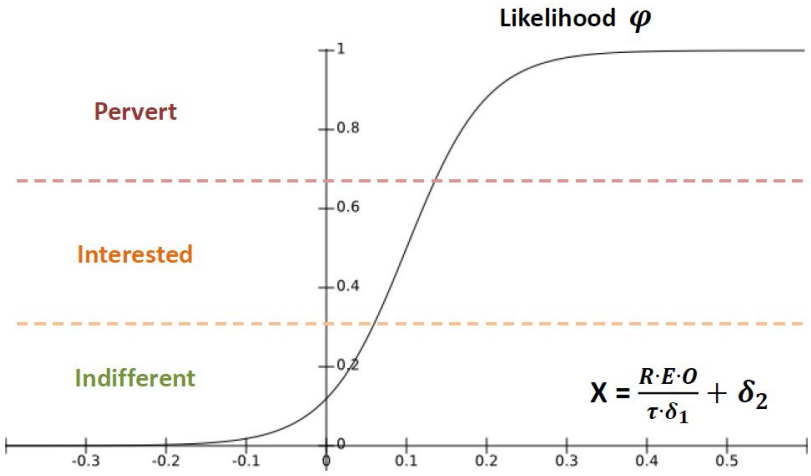

Figura 12. Distribución de la función Sigmoide

Los resultados individuales se pueden resumir en la figura 13, en la cual podemos ver cuál fue el comportamiento de cada métrica que intervino en la clasificación de los sospechosos. De los 35 chats obtenidos, la clasificación realizada arrojó los siguientes resultados: 26 chats fueron clasificados en el perfil "indiferentes", con una mediana de φ≈0.12; 4 chats fueron clasificados en el perfil de interesado con una mediana de φ≈0.41; y finalmente 5 chats fueron clasificados como "pervertidos", teniendo una mediana de φ≈0.99. Es de importancia aclarar que la métrica que estamos usando no es una probabilidad, ya que carece de las formalidades matemáticas para ello, pero si es una buena aproximación a la clasificación necesaria para identificar sospechosos de delitos sexuales.

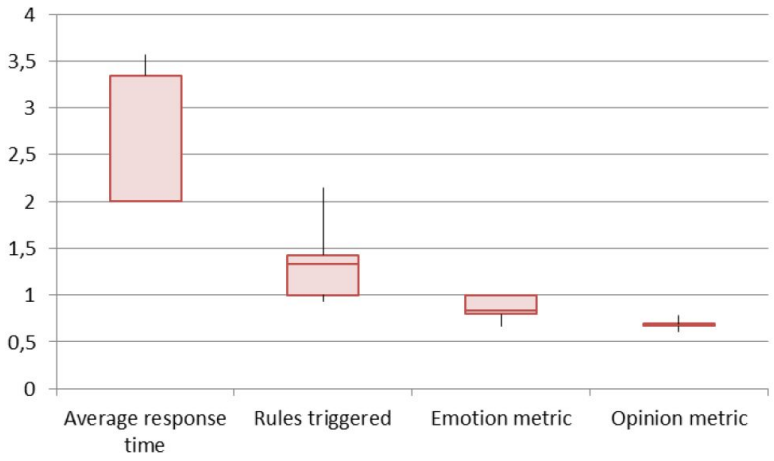

Figura 13. Evaluación de las métricas propuestas

5. Diseminación de resultados

Una versión extendida de los resultados de este proyecto puede encontrarse en el siguiente artículo: *J. Murcia, S. Moreno, D. Díaz Lopez, and F. Gómez Mármol . C3-Sex: a Chatbot to Chase Cyber perverts. The 4th IEEE Cyber Science and Technology Congress. Track Cyber Security, Privacy & Trust (Cyber Crime, Fraud, Abuse & Forensics). Fukuoka, Japan, 2019.*

Conclusiones

El trabajo realizado fue desarrollado integrando diferentes conceptos y herramientas de la Inteligencia Artificial para aportar a la solución de un problema en concreto, y consideramos que tiene un amplio potencial a partir del mejoramiento de cada componente individual.

Por nuestra parte, la mejor aproximación que encontramos fue la integración de los dos modelos propuestos: un modelo generativo (LSTM) y un modelo de conocimiento basado en reglas. En la medida que íbamos encontrando nuevas conversaciones y ajustamos los modelos, especialmente las reglas AIML, la interacción fue mejorando, por lo que proponemos como trabajo futuro estudiar la composición y mejora de estos modelos más a profundidad. Si se mejora la precisión de estos modelos individualmente, se mejorará el desempeño general del chatbot.

Las métricas propuestas fueron buenas, es necesario ajustar la métrica de emociones brindando una mejor riqueza en cuanto a expresividad, pero en general los resultados mostraron que φ podría ser confiable.

De las reglas AIML propuestas solo unas pocas fueron efectivamente utilizadas en la interacción, lo que implica que el comportamiento de un sospechoso es más complejo de lo actualmente modelado. Por lo que, como se comentó al inicio, es necesario contar con un experto en pornografía infantil para extraer el conocimiento específico y plasmarlo en nuestro sistema.

Finalmente, el trabajo realizado tiene mucho por mejorar, sin embargo, confiamos en que la idea podría ser usada como base y ser mejorada por el avance de la

tecnología y la investigación. Este es un campo realmente abstracto en el cual se han logrado cosas bastante interesantes y creemos que es posible modelar y abstraer patrones más complejos a los actualmente encontrados, integrando en alguna medida patrones y sistemas biológicos como en las investigaciones actuales que buscan integrar células a los procesadores. La calidad de nuestro saber, es fundamentada en la calidad de nuestro ser, por esto los avances que han surgido se han dado con pasión y de igual forma que las investigaciones futuras, esperamos que quienes continúen con nuestra investigación encuentren la pasión de la búsqueda del conocimiento.

Referencias

Karaian, L. (2014). Policing ‘sexting’: Responsibilization, respectability and sexual subjectivity in child protection/crime prevention responses to teenagers’ digital sexual expression. Theoretical criminology, 18(3), 282-299.

Ybarra, M. L., & Mitchell, K. J. (2008). How risky are social networking sites? A comparison of places online where youth sexual solicitation and harassment occurs. Pediatrics, 121(2), e350-e357.

Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) (pp. 70-74).

Kotzias, D., Denil, M., De Freitas, N., & Smyth, P. (2015, August). From group to individual labels using deep features. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 597-606). ACM.