ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

MODELAMIENTO DE TÓPICOS APLICADOS AL ESTUDIO DE ATAQUES DE INGENIERÍA SOCIAL

PROYECTO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

LUIS ENRIQUE VELÁSQUEZ GARCÍA
luis.velasquez@epn.edu.ec

DIRECTOR: M.Sc. PATRICIO XAVIER ZAMBRANO RODRÍGUEZ
patricio.zambrano@epn.edu.ec

CODIRECTOR: PhD. JENNY GABRIELA TORRES OLMEDO
jenny.torres@epn.edu.ec

Quito, septiembre 2021

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por **Luis Enrique Velásquez García,** bajo mi supervisión.

M.Sc. Patricio Xavier Zambrano Rodríguez

DIRECTOR DE PROYECTO

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por **Luis Enrique Velásquez García,** bajo mi supervisión.

PhD. Jenny Gabriela Torres Olmedo

CODIRECTOR DE PROYECTO

DECLARACIÓN

Yo, **Luis Enrique Velásquez García** declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normativa institucional vigente.

Luis Enrique Velásquez García

DEDICATORIA

A mis padres Fausto y Marisol quienes han sido el pilar fundamental de mi formación humana y académica, a mis hermanos Diego, Guillermo y Rafael quienes son el mejor regalo que he recibido de la vida. Y a mi inspiración Nicole, quien me ha brindado su apoyo incondicional en todo momento.

Para ustedes, mi familia, quienes nunca dejaron de creer en mí.

Luis Enrique Velásquez García

ÍNDICE DE CONTENIDO

RESUMEN

Hoy en día, el intercambio de información a través de la red es abismal, dando lugar a que exista mayor riesgo de sufrir ataques como acoso, intimidación, chantaje y amenazas en los canales de comunicación que ofrece internet. Los estudios relacionados con la ciberseguridad sobre el uso de técnicas de ingeniería social son limitados. Diversos factores motivarían la disminución de aportaciones significativas a las problemáticas en el campo de la seguridad de la información. Los investigadores de esta área se ven afectados por un acceso limitado a bases de datos de ataques en línea, ausencia de una estandarización que aborde la naturaleza del problema, o la falta de propuestas claras que ayuden a mitigar y prevenir el problema. Este proyecto tiene como finalidad ser un soporte en la investigación realizada por el Ing. Patricio Zambrano, quien propone establecer un procedimiento que denote el ciclo de vida de ataques relacionados a la ingeniería social. Por esta razón, se presenta un análisis detallado junto a su respectiva implementación, de diferentes técnicas de modelamiento de tópicos como NMF (Factorización de Matrices No Negativas), LSI (Indexación Semántica Latente), LDA (Asignación Latente de Dirichlet) y HDP (Proceso Jerárquico Dirichlet) para comprender que modelos obtienen los mejores resultados a la índole de los datos de mensajería instantánea. El proyecto se encuentra bajo el enfoque del marco de trabajo Scrum dando formalidad tanto al levantamiento de requerimientos como al desarrollo de la investigación. Conjuntamente, se utilizó la metodología CRISP-DM para el manejo y tratamiento de los datos.

Palabras claves: ciberseguridad, ciberataques, ingeniería social, modelamiento, tópicos, scrum.

ABSTRACT

Nowadays, the exchange of information through the network is abysmal, which results in a greater risk of suffering attacks such as harassment, intimidation, blackmail and threats in the different communication channels offered by the Internet. Cybersecurity-related studies on the use of social engineering techniques are still limited. Several factors could motivate the decrease of significant contributions to the problems that arise in the field of information security. Researchers in this area are affected by limited access to online attack databases, the absence of standardization that addresses the nature of the problem, or the fact that no clear proposals are put forward to help mitigate and prevent the problem. The purpose of this project is to support the research carried out by Eng. Patricio Zambrano, who proposes to establish a procedure to denote the life cycle of attacks related to social engineering. For this reason, a detailed analysis is presented together with its respective implementation, of different modeling techniques of topics such as NMF (Non-Negative Matrix Factorization), LSI (Latent Semantic Indexing), LDA (Latent Dirichlet Allocation) and HDP (Hierarchical Dirichlet Process) to understand which models obtain the best results in instant messaging data. The project is under the Scrum framework approach to give formality to both the requirements gathering and the research development. In conjunction, the CRISP-DM methodology was used for data management and processing.

Keywords: cybersecurity, cyber attacks, social engineering, modeling, topics, scrum.

CAPÍTULO 1. ENTENDIMIENTO DEL PROBLEMA

1.1. Introducción

El ciberacoso es un concepto que ha estado presente en diversas plataformas de tecnologías digitales, sin embargo, en la actualidad ha tomado más fuerza debido a la gran expansión de sitios web y redes sociales. Los diferentes tipos de ataques como *bullying, grooming,* violencia de género, fraude bancario, entre otros, poseen un patrón: la manipulación psicológica para alcanzar una satisfacción personal o económica. Desde el punto de vista de la seguridad de la información, esta entidad se analiza como parte del campo de la ciberseguridad. El crecimiento de las comunicaciones digitales ha permitido que las personas tengan mayor libertad de expresión. No obstante, en algunas ocasiones el contenido de estas contiene rasgos de acoso de todo tipo, como el odio y el rechazo. Esta carga emocional que está presente en contenidos textuales digitales precisa el apoyo de mecanismos informáticos para su estudio.

La falta de prevención en el inicio de los diferentes ataques ha desencadenado problemas psicológicos en las víctimas. En algunos casos llevando a las personas afectadas a perjudicar su integridad física. En definitiva, es necesario que exista soluciones innovadoras que ayuden a mitigar y combatir este tipo de ataques. Por otro lado, es imprescindible el trabajo colaborativo de los diferentes entornos como familiar, policial y judicial para construir defensas eficaces contra los ciberataques.

La cantidad exponencial de comunicaciones virtuales que se dan en la actualidad es visible. Proponer mecanismos que permitan obtener el conocimiento existente en las comunicaciones se ha convertido en un reto para los investigadores. Con este objetivo, es necesario emplear técnicas de procesamiento de lenguaje natural (PLN). Dentro de PLN existen técnicas conocidas como modelamiento de tópicos o temas. Para ilustrar, se tienen varios algoritmos como LSA (*Latent Semantic Analysis*), LDA (*Latent Dirichet Allocation*), CTM (*Correlated Topic Model*), entre otros. Estos algoritmos determinan relaciones entre las palabras que componen los textos digitales. Ahora bien, es necesario implementar y evaluar cuales de estos modelos son los que arrojan mejores resultados tomando en cuenta la naturaleza de los textos.

Muchos de los ciberataques surgen de conversaciones en chats, blogs, entre otros, denominándose estos como “textos cortos”. El objetivo de este proyecto de titulación es analizar e implementar diferentes tipos de modelamiento de tópicos con la finalidad de soportar la investigación llevada a cabo por el Ing. Patricio Zambrano en su afán de plantear un procedimiento que permita establecer ciclos de vida de los ataques

relacionados a la ingeniería social en función de los patrones conductuales de los atacantes.

Para la formalización de este requerimiento y la descripción de los resultados se tomó en cuenta el conjunto de procedimientos recomendados por Scrum.

El resto del documento se organiza como sigue. El capítulo 2 indica los objetivos planteados para la presente investigación. El capítulo 3 incluye la definición de algunos términos, aspectos teóricos y metodologías requeridas para el desarrollo formal de esta investigación. El capítulo 4 muestra la utilización de la metodología Scrum para el análisis y puesta en marcha de modelos de tópicos necesarios para desarrollo de este trabajo. El capítulo 5 presenta la implementación y posterior comparación de los resultados que arrojan los diferentes modelos utilizados. Por último, se encuentran las conclusiones y anexos de la investigación.

CAPÍTULO 2. PREGUNTA DE INVESTIGACIÓN Y OBJETIVOS DEL PROYECTO

2.1. Pregunta de Investigación

¿Se puede establecer patrones conductuales en ataques que utilizan ingeniería social usando técnicas de modelado de tópicos?

2..2. Objetivo General

- Implementar modelamiento de tópicos para el análisis de ataques asociados a la ingeniería social

2.3. Objetivos Específicos

- Determinar que modelos se van a implementar en función de su relevancia teórica y acceso a implantaciones informáticas.
- Implementar los modelos seleccionados y evaluar sus resultados en relación con la pertinencia de cada agrupación de palabras a las diferentes fases de los ataques descritos.
- Realizar una comparación cuantitativa y cualitativa de cada modelo de tópicos aplicados a los ataques descritos.
- Seleccionar el modelo de tópicos que genera la mayor precisión predictiva de un conjunto de datos.

2.4. Hipótesis

El uso de modelamiento de tópicos permite analizar y establecer patrones conductuales de atacantes que utilizan la ingeniería social en los ataques de Bullying y Grooming.

CAPÍTULO 3. MARCO TEÓRICO

Antecedentes

El activo más valioso que las organizaciones manejan hoy en día es la información. Los ataques cibernéticos se han enfocado en capturar dicha información. Por el contrario, las organizaciones han invertido grandes cantidades económicas para crear mecanismos e infraestructuras robustas que salvaguarde dichos activos. Estas acciones se alinean al ámbito de la seguridad informática mas no de la ciberseguridad. En [1] se expone que la seguridad de la información se centra en preservar a la información siendo esta un activo que puede sufrir ataques, mientras que la ciberseguridad considera la protección del ciberespacio es decir la protección de cada elemento que genera información (activos).

Muchos ciberataques han sido dirigidos al acoso de personas. Así, por ejemplo, el ciberbullying y el grooming han estado en auge en los últimos años, como lo demuestra en [1], [2], [3]. El fin de estos ataques es golpear la estabilidad emocional y psicológica de las víctimas. Utilizan estrategias de intimidación, así como de persuasión para que las victimas queden comprometidas. En casos de mayor gravedad, el daño físico irreversible se hace presente en las victimas. Estos ataques se los lleva a cabo a través de medios tecnológicos. Si bien el ciberacoso no afecta la confidencialidad, la integridad o disponibilidad de la información, su objetivo es dañar la estabilidad emocional y física de las víctimas. Por esta razón las personas deben ser consideradas como parte de la noción de activos en el campo de la ciberseguridad.

Ingeniería Social

En el campo de la seguridad informática se encuentra la Ingeniería Social. Hadnagy en [4] define a la ingeniería social como la acción de manipular a las personas para que realicen actos que no necesariamente están dentro de sus intereses. Por consiguiente, los ciber atacantes obtienen beneficios como contraseñas o información confidencial. La implementación de esta actividad en el ámbito del ciberespacio ha llevado a vincular la psicología con la tecnología tal como se aprecia en [5][6].

Tecnología vs Psicología

La aplicación de la tecnología ha ayudado a que varias áreas de las ciencias sociales generen más conocimiento, entre ellas la psicología. La cual se enfoca tanto en el

entendimiento como en el análisis del comportamiento humano y social. El uso de herramientas tecnológicas en el ámbito de la piscología beneficia a las personas a través de la modernización de técnicas de análisis y evaluación de la conducta humana [7,8]. Naturalmente, el creciente aumento de redes sociales ha permitido varios estudios [9] que relacionan el uso de la tecnología y los efectos psicológicos que causa en las personas. De esta manera se observa, que la psicología incluye a las tecnologías de la información para amplificar los conocimientos en esta rama. Por otro lado, dentro de la informática existe grandes cantidades de información digital que se relaciona con el comportamiento social. Esta información requiere de procedimientos formales que permitan relacionarlos a patrones conductuales en el campo de la psicología.

Subjetividad en los resultados científicos

Existe un componente psicológico que está presente en el acoso en línea. Este componente ha permitido establecer y diferenciar las fases que forman parte del ataque. Sin embargo, al ser un componente subjetivo es complicado estandarizar un procedimiento que permita modelar otros tipos de ataques. En [10] se muestra que no existen suficientes datos, se evidencia la falta de herramientas tecnológicas que ayuden a crear un estándar de los ataques de ingeniería social y las dificultades al momento de cuantificar datos subjetivos. En contraste a lo anterior, en [2,3] presentan la existencia de herramientas que basan sus resultados en algoritmos estadísticos, reduciendo así la subjetividad de los datos. Estas herramientas son conocidas como modelos de tópicos. Permiten obtener información de un conjunto de datos. En otras palabras, agrupan palabras (datos) que posean características similares dentro de un contexto, en este caso, fases de un ataque de ingeniería social.

Modelado de amenazas persistentes avanzadas - APT

Existen diversas propuestas de modelamiento de amenazas persistentes avanzadas (APT) como Mandiant, LogRhythm, Lockheed, entre otras. Sin embargo, cada una tiene sus distintas fases. Como se mencionó, esto debido a la falta de acuerdos para una estandarización. Por otro lado, desde el punto de vista de la ingeniería social se proponen modelos alternos para los ataques APT, como se muestra en [11] Mitnick presenta cuatro fases (recopilación de información, desarrollo de relaciones, explotación de relaciones y ejecución para conseguir el objetivo), además Mounton en [12] plantea como fases (formulación del ataque, recopilación de información, preparación, desarrollo de relaciones, explotación de relaciones y debrief). Teniendo en cuanta esta falta de convenios para modelar ataques relacionados a la ingeniería social, en [2] Zambrano propone una relación entre las fases de los modelos y las características léxicas obtenidas en sus casos de uso, como lo es el *bullying* y el *grooming.*

¿Qué es un modelo?

Un modelo es una construcción conceptual simplificada de una realidad más compleja, es una representación simple que permite plasmar una idea, por ejemplo, las ecuaciones, los gráficos, unos planos o los mapas son modelos del mundo que percibimos. Una cualidad de los modelos es que se aproximan correctamente a la realidad y permiten un entendimiento simple para utilizarlos. Los modelos se construyen a partir de observaciones y evidencias. Pero en el mundo de la informática, los modelos que se utilizan como base para lo que hoy se conoce como *machine learning*, *deep learning* o inteligencia artificial, son los modelos probabilísticos. Los cuales tienen como herramienta principal a la probabilidad.

Existen tres conceptos importantes dentro de los modelos que se deben tener en cuenta. El primero es los datos, son las mediciones que hacemos de la realidad, cabe mencionar que los datos son multidimensionales. El segundo es los parámetros, son los valores que se pueden modificar para ajustar el modelo a los datos. Y el tercero es el error, se necesita una función de error que sirve para conocer si un modelo se ajusta o no a los datos. Los diferentes parámetros del modelo van cambiando según el modelo para optimizarlo. Este proceso también se lo conoce como ajuste de modelo o entrenamiento.

Revisión de aplicaciones de modelado de temas.

El modelado de temas es una herramienta analítica para la evaluación de un conjunto de datos. Forma parte del concepto de inteligencia artificial específicamente del procesamiento de lenguaje natural (PLN) dado que también se trata de una técnica de aprendizaje automático no supervisado. Se enfoca en la detección de patrones tomando en cuenta las relaciones y restricciones dentro de los conjuntos de datos. La coherencia y la perplejidad son los factores que permiten estimar los resultados obtenidos. Existen varios artículos que muestran diferentes aplicaciones de los modelos de temas, revisar Tabla 8.

Tipos de metodologías aplicables.

En primer lugar, se estableció solo la metodología CRIPS-DM para llevar a cabo el presente proyecto el cual conlleva un imprescindible tratamiento de datos, sin embargo, es necesario complementarla con un marco de trabajo que aplique un conjunto de buenas prácticas enfocado en proyectos como Scrum. Por esta razón, para tener un mejor entendimiento de cada una se describen a continuación tanto la metodología CRISP-DM (tratamiento de datos) como Scrum.

Scrum

Es un marco de desarrollo ágil, en los años 90 Jeff Sutherland y Ken Schwaber formalizaron Scrum [13] como un marco de trabajo con reglas enfocadas al desarrollo de productos o proyectos de forma incremental e iterativa. Hoy en día el enfoque de Scrum es utilizado en diferentes áreas, no solo en el desarrollo de Software [14]. Scrum también es conocido por ser un método para la gestión de proyectos que sean iterativos, por tiempo e incrementales, bajo un enfoque de inspección y adaptación [15].

Scrum provee las reglas y practicas necesarias para llevar a cabo el desarrollo de un proyecto. Pero el compromiso de la parte humana es fundamental. No solo se requiere de seguir el marco de trabajo, si no de un cambio de mentalidad en las personas que apliquen este marco de trabajo. Scrum adopta un conjunto de valores y principios que son la base para una buena gestión de proyectos.

El proceso de Scrum consta de varios elementos para su implementación. Se conoce como Sprint a los ciclos de trabajo iterativos. Los Sprint duran alrededor de 4 semanas. Antes de iniciar el Sprint, el equipo de trabajo realiza una recopilación de requerimientos a través de reunión. Esta recopilación de requerimientos por lo general se lo realiza a través de historias de usuario, en las cuales consta las necesidades de los clientes, el equipo y las partes interesadas. Además, en esta reunión el equipo acuerda un objetivo colectivo pequeño, estable y claro que debe ser cumplido al finalizar el Sprint. Durante el Sprint no se pueden añadir nuevos requerimientos. Cada día, durante el Sprint, el equipo se reúne brevemente para discutir, informar y ajustar las tareas que llevaran a cabo cada día para cumplir el objetivo colectivo. Al final del Sprint, el equipo repasa el Sprint con las partes interesadas y mostrarán el producto funcional que desarrollaron en el tiempo que demoró el Sprint. Adicional, reciben comentarios que pueden ser incorporados en el siguiente Sprint [16].

Roles de Scrum

Existen 3 roles definidos que se detallan a continuación:

Product Owner: El propietario del producto es la persona comprometida a identificar las características del producto, priorizando sacar el máximo retorno de la inversión. Es quien da las preferencias a las características que van a ser implementadas por el equipo en cada Sprint. Tomando en cuenta lo anterior es el responsable de actualizar el *Product Backlog* [16].

Team: El equipo está conformado por un grupo de personas especialistas en varias áreas. El equipo es multifuncional y auto organizado, posee un alto grado de

responsabilidad y autonomía. Los integrantes tienen diferentes habilidades de diseño, análisis, desarrollo, arquitecturas, pruebas, entre otros. Dentro del equipo no hay etiquetas, no hay un DBA (administrador de bases de datos), un programados, o un jefe de equipo, simplemente son miembros del equipo. Esto provee al equipo de un ambiente de multi aprendizaje puesto que cada miembro del equipo posee cierta experticia en algún campo, sin embargo, sigue aprendiendo de las otras especialidades de los miembros del equipo. El equipo es el responsable de seleccionar de forma crítica los elementos del *Product Backlog* que implementarán en cada Sprint para lograr el objetivo colectivo. Además, de esto el equipo tiene la facultad de proporcionar ideas al *Product Owner* para aumentar y mejorar la productividad [16].

Scrum Master: Es la persona que ayuda al equipo y al *Product Owner* a entender, aprender y aplicar Scrum. Es el encargado de facilitar las reuniones, garantizando la resolución de posibles conflictos que se generen a lo largo del proyecto. Dado que el equipo es un ente autoorganizado, el *Scrum Master* mejora el flujo de trabajo, la autogestión y empoderamiento de los miembros del equipo [16].

Artefactos Scrum

Product Backlog

Es una lista priorizada de los requisitos del proyecto que se va a desarrollar hasta alcanzar una funcionalidad completa del producto. El *Product Backlog* se encuentra en constate actualización durante la vida del producto. En otras palabras, son todas las actividades que el equipo desarrollará, organizadas en orden de prioridad [16]. El *Product Owner* es el encargado de mantener al día esta lista de requerimientos.

Sprint Backlog

Consta de las tareas seleccionadas por el *team,* las cuales serán completadas durante el Sprint. En el *Sprint Planning Meeting* el equipo elije las tareas con mayor reelevancia y de acuerdo a su capacidad, para implementarlas de manera correcta. En la mayoría de las ocasiones, se asigna un tiempo determinado para el cumplimiento de cada tarea. Las tareas incluyen información sobre el trabajo que se va a realizar [16].

Sprint Burn-down Chart

Es una gráfica que indica la cantidad de trabajo que queda en el futuro para que el equipo termine las tareas del Sprint. Comúnmente es un gráfico que tiende a descender, hasta llegar a cero 0 en el último día del Sprint. Lo importante de este grafico es que muestre el progreso que han tenido hacia el objetivo [16].

Eventos Scrum

Sprint Planning Meeting

La reunión de planificación del Sprint tiene lugar al comienzo de cada Sprint. Aquí, el *Product Owner* indica al equipo, que elementos del *Product Backlog* tienen mayor preferencia sobre los otros. El equipo junto con la ayuda del *Product Owner* estiman la cantidad de tareas que se llevaran a cabo para el próximo Sprint. Luego, es el equipo quien decide las tareas del *Product Backlog* que se desarrollarán en el presente Sprint [16].

Daily Scrum Meeting

La reunión diaria de Scrum tiene una duración de 15 minutos. Los miembros del equipo informan a los otros miembros del equipo sobre los avances que tuvieron, lo que van a desarrollar ese día, y los impedimentos u obstáculos que tuvieron al cumplir cierta tarea. Esto con el fin de mejorar la comunicación y resolver inconvenientes que puedan llegar a retrasar el cumplimiento de las tareas [16].

Sprint Review

Una vez finalizado el Sprint, toma lugar la revisión de *Sprint, Scrum Master, Product Owner*, el equipo las partes interesadas, se reúnen para discutir lo que se ha logrado en el Sprint, es una actividad de revisión y adaptación para el producto [16]. Tiene una duración máxima de 30 minutos.

Sprint Retrospective

Es la actividad siguiente al *Sprint Review*, se enfoca en el proceso y el entorno del Sprint. Se verifica el desempeño del equipo, y se identifica las actividades que resultaron beneficiosas y las que no. Además, aquí se analiza el *BunrDown Chart* para comprobar si el modo en que se llevó a cabo las tareas permitió ejecutarlas de forma eficiente.

CRISP-DM

Cross Industry Standard Process for Data Mining, es un modelo estandarizado ampliamente utilizado en la minería de datos. La norma estableces 6 fases diferentes que pueden llevarse a cabo una o varias veces como se muestra en la Figura 1.

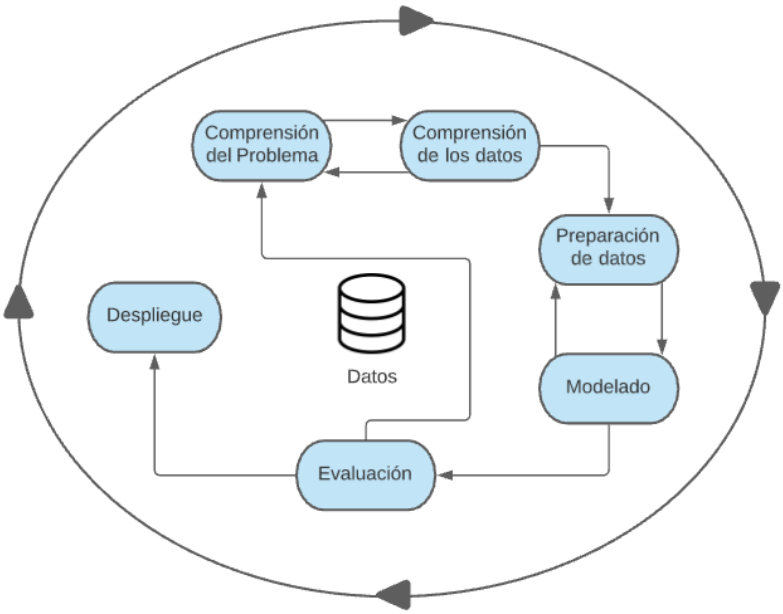


Figura 1. CRISP-DM

Fase de comprensión del problema

Es la fase más importante de la metodología puesto que es aquí donde se entiende de manera clara el problema que se pretende resolver [17]. La guía especifica que se deben definir los objetivos y metas que se quieren alcanzar. Además, sugiere una revisión e investigación previa acerca de la problemática que se quiere resolver [17].

Fase de comprensión de los datos

Se da el primer contacto con el problema ya que comprende la recolección, descripción y exploración de datos que ayuden a cumplir el objetivo [17]. Es aquí donde se familiariza e identifica la calidad de estos [17].

Fase de preparación de datos

Los datos requieren un procesamiento tomando en cuenta la técnica de modelado que sea elegida [17]. Por esa razón aquí se da una selección, limpieza, normalización, discretización de datos, entre muchas otras [17]. Con el fin de tener un conjunto de datos adecuado al modelo que se implementará.

Fase de modelado

Aquí se plantean las técnicas de modelado más convenientes para el proyecto de minería de datos [17]. Luego se construye el modelo con los parámetros óptimos de

cada modelo en un proceso iterativo comparando y evaluado los resultados que arrojan los modelos [17].

Fase de evaluación

En esta fase se trata de averiguar si por alguna razón el modelo es deficiente, si es así se debe revisar si se tiene que rehacer pasos anteriores en los que posiblemente se hayan cometido un error [17].

Fase de despliegue

Finalmente, el modelo listo y probado, se transforma en conocimiento [17]. Se debe documentar y presentar los resultados de forma inteligible, con el propósito de aumentar el conocimiento [17].

CAPÍTULO 4. METODOLOGÍA

Para el presente proyecto de investigación, se utilizó las prácticas y principios que fomenta Scrum para el desarrollo de proyectos o productos. Por lo tanto, se define los roles de responsabilidad para el presente proyecto, esta descripción se encuentra en la Tabla 1.

Tabla 1. Roles Scrum

Roles de Scrum	
Scrum Master	Luis Velásquez
Product Owner	Patricio Zambrano
Team	Luis Velásquez

Definición de Requerimientos

En el inicio del desarrollo del proyecto de investigación, se pactó la primera reunión con el *Product Owner*, quien manifestó la necesidad de realizar una investigación que permita identificar cual o cuales técnicas de modelamiento de tópicos arrojan mejores resultados al ser aplicados al estudio de ataques de Ingeniería Social. Esto con el fin de ser un aporte valioso para el contenido del *paper "On the modeling of cyber-attacks associated with social engineering: A parental control prototype"*. Los requerimientos acordados se muestran en la Tabla 2.

Tabla 2. Requerimientos Funcionales

ID	REQUERIMIENTO

R01	Entender qué son los modelos de tópicos y el ciclo de un ataque de ingeniería social.
R02	Investigar, recopilar de fuentes científicas o académicas información que permita verificar el uso de modelos de tópicos.
R03	Seleccionar cuatro modelos de tópicos que tengan un respaldo académico o científico y su implementación en lenguaje de programación Python.
R04	Comprender el funcionamiento completo de cada uno de los modelos seleccionados.
R05	Implementar los modelos seleccionados.
R06	Analizar los resultados obtenidos
R07	Concluir cuál es el modelo que mejor se adapta al estudio de ataques de Ingeniería Social

Historias épicas de Usuario

Ahora, se presenta en la Tabla 3 las historias épicas de usuario creadas a partir de los requerimientos antes descritos.

Tabla 3. Historias épicas de usuario

ID	Épica de Usuario
U01	Como investigador requiero la recopilación de información de artículos científicos que sustenten la utilización de modelos de tópicos para seleccionar cuatro modelos que tengan un sustento científico y su aplicación con Python.
U02	Como investigador requiero un análisis riguroso de los cuatro modelos seleccionados para tener un mayor nivel de precisión y análisis de los resultados que arrojen los modelos.
U03	Como investigador requiero la implementación en Python de los cuatro modelos de tópicos en el ámbito de ataques de ingeniería social (bullying y grooming) para obtener agrupaciones de palabras de acuerdo a cada modelo.
U04	Como investigador requiero el análisis y comparación de los resultados arrojados por los modelos seleccionados para distinguir el o los modelos de tópicos que mejor se adaptan a la naturaleza de los datos (ataques de buillying y grooming).

Product Backlog

A continuación, se procede a crear el *Product Backlog* del proyecto de investigación, el cual consta de la historia de usuario, la estimación y el nivel de prioridad. En la Tabla 4 se muestran los valores de prioridad.

Tabla 4. Valores de prioridad

VALORACIÓN	PRIORIDAD
1	Media
2	Alta
3	Muy alta

Por tanto, en la Tabla 5 se muestra el *Product Backlog* inicial.

Tabla 5. Product Backlog

Product Backlog				
Épica de Usuario	ID	Historia de Usuario	Estimación (días)	Prioridad
U01	U01-01	Recopilar artículos científicos de modelamiento de temas	5	3
	U01-02	Analizar los artículos científicos	4	3
	U01-03	Clasificar los artículos científicos	3	2
	U01-04	Seleccionar 4 modelos de tópicos	1	3
U02	U02-01	Análisis de modelo NMF	4	3
	U02-02	Análisis de modelo LSI	4	3
	U02-03	Análisis modelo LDA	4	3
	U02-04	Análisis modelo HDP	4	3
U03	U03-01	Implementación de los cuatro modelos	4	3
U04	U04-01	Análisis y comparación de resultados	1	2

Planificación del *Release*

Como se pude apreciar en la Tabla 6, la duración del proyecto es aproximadamente de cinco semanas. El primer Sprint es el que lleva más tiempo pues aquí se realiza una investigación profunda a cerca de los artículos que se utilizarán para la creación del *paper*. Además, se investiga acerca del ciclo de vida de un ataque de ingeniería social.

El Sprint 2 y 3 tienen una duración de una semana cada uno. El sprint 4 tiene una duración aproximada de una semana.

Tabla 6. Planificación Release

Historia de Usuario	Estimación
Sprint 1	
U01-01	5
U01-02	4
U01-03	3
U01-04	1
Total	13
Sprint 2	
U02-01	4
U02-02	4
Total	8
Sprint 3	
U02-03	4
U02-04	4
Total	8
Sprint 4	
U03-01	4
U04-01	1
Total	5

Además, se añadió un Sprint 0, el cual tiene como objetivo familiarizarse con las definiciones fundamentales de modelamiento de tópicos, además de buscar fuentes de consultas para artículos científicos, y la instalación de las herramientas de software que permitan implementar los modelos en Python.

CAPÍTULO 5. IMPLEMENTACIÓN

Continuando con la metodología Scrum se tiene la puesta en marcha de cada uno de los Sprint planificados, por esa razón se detalla a continuación las actividades realizadas en cada uno.

Sprint 0

El sprint 0 por lo general se lo utiliza para tareas que no aportan un valor tangible en el desarrollo del proyecto. Se utilizó este tiempo para la familiarización del modelamiento de temas.

Además, se definió e instaló las herramientas y software necesario para la implementación de los modelos en el lenguaje de programación Python. En la Tabla 7 se detallan las herramientas empleadas.

Tabla 7. Descripción de software utilizado

Software	Descripción
 Anaconda	Es un software de distribución libre y abierta para los lenguajes de programación R y Python. Comúnmente utilizada para aprendizaje automático, ciencia de datos, procesamiento de grandes volúmenes de datos, entre otros.
 Jupyter Notebook	Es una aplicación que se ejecuta en un navegador web, es de código abierto. Soporta la creación de documentos que contienen código en Python, además soporta visualizaciones y texto narrativo.

Se obtuvieron las bases de conceptos, definiciones tanto del modelamiento de tópicos como de la ingeniería social. Además de un entendimiento general del ciclo de vida de ataques como el grooming y el bullying, principalmente de los papers [2,3,6].

Sprint 1

Basándose en la metodología CRISP-DM y acorde a lo planteado en la planificación del *Release*, en esta fase de la investigación, se recopilo y eligió artículos científicos que comparan al igual que describen diferentes modelos que forman parte de la modelización de temas. Los artículos elegidos fueron obtenidos de fuentes confiables tales como IEEE Xplore y Springer Link. Los artículos científicos tuvieron una clasificación, relacionándolos de acuerdo con el tema que abarcan en su contenido. La Tabla 8 muestra los artículos que sirvieron como revisión de literatura.

Tabla 8. Revisión de Literatura de modelos de temas

Revisión de Literatura	Referencia
Modelos de temas que utilizan el contexto semántico para mejorar la clasificación de documentos	[20][21][22][23][24] [25][26][27][28]
Modelamiento de temas en textos cortos basado en redes de coocurrencia de palabras	[29]
Modelos de tópicos que identifican los temas a medida que surgen en el tiempo	[30][18][31][19][32]
Incorporación de características temáticas para mejorar la precisión de la agrupación de documentos	[33]
Modelo temático multi partes mejora el restablecimiento de información y la clasificación de documentos	[34]
Modelos de temas que detectan automáticamente patrones recurrentes de expresiones	[35]
Modelamiento de temas mediante patrones de coocurrencia de palabras.	[36]
Clasificación de textos mediante LDA o variantes de LDA	[37][38][39][40][41] [42][43][44][45][46] [47][48][49][50][51] [52][53]
La incrustación de palabras para mejorar el modelamiento de temas	[54][55][56][57][58] [59]
Incorporación de frases en los modelos de temas para añadir coherencia	[60][61]
La computación paralela como modelo de temas	[62]
Modelización de temas con redes neuronales artificiales	[63]
Modelamiento de temas para textos breves mediante patrones de palabras	[42][64]
Modelo LDA aplicando distancias con distribuciones de probabilidad	[65][66][67]
Modelo LDA aplicando ponderaciones a las palabras de la muestra	[68][69]
Modelos de temas que obtienen resúmenes contextualizados	[70]
Modelos de tópicos que realizan un análisis teniendo en cuenta la atención humana	[71]

Modelos de temas que filtran y reducen la generalidad e impurezas de los documentos	[72][73][74]
Modelo de tópico que añade un parámetro de distribución de categorías a LDA	[75]
Modelo temático que tiene en cuenta las entidades presentes en documentos	[76]
Modelamiento de temas que aprovecha el orden secuencial de las frases y la relación entre frases sucesivas	[77]
Modelo de tema centrado en resolver la escasez de datos y mejorar la seguridad de estos	[78]
Modelado de temas centrado en las citas y los títulos de los documentos	[79]
Modelos de temas que implementan árboles temáticos jerárquicos	[80]
Modelo de temas que relaciona textos cortos con documentos largos	[81]

Luego de la revisión de la literatura se puede comprobar que no existen artículos científicos que se apliquen o relacionen con ataques de ingeniería social.

Por otro lado, los modelos seleccionados fueron NMF (Factorización de Matrices No Negativas), LSI (Indexación Semántica Latente), LDA (Asignación Latente de Dirichlet) y HDP (Proceso Jerárquico Dirichlet). Cada uno de estos modelos fue seleccionado porque tienen un fuerte respaldo teórico y práctico, es decir existen investigaciones en artículos científicos dónde se utilizan los modelos, además, cuentan con librerías de código abierto para su implementación en Python.

Sprint 2

Siguiendo con la planificación del *Release,* el principal objetivo de esta fase es el entendimiento detallado de cada uno de los modelos que se van a implementar. Es así como se presenta un análisis detallado de los modelos NMF y LSI.

NMF Factorización de Matriz No Negativa

Es una técnica de reducción dimensional [82], propuesto por Lee en [83] la cual consiste en descomponer (factorizar) una matriz de números no negativos en dos matrices denominadas como W (matriz base) y H (matriz de expansión). En la Figura 2, se puede observar una representación del modelo NMF.

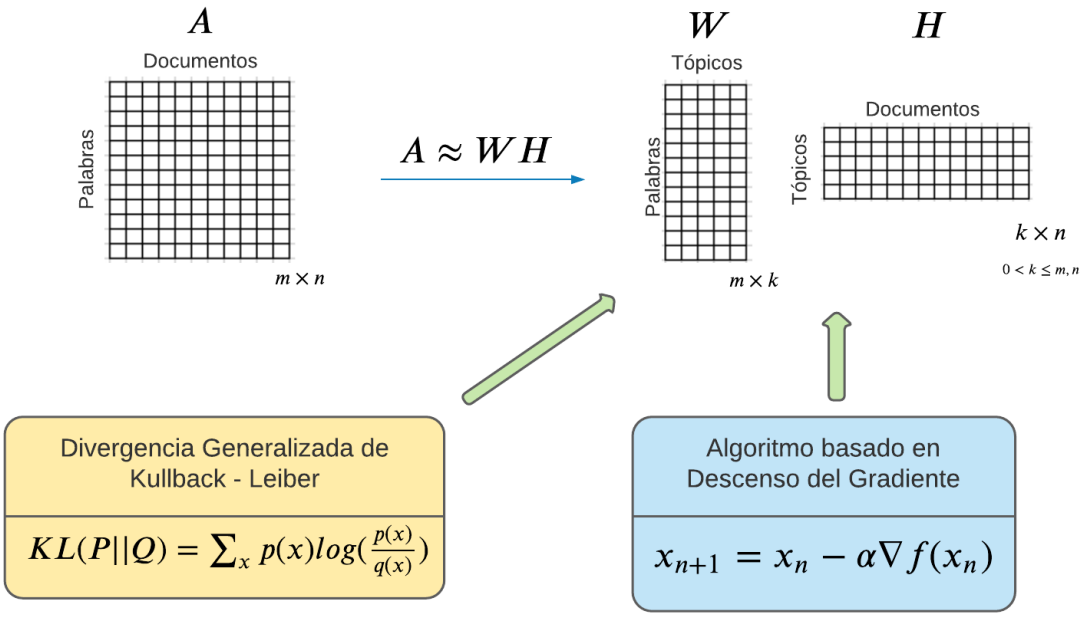


Figura 2. Modelo NMF

Los valores que conforman las matrices W y H se obtienen de la factorización de la matriz de entrada, ver Figura 3. Estos valores son comparados con la matriz original mediante la divergencia generalizada de Kullback-Leiber, la cual no es más que una métrica para mediar la discrepancia entre los datos de entrada y la aproximación que se obtiene después de la factorización. Además, durante la optimización de valores se aplica el descenso del gradiente proyectado (PGD) para evitar que los números sean conviertan en negativos, pero sobre todo es lo que permite que los valores se aproximen al mínimo error [82].

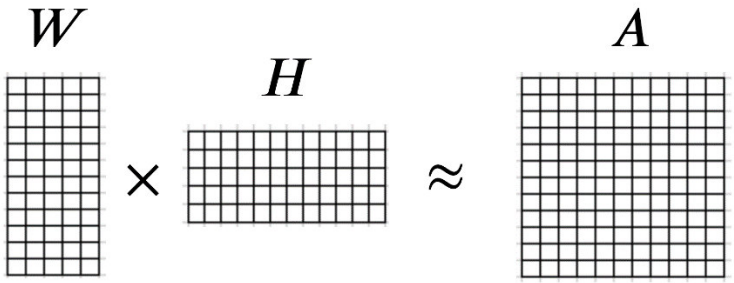


Figura 3. Factorización NMF

La divergencia generalizada de Kullback-Leiber, es la alternativa con mejores resultados en el campo del modelamiento de temas en NMF [82], ver Ecuación 1. Otras métricas aplicadas en el modelo NMF no han dado resultados satisfactorios en el campo del análisis de texto. El descenso del gradiente es el algoritmo que le permite al modelo NMF ser considerado un modelo no supervisado, ver Ecuación 2. Este algoritmo itera en los datos hasta conseguir la convergencia, es decir llegar al modelo óptimo.

$$KL(P||Q) = \sum_{x} p(x) log\left(\frac{p(x)}{q(x)}\right)$$

Ecuación 1. Divergencia Kullback – Leiber

$$x_{n+1} = x_n - \alpha \nabla f(x_n)$$

Ecuación 2. Descenso del gradiente

Otra alternativa para el algoritmo PGD es Alternating Direction Method of Multipliers o ADMM por sus siglas en inglés, es un algoritmo que utiliza la descomposición en partes pequeñas o subproblemas de un problema general puesto que cada subdivisión es más fácil de resolver [84]. Tiene un mayor costo computacional, sin embargo, las iteraciones que debe alcanzar para converger son menores. Al igual que el PGD tiene como objetivo optimizar los datos de la matriz hasta encontrar los datos que mejor se acerquen a la realidad.

LSI Indexación Semántica Latente

Es un modelo de tópico propuesto por Deerwester en [85]. Aprovecha la existencia de una estructura semántica entre documentos de textos y las palabras que los componen.

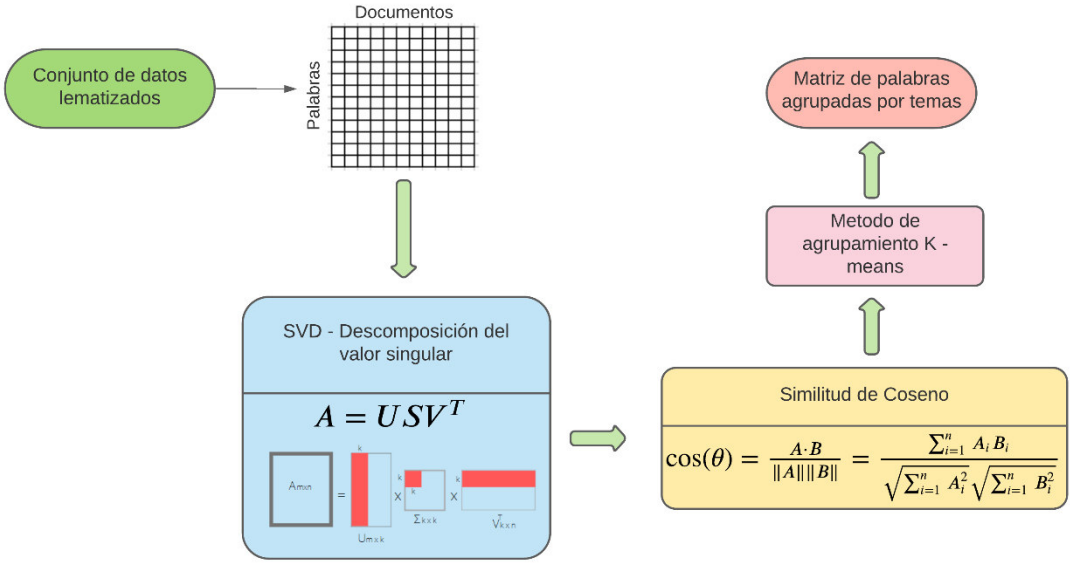


Figura 4. Modelo LSI

Como se puede apreciar en la Figura 4, LSI utiliza una técnica matemática conocida como SVD (Descomposición de valor singular). SVD requiere de una matriz de entrada comúnmente denominada una matriz de términos (palabras) y documentos. Esta matriz se descompone en tres matrices como lo dicta la misma técnica. Esta descomposición es similar al modelo NMF sin embargo la diferencia más notoria es que permite los valores negativos. El análisis estadístico que brinda LSI es netamente matemático, es decir no requiere de estructuras de ayuda como diccionarios, además lo hace independiente de un lenguaje [86].

Dentro del modelo LSI, se aplican varios algoritmos que permiten el entrenamiento y optimización de resultados, entre estos se tiene la similitud del coseno, la cual permite encontrar los valores más cercanos entre los elementos de la matriz. También se puede encontrar el algoritmo de agrupamiento *K-means* el cual permite que LSI se convierta en un modelo no supervisado. EL objetivo de este algoritmo es encontrar grupos en datos que aparentemente no tienen relación.

SVD

La descomposición de valor singular tiene como objetivo el representar una matriz de entrada en tres matrices diferentes conocidas como USV^T en algunas investigaciones las matrices son nombradas $U\Sigma V^T$. Lo que pretende esta técnica es reducir la matriz de entrada a un espacio donde se puedan realizar cálculos, análisis y comparaciones, ver Figura 5. Es decir, SVD presenta a las palabras y a los documentos como vectores en

un espacio dimensional reducido. Se puede encontrar patrones de relación entre los términos de una colección de datos, como por ejemplo que tan similares son dos documentos, o dos palabras y analizar la relación de la palabra en el documento. Hay que recalcar que SVD tiene un coste computacional elevado.

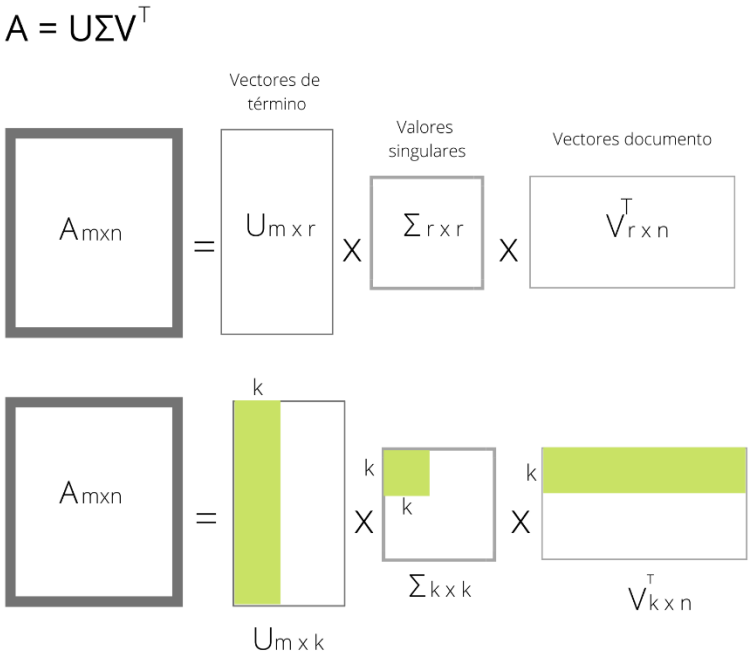


Figura 5. Descomposición de valores singulares

Las matrices U, V son matrices unitarias es decir cumplen la propiedad $U^TU = UU^T = I_{nxn)}$ (matriz identidad). La matriz S o Σ es una matriz diagonal cuyos valores son decrecientes.

Similitud de cosenos

Dentro de LSI los documentos son vectores que se comparan a través de la métrica de similitud de cosenos. Es una ecuación que mide el coseno del ángulo que existe entre dos vectores (documentos) proyectados en un espacio multidimensional. Los valores más cercanos a 1 son los que tienen un mayor grado de similitud con respecto a los valores que se acercan más al 0.

Para encontrar esta medida se aplica la siguiente fórmula a los vectores que se quieren comparar, Ecuación 3.

$$\cos(\theta) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Ecuación 3. Similitud Coseno

Algoritmos para factorizar SVD

Como se mencionó anteriormente, la técnica SVD demanda cierta capacidad computacional cuando se trata de matrices con un extenso número de datos. A lo largo del tiempo se han ido implementado diferentes algoritmos tratando disminuir O(n). Los algoritmos que utiliza la librería gensim se basan en [87], y se describen a continuación.

El primer algoritmo, se lo denomina algoritmo estocástico de dos pasadas [87], no es la mejor opción en descomposiciones de gran escala. Su computo requiere O(nk + mk) de memoria. Sin embargo, los autores en [87] optimizan el algoritmo para que los datos que se almacenan en memoria no la sobrepasen, alcanzando una notación aceptable de O(mk) que es independiente del flujo de entrada de n para poder realizar dos pasadas a la matriz.

El segundo algoritmo, se lo conoce como algoritmo de una sola pasada y se caracteriza porque mientras se mantenga constantes los requisitos de memoria se pueden procesar infinitos flujos de entrada [87].

Existe un tercer algoritmo, denominado hibrido, el cual utiliza el núcleo del algoritmo estocástico en el marco de una sola pasada [87]. Sin embargo, el rendimiento no es óptimo, aunque el tiempo de procesamiento es menor también se pierde la precisión de los datos.

Teniendo en cuenta los resultados en [87] los creadores de LSI para la librería Gensim proponen solo los dos primeros algoritmos para su utilización. El algoritmo de una sola pasada es el que se instancia por defecto para el modelo LSI, dependiendo de las necesidades puede ser intercambiado por el algoritmo estocástico de dos pasadas. En el presente trabajo se utilizó el algoritmo por defecto.

Algoritmo *K-means*

Es un algoritmo de aprendizaje no supervisado, recomendado en datos que no están etiquetados, en otras palabras, datos que no están definidos por una característica o en algún grupo [88]. Es un algoritmo rápido, robusto y simple. Primero se selecciona el número de agrupaciones *k* que se quiere identificar, si no se especifica, Gensim creará por defecto 200 agrupaciones (tópicos). Luego el algoritmo elegirá aleatoriamente k números (puntos - centroides) de los datos que arroja SVD. Después, mide la distancia (similitud de coseno) de cada uno de los datos con los puntos aleatorios y los asigna al punto que se encuentra más cerca. Luego se vuelve a seleccionar k puntos aleatorios y se repite el proceso, ver Figura 6. Al final se compara cada análisis, y se elige el que tuvo la mejor distribución entre puntos.

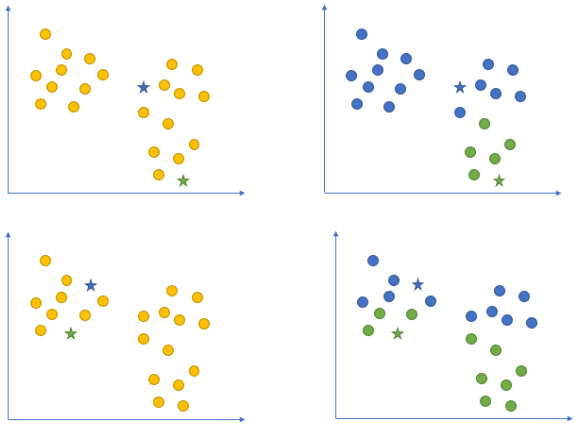


Figura 6. Ejemplo de K-mean

Sprint 3

En esta fase se completa el análisis de los dos modelos restantes, LDA y HDP. Se describe cada modelo de tópico, se entiende y fabrica un diagrama que muestre el funcionamiento de cada uno. A continuación, se describen los modelos.

LDA Asignación de Dirichlet Latente

Asignación de Dirichlet Latente o por sus siglas en ingles LDA, fue desarrollado por Blei [89]. Es un modelo probabilístico generativo de un conjunto de textos. Su flujo representativo se muestra en Figura 7.

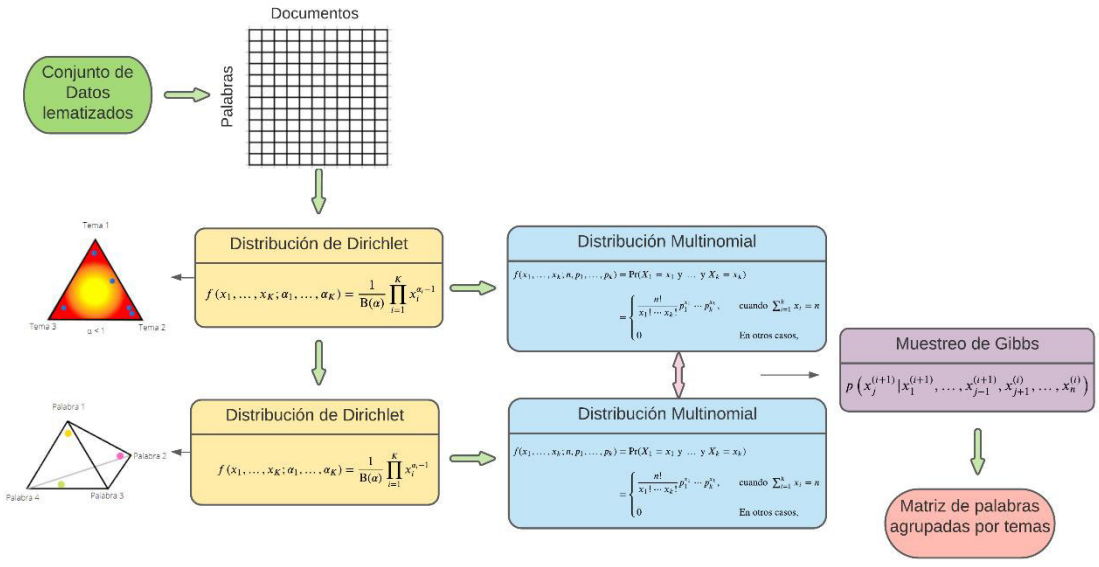


Figura 7. Modelo LDA

Utiliza la distribución de Dirichlet para modelar los documentos con temas, y temas con palabras como se indica en Figura 9. Además, utiliza distribuciones multinomiales para cuantificar esa relación. Para una optimización de los resultados implementa el algoritmo

de muestreo de Gibbs sobre las distribuciones de Dirichlet para deducir la asignación de temas a los datos [90], ver Figura 10.

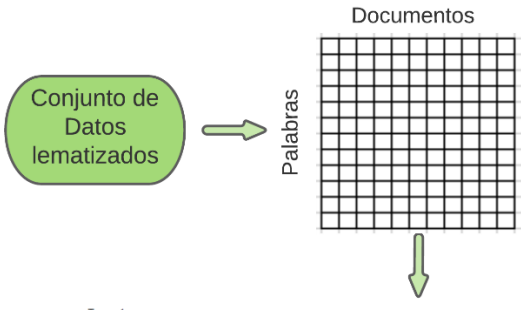


Figura 8. Modelo LDA - Fase 1

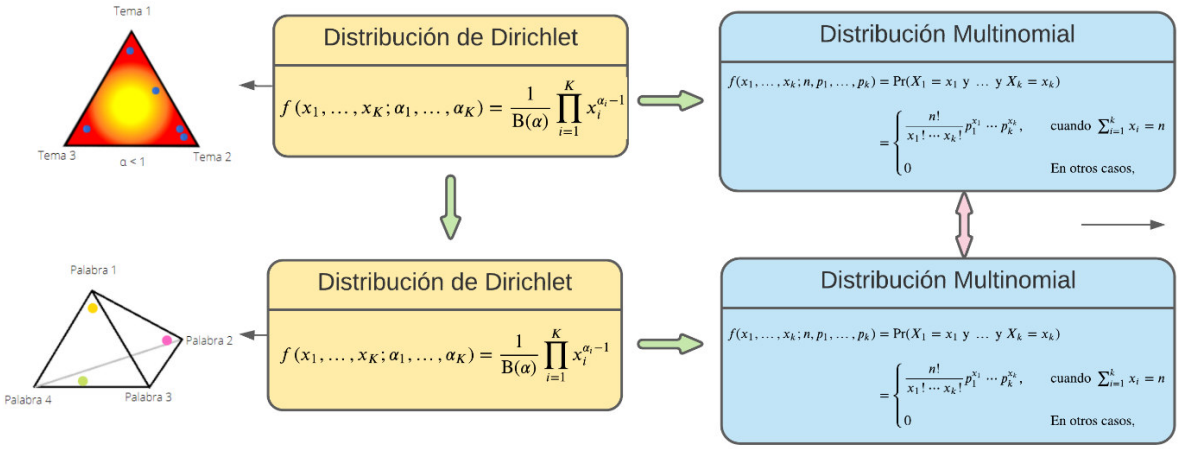


Figura 9. Modelo LDA - Fase 2

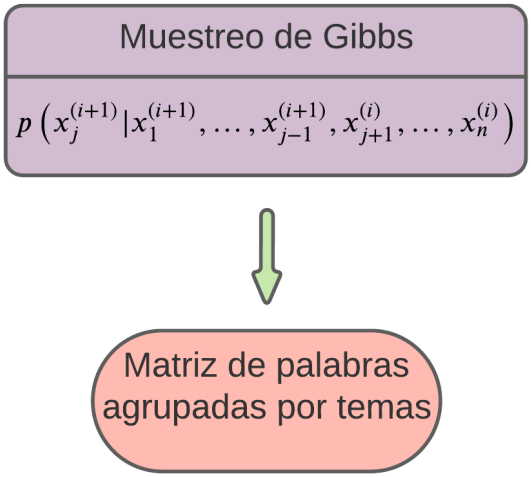


Figura 10. Modelo LDA - Fase3

Un diagrama conocido para la representación de LDA se muestra en Figura 11.

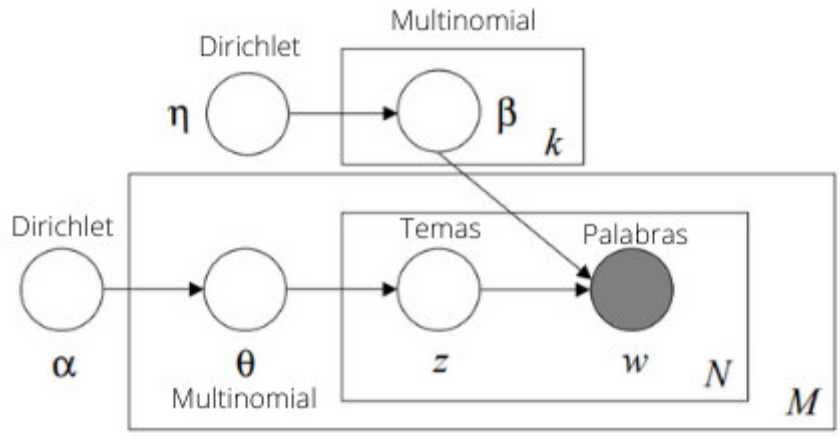


Figura 11. Representación gráfica del modelo LDA [81]

Distribución de Dirchlet

Es conocida como una distribución de distribuciones. Dentro del modelamiento de temas, específicamente en LDA, se aplica dos distribuciones Dirichlet, una para cuantificar la probabilidad de pertenencia de un documento a los diferentes temas encontrados. La segunda distribución, de igual manera permite ubicar la probabilidad de relación que tiene un tema dependiendo de cada palabra [90]. Una particularidad de la distribución de Dirichlet es que los valores tienden a los extremos.

Por ejemplo, en una distribución de 3 dimensiones esta formará un triángulo, por tanto, mientras los hiper parámetros sean menores que 1 los valores tienden a las esquinas. Respectivamente en LDA, el hiper parámetro alfa α corresponde a la primera distribución Dirichlet y el hiper parámetro beta β para la segunda. Estos hiper parámetros sirven para ajustar la distribución en el espacio.

Siguiendo con el ejemplo anterior, cada vértice del triángulo representa un tema, ver Figura 12. Cada punto dentro del triángulo es un documento. El documento en el espacio es un vector con 3 valores de probabilidad, uno por cada tema.

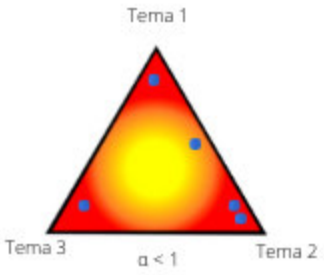


Figura 12. Representación Dirichlet Temas – Documentos

Para la siguiente distribución de Dirichlet, en la Figura 13, se aprecia un tetraedro (4 dimensiones), cada vértice representa una palabra, y cada punto representa un tema. De la misma manera, para este caso, cada punto (tema) cuenta con un vector de 4 valores de la probabilidad dependiendo de cuan probable es que se relacione con la palabra.

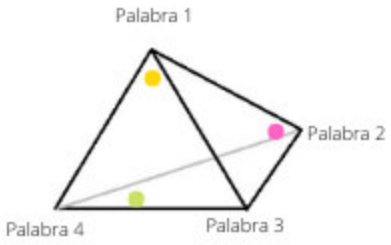


Figura 13.Representación Dirichlet Palabra - Temas

Distribución Multinomial

Es la generalización de una distribución binomial, en la que cada ensayo de Bernoulli tiene 3 o más valores como resultados posibles. Es decir, una variable al azar “x” que

precisa el número de veces que se ha presentado el valor “i” sobre el suceso “n”, donde cada suceso tiene un único valor de los posibles “k” de las probabilidades “p” [91]. La función de probabilidad de una distribución multinomial se muestra en la Ecuación 4.

$$\mathrm{f}(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \Pr(X_1 = x_1 y \ldots y X_k = x_k)$$

$$= \begin{cases} \dfrac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_i} \\ 0 \end{cases}$$

Ecuación 4.Distribución Multinomial

Muestreo de Gibbs

Es un caso especial de la cadena de Markov Monte Carlo por sus siglas en inglés (MCMC). El algoritmo reasigna la probabilidad de una palabra de pertenecer a un tema, basándose en un muestreo aleatorio de cualquier palabra tomando como valores iniciales la distribución Dirichlet [90]. Esto con el fin de que el modelo aprenda y clasifique de forma correcta las palabras en cada tema.

HDP Proceso de Dirchlet Jerárquico

Un proceso de Dirchlet Jerárquico es una colección de procesos de Dirichlet, propuesto por Teh [92].

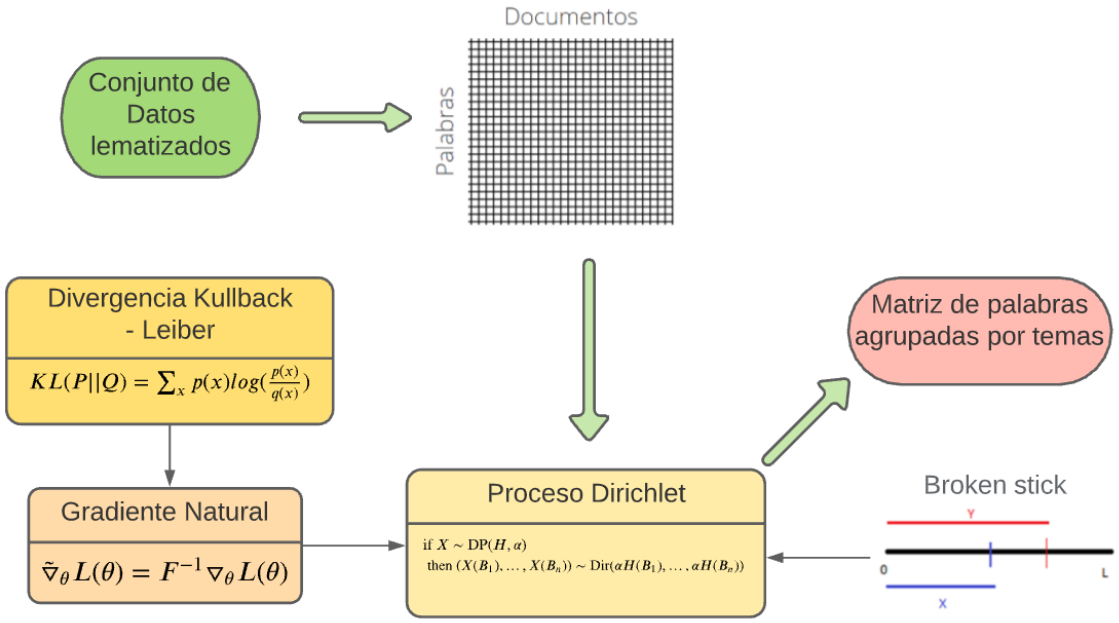


Figura 14. Modelo HDP

La estructura de HDP está definida por distribuciones embebidas. Consta de una distribución Dirchelet que tiene como un parámetro otra distribución Dirchlet, ver Figura 14. Se crean dos niveles, en el primero constan los números de temas que se generan con el conjunto de documentos (corpus). El segundo nivel son los temas que se forman en cada documento [93]. Para un mejor entendimiento se plantea el problema del rompimiento de un palo (Stick-breaking).

Propone que se tiene un palo de tamaño L el cual es partido en dos partes de manera aleatoria Y y L-Y, como se indica en Figura 15. Uno de los dos pedazos de nuevo es partido en dos de manera aleatoria. Entonces se tienen 3 pedazos X, Y -X y L-Y, Figura12. El problema radica en que, si se quiere calcular cualquier valor de la variable aleatoria X, como su varianza o esperanza, va a depender del primer rompimiento aleatorio Y que tuvo el palo y así sucesivamente.

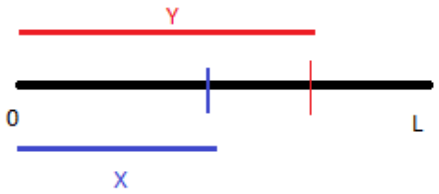


Figura 15. Stick Breaking

Es decir, dentro del modelado de temas, para hallar un tema dentro de un documento, primero se deben inferir el número de temas en el corpus total. HDP infiere el número de temas a través de los datos y genera dichos datos a través de una distribución Multinomial.

El proceso de Dirichlet empieza con una distribución Dirichlet G0, con dos parámetros como se puede ver en Ecuación 5. El primero parámetro γ conocido como parámetro de concentración es un número real positivo, y H es una distribución de probabilidad Dirichlet simétrica [93], esto con el fin de que el primer nivel (temas en el corpus) sea un número discreto. En la segunda distribución Dirchlet, Gj se pasa como parámetro G0 para garantizar que todos los documentos comparten el mismo número de temas [93]. Gj hereda los temas de G0, pero los pondera de acuerdo con las proporciones de los temas específicos del documento [93].

$$G_0 \sim DP(\gamma, H)$$

$$\mathrm{G_j}|\mathrm{G_0} \sim D\mathrm{P}(\alpha, \mathrm{G_0})\mathrm{forj} = 1, \ldots, \mathrm{J}$$

Ecuación 5. Proceso Dirichlet

Luego se genera el tema asociado a la enésima palabra del jº documento; para después generar la palabra a partir de ese tema a través de una distribución Multinomial, como se indica en Ecuación 6 [93].

$$\theta_{jn} \sim G_j, \quad w_{jn} \sim \mathrm{Mult}(\theta_{jn})$$

Ecuación 6. Distribución Multinomial en modelo HDP

Posterior a esto se calculan varios parámetros para actualizar los datos a nivel de documento, mediante la Divergencia KL la cual ayuda a comparar valores para su posterior actualización. Se calcula el gradiente natural a nivel de corpus para definir la tasa de aprendizaje y actualizar datos a nivel de corpus. Esto de manera infinita hasta que no se cumpla una condición de parada, el límite de tiempo expire o todo el corpus haya sido procesado [93].

Sprint 4

En esta etapa, tomando en cuenta la planificación de *Release*, se ponen en marcha los modelos previamente seleccionados y analizados. Como se menciona en el Sprint 1, estos modelos cuentan con su aplicativo a través de una librería en Python. Esta librería es conocida como Gensim.

Es fundamental en esta etapa de la implementación contar con los datos que van a ayudar a construir e implementar los diferentes modelos de tópicos. Por esta razón nos valemos nuevamente de la metodología CRISP-DM. Se tiene la etapa de comprensión de los datos. Como indica la guía, aquí se da el primer contacto con datos reales del problema. Una recopilación de datos de mensajería instantánea relacionados con ataques de bullying (acoso) y grooming (acoso, abuso sexual en línea). Los datos fueron proporcionados por el *Product Owner.* alineándose a los requerimientos necesarios para ser un aporte a su investigación.

Una vez que los datos hayan sido recopilados viene la etapa de preparación de datos. Para esto se describe el procedimiento para el tratamiento previo de los datos antes de la fase de modelado.

En primer lugar, se importan las librerías necesarias para el tratamiento de datos y la construcción de los modelos, ver Figura16.

```
import re
import numpy as np
import pandas as pd
from pprint import pprint

# Gensim
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel

# spacy for lemmatization
import spacy
```

Figura 16. Importar librerías

Después se define el grupo de palabras que deben ser eliminadas al no tener un aporte significativo para el entrenamiento de los modelos. A estas palabras se las conoce como *stop words* o palabras vacías (en su mayoría son artículos). Esto se evidencia en Figura 17.

```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use'])
```

Figura 17. Definir palabras vacías

Luego se carga el conjunto de datos obtenidos en la fase anterior, este paso se lo realizó con los dos conjuntos de datos, tanto de bullying como de grooming, ver Figura18

```
df = pd.read_csv('chatsLogs.csv', sep=';')
df.head()
```

	chat_no	message_no	message_text	message_stage
0	1	1.0	hello there my sweet	S4
1	1	2.0	how are you doing	NaN
2	1	3.0	yes it is	S1
3	1	4.0	glad to meet you	S6
4	1	5.0	did you look at my profile	S1

Figura 18. Cargar datos de mensajería instantánea

Posterior a esto se separan las oraciones en palabras a través de las funciones descritas en las Figuras 19 y 20.

```
# Convert to list
data = df.message_text.values.tolist()
```

Figura 19. Convertir datos en una lista

```
def sent_to_words(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))

data_words = list(sent_to_words(data))

print(data_words[:1])
```

```
[['hello', 'there', 'my', 'sweet']]
```

Figura 20. Separar en palabras

Continuando la preparación de datos, Se construyen bigramas y trigramas con la finalidad de proporcionar una probabilidad condicionada entre palabras subsecuentes, ver Figura 21.

```
# Build the bigram and trigram models
bigram = gensim.models.Phrases(data_words, min_count=5, threshold=100)
trigram = gensim.models.Phrases(bigram[data_words], threshold=100)

# Faster way to get a sentence clubbed as a trigram/bigram
bigram_mod = gensim.models.phrases.Phraser(bigram)
trigram_mod = gensim.models.phrases.Phraser(trigram)

# See trigram example
print(trigram_mod[bigram_mod[data_words[0]]])
```

```
['hello', 'there', 'my', 'sweet']
```

Figura 21. Construcción de Bigramas y Trigramas

Luego como indica Figura 22 se construyen funciones que permitan crear esos bigramas y trigramas en el conjunto de datos. Además de una función que permite la lematización de las palabras para una mejor construcción de los modelos de temas.

```
# Define functions for stopwords, bigrams, trigrams and lemmatization
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc)) if word not in stop_words] for doc in texts]

def make_bigrams(texts):
    return [bigram_mod[doc] for doc in texts]

def make_trigrams(texts):
    return [trigram_mod[bigram_mod[doc]] for doc in texts]

def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV']):
    """https://spacy.io/api/annotation"""
    texts_out = []
    for sent in texts:
        doc = nlp(" ".join(sent))
        texts_out.append([token.lemma_ for token in doc if token.pos_ in allowed_postags])
    return texts_out
```

Figura 22. Funciones de tratamiento de datos

Se guardan en variables los resultados que arrojan las funciones antes creadas, para que puedan ingresar como parámetros en la función de lematización, y obtener un conjunto de datos depurado, como se indica en Figura 23.

```
# Remove Stop Words
data_words_nostops = remove_stopwords(data_words)

# Form Bigrams
data_words_bigrams = make_bigrams(data_words_nostops)

# Initialize spacy 'en' model, keeping only tagger component (for efficiency)
# python3 -m spacy download en
nlp = spacy.load('en', disable=['parser', 'ner'])

# Do lemmatization keeping only noun, adj, vb, adv
#data_lemmatized = lemmatization(data_words_bigrams, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV'])
#data_lemmatized = lemmatization(data_words, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV'])
data_lemmatized = lemmatization(data_words_nostops, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV'])
```

Figura 23. Conjunto de datos lematizados

A continuación, se crean las variables necesarias que necesita los modelos de tópicos según la librería Gensim, esto se visualiza en Figura 24.

```
# Create Dictionary
id2word = corpora.Dictionary(data_lemmatized)

# Create Corpus
texts = data_lemmatized

# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
```

Figura 24. Datos depurados

Como plantea la metodología CRISP-DM, ahora se ingresa a la fase de modelado. Es decir, se construye el modelo a partir de los datos depurados. Cabe recalcar que en

todos los modelos de tópicos implementados en esta investigación se utilizó el mismo conjunto de datos depurados de bullying y de grooming.

La librería Gensim ofrece documentación detallada de cada modelo para un mejor entendimiento de los parámetros que se pueden definir

Se utiliza los modelos de Gensim NMF, LSI, LDA y HPD con los mismos parámetros por defecto, salvo el número de tópicos, dado que por la naturaleza del modelo HDP no permite definir esa variable. Los demás modelos se ejecutan definiendo el número de tópicos en 4 y 6 puesto que arroja mejores resultados.

Ahora se muestra el entrenamiento de los modelos. La sintaxis del código es similar para los 4 modelos, ver Figuras 25 y 26. Este modelo se almacena en una variable que posteriormente arrojará los resultados.

```
lsi_model = gensim.models.lsimodel.LsiModel(corpus=corpus,
                                            id2word=id2word,
                                            num_topics=6,
                                            chunksize=100)
```

Figura 25.Modelado LSI Gensim

```
# Build HDP model
hdp_model = gensim.models.hdpmodel.HdpModel(corpus=corpus,
                                            id2word=id2word,
                                            random_state=100,
                                            chunksize=100)
```

Figura 26. Modelado HDP Gensim

Una vez creados los modelos se pasa como parámetro los datos para que el modelo arroje sus predicciones. Cada modelo tiene como salida una matriz de palabras agrupadas por temas. Como ejemplo, en la Figura 27 y 28 se aprecia la salida de los modelos LSI y HDP respectivamente. Se aprecia como ejemplo, las diez primeras palabras, junto a su ponderación de coherencia, que pertenecen a los seis primeros tópicos desde el 0 al 5. La coherencia indica en qué medida una palabra pertenece a cierto tópico.

```
pprint(lsi_model.print_topics())
doc_lsi = lsi_model[corpus]
```

```
[(0,
  '0.601*"want" + 0.482*"go" + 0.342*"would" + 0.212*"know" + 0.180*"be" + '
  '0.162*"see" + 0.128*"lol" + 0.113*"get" + 0.109*"think" + 0.093*"make"'),
 (1,
  '0.730*"go" + -0.623*"want" + 0.224*"be" + -0.104*"would" + 0.063*"get" + '
  '-0.040*"see" + 0.037*"bed" + 0.036*"back" + 0.027*"think" + 0.023*"lol"'),
 (2,
  '0.853*"would" + -0.416*"want" + -0.226*"go" + 0.102*"love" + 0.100*"think" '
  '+ 0.090*"like" + 0.049*"see" + 0.043*"baby" + 0.043*"could" + 0.035*"let"'),
 (3,
  '0.641*"know" + -0.341*"go" + -0.277*"would" + 0.270*"be" + -0.228*"want" + '
  '0.215*"lol" + 0.210*"get" + 0.201*"baby" + 0.192*"see" + 0.101*"good"'),
 (4,
  '0.679*"be" + -0.486*"know" + 0.293*"get" + -0.219*"baby" + -0.214*"girl" + '
  '-0.209*"go" + 0.205*"see" + 0.100*"lol" + 0.064*"sure" + -0.061*"good"'),
 (5,
  '0.820*"get" + -0.407*"be" + -0.196*"see" + -0.183*"know" + 0.162*"girl" + '
  '0.133*"lol" + 0.075*"good" + 0.057*"time" + 0.049*"think" + 0.047*"take"')]
```

Figura 27. Resultado del modelo LSI Gensim

```
pprint(hdp_model.print_topics())#all topics
doc_hdp = hdp_model[corpus]
```

```
[(0,
  '0.008*want + 0.007*go + 0.006*would + 0.006*know + 0.005*see + 0.005*be + '
  '0.004*think + 0.004*good + 0.004*get + 0.004*lol'),
 (1,
  '0.008*want + 0.007*go + 0.006*would + 0.005*know + 0.004*be + 0.004*think + '
  '0.004*see + 0.004*lol + 0.003*get + 0.003*good'),
 (2,
  '0.007*want + 0.007*go + 0.006*would + 0.005*know + 0.004*see + 0.004*be + '
  '0.004*think + 0.003*get + 0.003*good + 0.003*lol'),
 (3,
  '0.007*want + 0.007*go + 0.006*would + 0.005*know + 0.005*think + 0.004*be + '
  '0.004*good + 0.004*get + 0.004*see + 0.003*lol'),
 (4,
  '0.009*want + 0.008*would + 0.007*go + 0.005*know + 0.004*be + 0.004*lol + '
  '0.004*see + 0.004*think + 0.004*get + 0.003*good'),
 (5,
  '0.008*want + 0.007*go + 0.006*would + 0.005*know + 0.004*be + 0.004*think + '
  '0.004*get + 0.004*see + 0.003*good + 0.003*lol'),
```

Figura 28. Resultado del modelo HDP Gensim

Con estos resultados se procede a realizar una evaluación y comparación de los resultados que arrojaron los 4 modelos seleccionados, como lo estipula en CRISP-DM

6. RESULTADOS Y DISCUSIÓN

Al ser los modelos de temas métodos no supervisados, no se pueden realizar una comparación con métricas de rendimiento comunes como el cálculo del error cuadrático medio. En cambio, se debe utilizar métricas como la coherencia, la cual manifiesta de manera objetiva si las palabras agrupadas en un tema tienen sentido entre ellas.

La Figura 29 esquematiza el comportamiento de los datos (relacionados con el grooming) aplicados a los modelos propuestos LSI, LDA, NMF y HDP. Los resultados muestran que los modelos LSI y NMF generan una mayor dispersión entre los valores

de coherencia obtenidos en cada tema. Descartando estos dos modelos del análisis, el modelo LDA y su variante HDP concentran mejor los datos procesados.

Esto podría deberse a que los modelos LDA y HDP utilizan algoritmos robustos para su aprendizaje, muestreo de Gibbs e inferencia variacional respectivamente. Mientras que LSI emplea el algoritmo K-Means el cual posee una tendencia de agrupamiento en forma circular al centroide que define de referencia. Para mejorar los resultados se podrían aplicar otros algoritmos de agrupamiento como los modelos de mezcla Gaussiana por sus siglas en inglés GMM, o el agrupamiento espacial basado en densidad de aplicaciones con ruido (DBSCAN) los cual tiene un diferente criterio de agrupación.

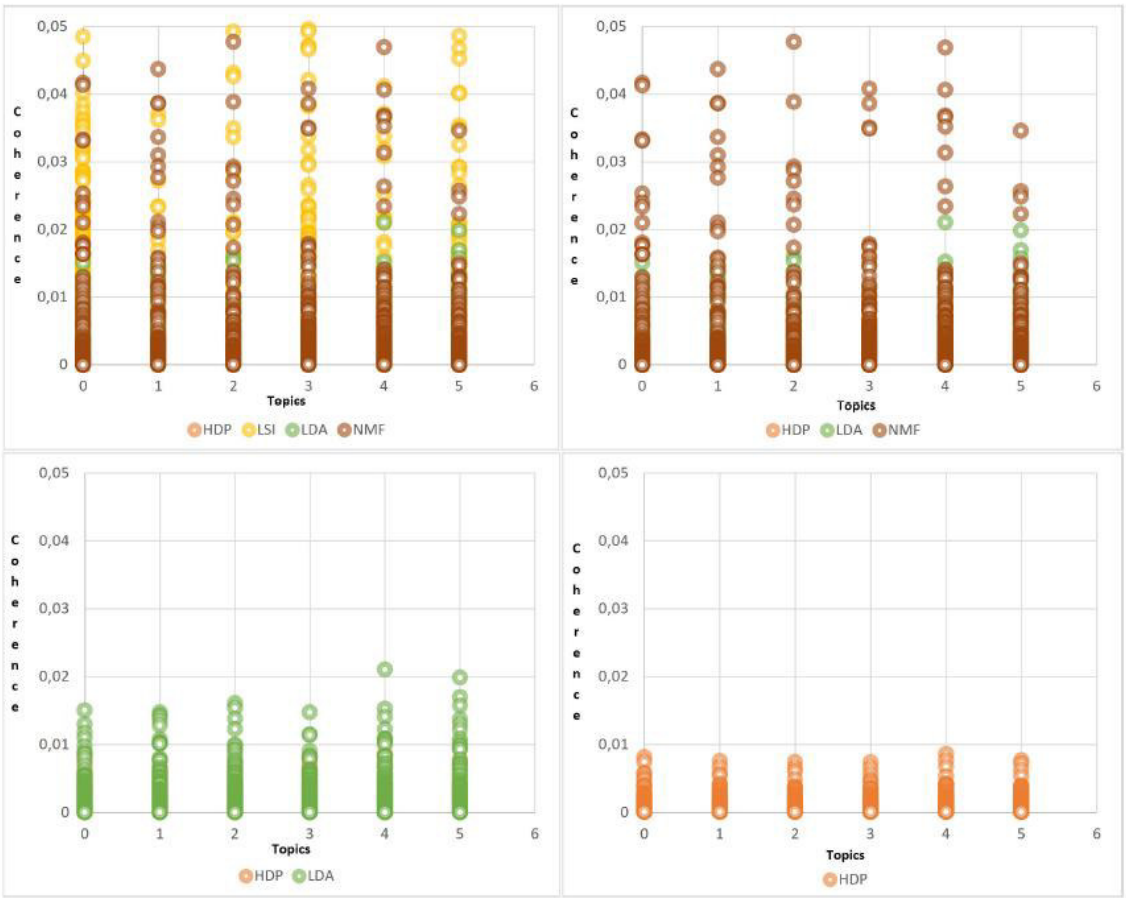


Figura 29. Comparación gráfica del comportamiento de los valores de coherencia en 4 modelos temáticos - Grooming

En cuanto al Bullying, ver Figura 28, el comportamiento de los modelos es similar al caso del Grooming con la diferencia de que tiene cuatro temas. Del mismo modo, se descartan los modelos LSI y NMF ya que los valores de coherencia son muy dispersos respecto a los resultados de LDA y su predecesor HDP. El modelo HDP, aplicado en los 2 casos de estudio, ofrece mejores resultados en la clasificación de palabras por temas. Cabe destacar que el HDP es un modelo mejorado respecto al LDA.

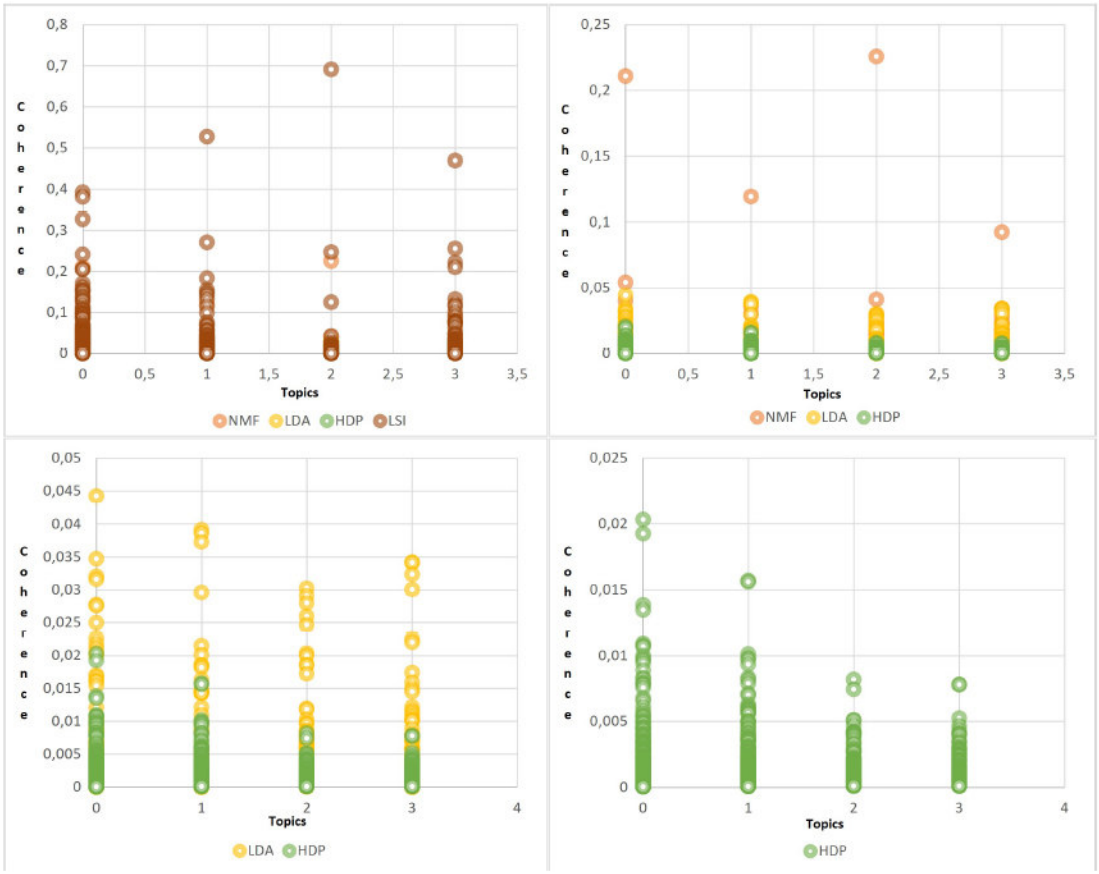


Figura 30. Comparación gráfica del comportamiento de los valores de coherencia en 4 modelos temáticos - Bullying

Con esta premisa, se concluye que el modelo LDA o sus variantes (HDP) se adaptan adecuadamente a los textos cortos.

7. CONCLUSIONES

- Las diferentes técnicas de modelamiento de tópicos que se implementaron en este proyecto arrojaron resultados admisibles en la investigación. En su mayoría se ajustan de manera adecuada a los datos relacionados con ataques de ingeniería social. Sin embargo, los modelos NMF y LSI contienen dentro de sus componentes varios algoritmos que pueden ser mejorados o sustituidos con otros superiores. Así se encontró que el algoritmo K-means de agrupamiento utilizado por LSI es susceptible de fallos si la distribución de los datos no se alinea con su tendencia al agruparlos.
- Los modelos LSI, NMF, LDA y HDP formaron parte de la investigación dado a su fuerte respaldo teórico/práctico en distintas investigaciones y que además cuentan con librerías en lenguajes de programación robustos, facilitando su uso en cualquier campo como en la presente investigación al emplearse en un conjunto de datos de ataques de ingeniería social.
- La coherencia es una métrica asignada a las palabras que forman un tópico, ayuda a medir cuan relacionadas están las palabras entre sí. Los resultados de la implementación de los modelos son variados, sin embargo, son aceptables en términos de coherencia, como se pudo observar en las gráficas de comparación, la mayoría de los datos mantienen mismo rango de coherencia entre sí, mientras el número de tópicos sea entre cuatro y seis, muy pocos datos resultaron atípicos, esto es un buen indicador teniendo en cuenta la naturaleza predictiva de los modelos.
- De manera cuantitativa los valores de coherencia son parecidos entre modelos, destacándose LDA y HDP por mantener una coherencia estandarizada entre cada tema. Desde el punto de vista cualitativo los modelos LSI y NMF presentan una mayor irregularidad entre sus datos, concluyendo que estos no son suficientemente precisos para ser implementarlos con datos de “textos cortos” estandarizar en el documento.
- Finalmente se seleccionó a los modelos de temas LDA y HDP como los modelos que ofrecen una mejor adaptación a la naturaleza de los datos de mensajería instantánea puesto que generaron una mayor precisión predictiva al conjunto de datos propio de ataques de ingeniería social.

8. REFERENCIAS BIBLIOGRÁFICAS.

[1] R. Von Solms, J. Van Niekerk, From information security to cyber security, computers & security 38 (2013) 97-102.

[2] P. Zambrano, J. Torres, L. Tello-Oquendo, R. Jácome, M. E. Benalcazar, R. Andrade, W. Fuertes, Technical mapping of the grooming anatomy using machine learning paradigms: An information security approach, IEEE Access 7 (2019) 142129-142146.

[3] P. Zambrano, J. Torres, _A. Yanez, A. Macas, L. Tello-Oquendo, Understanding cyberbullying as an information security attack|life cycle modeling, Annals of Telecommunications (2020) 1-19.

[4] C. Hadnagy, Social Engineering: The Art of Human Hacking, 1st ed. Indianápolis: Wiley Publishing, Inc, 2010, pp. 9,10,11.

[5] J. M. Hatfield, Social engineering in cybersecurity: The evolution of a concept, Computers & Security 73 (2018) 102-113.

[6] P. Zambrano, J. Torres, P. Flores, How does grooming fit into social engineering? in: Advances in Computer Communication and Computational Sciences, Springer, 2019, pp. 629-639.

[7] N. Morelli, D. Potosky, W. Arthur Jr, N. Tippins, A call for conceptual models of technology in io psychology: An example from technology-based talent assessment, Industrial and Organizational Psychology 10 (4) (2017) 634.

[8] R. F. Muñoz, Harnessing psychology and technology to contribute to making health care a universal human right, Cognitive and Behavioral Practice (2019).

[9] V.-L. Dao, C. Bothorel, P. Lenca, Community detection methods can discover better structural clusters than ground-truth communities, in: 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2017, pp. 395-400.

[10] Z. Amin, Q methodology: A journey into the subjectivity of human mind, Singapore medical journal 41 (8) (2000) 410-414.

[11] K. D. Mitnick, W. L. Simon, S. Wozniak, The art of deception: Controlling the human element of security. 2002, Paperback ISBN 0-471-23712-4 (2006).

[12] F. Mouton, L. Leenen, H. S. Venter, Social engineering attack examples, templates and scenarios, Computers and Security 59 (2016) 186-209. doi:10.1016/j.cose.2016.03.004.

URL http://dx.doi.org/10.1016/j.cose.2016.03.004

[13] Schwaber K., Beedle M. (2001). Agile software development with scrum. Prentice Hall PTR, Upper Saddle River, NJ, USA

[14] C. Rodríguez and R. Dorado, ¿Por qué implementar Scrum?. 2015, pp. 1-20.

[15] Herbsleb, J. D., & Moitra, D. (2001). Global software development. Software, IEEE, 18(2), 16-20

[16] Deemer,P., Benefield, G., Larman, C., Vodde, B.(2008). The Scrum Primer Version 2.0, Scrum Training Institute, available at http://www.scrumprimer.com ,last visited on August 19, 2015.

[17] P. Chapman et al., CRISP-DM 1.0. 2000, pp. 1-76.

[18] L. AlSumait, D. Barbará, C. Domeniconi, On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking, in: 2008 eighth IEEE international conference on data mining, IEEE, 2008, pp. 3-12.

[19] S. Qiao, A. Han, A way to construct evolution model of scientific papers based on the seed document and olda models, in: Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), IEEE, 2013, pp. 900{903.

[20] S. Jameel, W. Lam, L. Bing, Supervised topic models with word order structure for document classification and retrieval learning, Information Retrieval Journal 18 (4) (2015) 283-330.

[21] L. Li, Y. Sun, C. Wang, Semantic augmented topic model over short text, in: 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), IEEE, 2018, pp. 652-656.

[22] D. Peng, D. Guilan, Z. Yong, Contextual-lda: a context coherent latent topic model for mining large corpora, in: 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), IEEE, 2016, pp. 420{425.

[23] D. Liu, Y. Zeng, Y. Luo, H. Pang, X.-H.Wu, Window-based topic model for hdp, in: 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, IEEE, 2019, pp. 70-75.

[24] M. Allahyari, K. Kochut, Automatic topic labeling using ontology-based topic models, in: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), IEEE, 2015, pp. 259-264. [23] H. T. Le, L. N. Pham, D. D. Nguyen, S. V.

Nguyen, A. N. Nguyen, Semantic text alignment based on topic modeling, in: 2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), IEEE, 2016, pp. 67-72.

[26] H. Liu, R. He, H. Wang, B. Wang, Fusing parallel social contexts within flexible order proximity for microblog topic detection, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 875-884.

[27] T. Shi, K. Kang, J. Choo, C. K. Reddy, Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1105-1114.

[28] J. Bai, L. Li, D. Zeng, Activating topic models from a cognitive perspective, in: 2016 IEEE Conference on Intelligence and Security Informatics (ISI), IEEE, 2016, pp. 55-60.

[29] Y. Zuo, J. Zhao, K. Xu, Word network topic model: a simple but general solution for short and imbalanced texts, Knowledge and Information Systems 48 (2) (2016) 379-398.

[30] R. Churchill, L. Singh, C. Kirov, A temporal topic model for noisy mediums, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 42- 53.

[31] Q. Wu, X. Deng, C. Zhang, C. Jiang, Lda-based model for topic evolution mining on text, in: 2011 6th International Conference on Computer Science & Education (ICCSE), IEEE, 2011, pp. 946-949.

[32] S. A. Bahrainian, I. Mele, F. Crestani, Modeling discrete dynamic topics, in: Proceedings of the Symposium on Applied Computing, 2017, pp. 858-865.

[33] B. Liao, W. Wang, C. Jia, Clustering and recommendation of scientific documentation based on the topic model, in: Proceedings of the 2012 International Conference on Information Technology and Software Engineering, Springer, 2013, pp. 629-637.

[34] Z. Xie, L. Jiang, T. Ye, Z. He, Mptm: A topic model for multi-part

documents, in: International Conference on Database Systems for

Advanced Applications, Springer, 2015, pp. 154{168.

[35] A. Trabelsi, O. R. Zaïane, A joint topic viewpoint model for contention analysis, in: International Conference on Applications of Natural Language to Data Bases/Information Systems, Springer, 2014, pp. 114-125.

[36] X. Cheng, X. Yan, Y. Lan, J. Guo, Btm: Topic modeling over short texts, IEEE Transactions on Knowledge and Data Engineering 26 (12) (2014) 2928-2941.

[37] Z. Li, W. Shang, M. Yan, News text classification model based on topic model, in: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), IEEE, 2016, pp. 1-5.

[38] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, H. Yao,

Research on topic detection and tracking for online news texts, IEEE Access 7 (2019) 58407-58418.

[39] S. Sendhilkumar, M. Srivani, G. Mahalakshmi, Generation of word clouds using document topic models, in: 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), IEEE, 2017, pp. 306-308.

[40] A. U. Rehman, Z. Rehman, J. Akram, W. Ali, M. A. Shah, M. Salman, Statistical topic modeling for urdu text articles, in: 2018 24th International Conference on Automation and Computing (ICAC), IEEE, 2018, pp. 1-6.

[41] M. Hasan, M. M. Hossain, A. Ahmed, M. S. Rahman, Topic modelling: A comparison of the performance of latent dirichlet allocation and lda2vec model on bangla newspaper, in: 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE, 2019, pp. 1-5.

[42] X. Wu, C. Li, Short text topic modeling with flexible word patterns, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1-7.

[43] N. Sukhija, M. Tatineni, N. Brown, M. Van Moer, P. Rodriguez, S. Callicott, Topic modeling and visualization for big data in social sciences, in: 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), IEEE, 2016, pp. 1198-1205.

[44] P. Ahmadi, M. Tabandeh, I. Gholampour, Persian text classification based on topic models, in: 2016 24th Iranian Conference on Electrical Engineering (ICEE), IEEE, 2016, pp. 86-91.

[45] R. Pandey, G. O. Mohler, Evaluation of crime topic models: topic coherence vs spatial crime concentration, in: 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), IEEE, 2018, pp. 76-78.

[46] E. Laoh, I. Surjandari, L. R. Febirautami, Indonesians' song lyrics topic modelling using latent dirichlet allocation, in: 2018 5th International Conference on Information Science and Control Engineering (ICISCE), IEEE, 2018, pp. 270-274.

[47] S. ElShal, M. Mathad, J. Simm, J. Davis, Y. Moreau, Topic modeling of biomedical text, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2016, pp. 712-716.

[48] Y. Luo, H. Shi, Using lda2vec topic modeling to identify latent topics in aviation safety reports, in: 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), IEEE, 2019, pp. 518-523.

[49] S. Mifrah, B. L. El Habib, Semantic relationship study between citing and cited scientific articles using topic modeling, in: Proceedings of the 4th International Conference on Big Data and Internet of Things, 2019, pp. 1-8.

[50] C. Zhai, C. Geigle, A tutorial on probabilistic topic models for text data retrieval and analysis, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1395-1398.

[51] B. S. Kumar, V. Ravi, Lda based feature selection for document clustering, in: Proceedings of the 10th Annual ACM India Compute Conference, 2017, pp. 125-130.
[52] O. Mitrofanova, A. Sedova, Topic modelling in parallel and comparable fiction texts (the case study of english and russian prose), in: Proceedings of the International Conference IMS-2017, 2017, pp. 175-180.

[53] L. E. George, L. Birla, A study of topic modeling methods, in: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2018, pp. 109-113.

[54] A. Onan, Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering, IEEE Access 7 (2019) 145614-145633.

[55] L. Sun, J. Chen, J. Li, Y. Peng, Joint topic-opinion model for implicit feature extracting, in: 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), IEEE, 2015, pp. 208-213.

[56] F. Zhang, W. Gao, Y. Fang, B. Zhang, Enhancing short text topic modeling with fasttext embeddings, in: 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), IEEE, 2020, pp. 255-259.

[57] Z. Liu, T. Qin, K.-J. Chen, Y. Li, Collaboratively modeling and embedding of latent topics for short texts, IEEE Access 8 (2020) 99141-99153.

[58] W. Liang, R. Feng, X. Liu, Y. Li, X. Zhang, Gltm: A global and local word embedding-based topic model for short texts, IEEE access 6 (2018) 43612-43621.

[59] L. Li, Y. Sun, X. Han, C. Wang, Research on improve topic representation over short text, in: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), IEEE, 2018, pp. 848-853.

[60] M. Xu, R. Yang, S. Ranshous, S. Li, N. F. Samatova, Leveraging external knowledge for phrase-based topic modeling, in: 2017 Conference on Technologies and Applications of Arti_cial Intelligence (TAAI), IEEE, 2017, pp. 29-32.

[61] X. Li, C. Li, J. Chi, J. Ouyang, Short text topic modeling by exploring original documents, Knowledge and Information Systems 56 (2) (2018) 443-462.

[62] C. Dai, Y. Wang, Q. Wang, Topic model and similarity calculation of text on spark, in: 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, 2017, pp. 15-19.

[63] S. Subramani, V. Sridhar, K. Shetty, A novel approach of neural topic modelling for document clustering, in: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2018, pp. 2169-2173.

[64] B. Jadhav, D. Bhosale, D. Jadhav, Pattern based topic model for data mining, in: 2016 International Conference on Inventive Computation Technologies (ICICT), Vol. 2, IEEE, 2016, pp. 1-6.

[65] W. Hong, X. Zheng, J. Qi, W. Wang, Y. Weng, Project rank: An internet topic evaluation model based on latent dirichlet allocation, in: 2018 13th International Conference on Computer Science & Education (ICCSE), IEEE, 2018, pp. 1-4.

[66] Q. Chen, L. Yao, J. Yang, Short text classification based on lda topic model, in: 2016 International Conference on Audio, Language and Image Processing (ICALIP), IEEE, 2016, pp. 749-753.

[67] T. T. Dao, T. D. Thanh, T. N. Hai, V. H. Ngoc, Building Vietnamese topic modeling based on core terms and applying in text classification, in: 2015 Fifth International Conference on Communication Systems and Network Technologies, IEEE, 2015, pp. 1284-1288.

[68] S. Lee, J. Kim, S.-H. Myaeng, An extension of topic models for text classification: A term weighting approach, in: 2015 International Conference on Big Data and Smart Computing (BIGCOMP), IEEE, 2015, pp. 217-224.

[69] H. Guo, Q. Liang, Z. Li, An improved ad-lda topic model based on weighted gibbs sampling, in: 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE, 2016, pp. 1978-1982.

[70] P. Yang, W. Li, G. Zhao, Language model-driven topic clustering and summarization for news articles, IEEE Access 7 (2019) 185506-185519.

[71] J. Wang, L. Chen, L. Qin, X. Wu, Astm: An attentional segmentation based topic model for short texts, in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018, pp. 577-586.

[72] H. Azarbonyad, M. Dehghani, T. Kenter, M. Marx, J. Kamps, M. De Rijke, Hitr: Hierarchical topic model re-estimation for measuring topical diversity of documents, IEEE Transactions on Knowledge and Data Engineering 31 (11) (2018) 2124-2137.

[73] F. Wang, R. Liu, Y. Zuo, H. Zhang, H. Zhang, J. Wu, Robust word-network topic model for short texts, in: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2016, pp. 852-856.

[74] T. T. Wai, S. S. Aung, Enhanced frequent itemsets based on topic modeling in information filtering, International Journal of Software Innovation (IJSI) 5 (4) (2017) 33-43.

[75] D. Zhao, J. He, J. Liu, An improved lda algorithm for text classification, in: 2014 International Conference on Information Science, Electronics and Electrical Engineering, Vol. 1, IEEE, 2014, pp. 217-221.

[76] H. Kim, Y. Sun, J. Hockenmaier, J. Han, Etm: Entity topic models for mining documents associated with entities, in: 2012 IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 349-358.

[77] S. Li, Y. Zhang, R. Pan, Bi-directional recurrent attentional topic model, ACM Transactions on Knowledge Discovery from Data (TKDD) 14 (6) (2020) 1-30.

[78] D. Jiang, Y. Song, Y. Tong, X.Wu, W. Zhao, Q. Xu, Q. Yang, Federated topic modeling, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 1071-1080.

[79] T. Nguyen, P. Do, Citationlda++ an extension of lda for discovering topics in document network, in: Proceedings of the Ninth International Symposium on Information and Communication Technology, 2018, pp. 31-37.

[80] N. Kawamae, Topic chronicle forest for topic discovery and tracking, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 315-323.

[81] Y. Yang, F. Wang, F. Jiang, S. Jin, J. Xu, A topic model for hierarchical documents, in: 2016 IEEE First International Conference on Data Science in Cyberspace (DSC), IEEE, 2016, pp. 118-126.

[82]V. Y. F. T. Renbo Zhao, «Online Nonnegative Matrix Factorization with Outliers,» pp. 1-28, 2016.

[83] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788-791.

[84] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. 2010, pp. 3-5.

[85] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American society for information science 41 (6) (1990) 391-407.

[86] Anthony Zukas and Robert J. Price. Document categorization using latent semantic indexing. In 2003 Symposium on Document Image Understanding Technology, pages 87–91.

[87] R. Rehurek, Fast and Faster: A Comparison of Two Streamed Matrix Decomposition Algorithms. 2011, pp. 1-7.

[88] J. Yadav and M. Sharma, A Review of K-mean Algorithm. 2013, pp. 1-5.

[89] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993-1022.

[90] M. Hoffman, D. Blei and F. Bach, Online Learning for Latent Dirichlet Allocation. pp. 1-9.

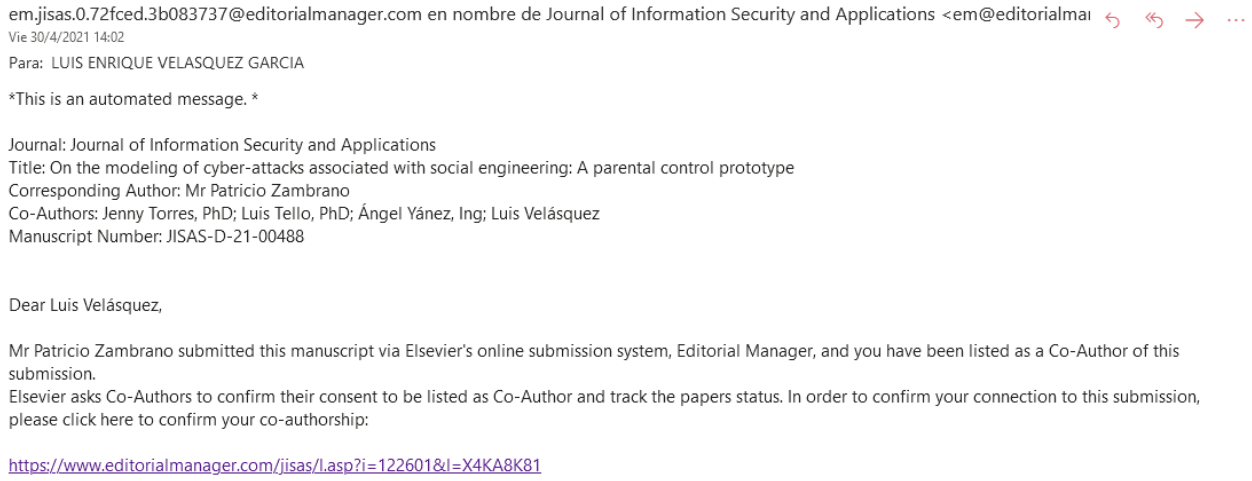
[91] S. Sinharay, Discrete Probability Distributions. International Encyclopedia of Education. 2010, pp. 132-134.

[92] Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei, Hierarchical Dirichlet processes, Journal of the american statistical association 101 (476) (2006) 1566-1581.

[93]C. Wang, J. Paisley and D. M. Blei, Online Variational Inference for the Hierarchical Dirichlet Process, 1st ed. Florida, USA: 14th International Conference on Artificial Intelligence and Statistics, 2011.

9. ANEXOS

9.1. Notificación de Recepción de revista



em.jisas.0.72fced.3b083737@editorialmanager.com en nombre de Journal of Information Security and Applications <em@editorialmai

Vie 30/4/2021 14:02

Para: LUIS ENRIQUE VELASQUEZ GARCIA

*This is an automated message. *

Journal: Journal of Information Security and Applications
Title: On the modeling of cyber-attacks associated with social engineering: A parental control prototype
Corresponding Author: Mr Patricio Zambrano
Co-Authors: Jenny Torres, PhD; Luis Tello, PhD; Ángel Yánez, Ing; Luis Velásquez
Manuscript Number: JISAS-D-21-00488

Dear Luis Velásquez,

Mr Patricio Zambrano submitted this manuscript via Elsevier's online submission system, Editorial Manager, and you have been listed as a Co-Author of this submission.
Elsevier asks Co-Authors to confirm their consent to be listed as Co-Author and track the papers status. In order to confirm your connection to this submission, please click here to confirm your co-authorship:

https://www.editorialmanager.com/jisas/l.asp?i=122601&l=X4KA8K81

Figura 31. Notificación de recepción de revista JISAS

9.2. Articulo Enviado

On the modeling of cyber-attacks associated with social engineering: A parental control prototype

Patricio Zambrano[a,*], Jenny Torres[a], Luis Tello-Oquendo[b,c], Ángel Yánez[a], Luis Velásquez[a]

[a]*Department of Informatics and Computer Science, Escuela Politécnica Nacional, Quito 170525, Ecuador*
[b]*College of Engineering, Universidad Nacional de Chimborazo, Riobamba 060108, Ecuador*
[c]*Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador*

Abstract

Nowadays, the psychological techniques used to harass, intimidate, threaten, steal information are more common due to free access to technological resources and the digitization of communications. Studies related to cybersecurity concerning the use of social engineering techniques are still limited. Several factors such as access to specific databases on cyber-attacks, the unification of scientific criteria that evaluate the nature of the problem, or the absence of accurate proposals that prevent and mitigate this problem could motivate researchers' lack of interest in the field of information security to generate meaningful contributions. This research presents the cyber-attack modeling process defining its stages through topic modeling. Additionally, it presents the tools, techniques, and mechanisms used for the modeling of grooming and bullying. This proposal is supported by a background of research related to the attacks, the modeling of topics, and a functional prototype of parental control that supports the proposed modeling process.

Keywords: APT, bullying, cyber-attack, grooming, pattern behavior, social engineering, topic model

*Corresponding author
Email address: patricio.zambrano@epn.edu.ec (Patricio Zambrano)

1. Introduction

At present, online harassment is a concept that has demonstrated its presence in various communication channels, although it has not been approached globally. Attacks such as grooming, bullying, gender violence, bank fraud, among others, share the same pattern: psychological manipulation to obtain benefits such as personal or financial satisfaction. From the information security perspective, this entity is analyzed as part of the cybersecurity field. The openness and availability of digital communications allow people to express their opinions freely. However, sometimes the content of these contains traits of harassment of all kinds, such as rejection and hatred. This emotional charge reflected in digital textual content requires new proposals supported by computerized mechanisms for its study. It should be noted that the increasing incidence of these cyber-attacks has psychological and economic repercussions. For the first case, a high risk of suicidal behavior has been evidenced in the absence of preventive mechanisms that allow the first signs of an attack to be observed. It is worth mentioning that innovative technological solutions are not enough to combat this type of attack. Collective and collaborative work between the family, police, and judicial spheres is necessary to generate more effective defenses against these cyber attacks.

In today's environment, the virtual communications usability trend is increasingly evident. A fundamental challenge for researchers is to propose mechanisms that allow managing the knowledge inherent in virtual communications. For this purpose, the use of natural language processing (NLP) techniques is necessary. One of the most widely used NLP techniques is topic detection and modeling. Several algorithm proposals, such as pLSA, LDA, HDP, TDM, NMF, among many others, determine connecting aspects (topics) between the words that make up digital texts. There are free access and licensed libraries that have been previously analyzed and tested in the scientific field for the implementation of these algorithms. In the classification of digital information are those texts considered "short texts"; these come from comments, blogs, chats, among others. Short texts do not have a defined structure; however, they allow defining common patterns in large quantities.

The focus of this research is to propose a modeling process of an attack related to psychological violence. Large amounts of texts related to online harassment were used for this purpose. They went through debugging and modeling processes to find intrinsic patterns that allow modeling

cyber-attacks such as grooming and bullying from information security. The results with a high degree of precision supported the feasibility of applying the process to new attacks related to the study of online harassment in short texts (chats).

The main contributions of this study are summarized as follows.

- Formalization of a process applied in the analysis and study of cyber attacks related to bullying and grooming;
- Propose current techniques for obtaining and processing data related to instant messaging;
- Propose alternative models of topics in short word processing;
- Point out functional tools that allow establishing lexical meanings to the topics obtained in the modeling;
- Show the architecture and operation of a parental control prototype based on the models proposed in the scientific field.

The remainder of this paper is organized as follows. Section 2 presents a background that allows clarifying the research problem. Section 3 describes the process carried out to model Grooming and Bullying, respectively. In Section 4, techniques, tools, models and other aspects that were used in each phase of the process are described. Section 5 describes the functioning of a functional parental control prototype based on the proposed models. Section 6 answers the research questions posed. Finally, Section VII draws conclusions and presents future work.

2. Problem Understanding

2.1. Background

2.1.1. Cyber-attacks vs. Assets

Today, one of the most valuable assets of organizations is information, and various strategies or controls are used to prevent it from being affected by unwanted attacks. However, this concept is exclusively aligned with information security and not with cybersecurity. The authors in [1] explore the different definitions of information security and cybersecurity. They describe that, although the concept of cybersecurity is aligned with that of information security, their coverage is different. Information security is

the protection of information considered as an asset, which is susceptible to potential attacks. On the other hand, cybersecurity covers the protection of cyberspace itself and the protection of each one of the elements that generate information (assets).

Currently, there are cyber attacks that guide their attention to the harassment of people; cases such as cyberbullying and grooming are present in this spectrum of unconventional attacks. In various studies [1], [2], [3], the increasing trend of these attacks is demonstrated. In their eagerness to compromise their victims' emotional and psychological stability, the aggressors use intimidation and harassment strategies so that their victims feel shame and normalize violence. Physical damage is achieved in the most advanced cases, generating severe, negative, and negative impacts irreversible in victims. All these processes are carried out with technological tools. Given the rise of cyberbullying, scientists and governments are developing proposals that progressively address and mitigate these cybersecurity attacks. It should be noted that being a victim of cyberbullying in cyberspace does not establish a loss of confidentiality, integrity, or availability of a type of tangible information. Instead, the target of these attacks is the victims and their emotional stability. Consequently, people and their physical-emotional stability must be part of the concept of assets within the field of cybersecurity.

2.1.2. Psychology vs. Social Engineering

In the area of computer security within Computer Science, there is Social Engineering. This field of study is responsible for studying and establishing the techniques or practices of psychological manipulation that attackers use to obtain confidential information from computer resources or people. The increase in attacks concerning tactics related to social engineering in cyberspace has led to research linking the fields of psychology and technology. Such is the case, in [4, 5] the authors generalize the basic principles of social engineering based on the behavior and susceptibility of victims. In these investigations, the authors adapted knowledge of experimental psychology to identify factors that increase the probability of success of a social engineer against a human victim.

2.1.3. Technology vs. Psychology

Solving social demands related to cognitive processes and interpersonal relationships is one of the objectives of scientists in the field of psychology. The application of information technologies has made it possible to improve

the processes of generating new knowledge in this area. For this reason, technology has become a fundamental tool that psychology uses for the benefit of people through different techniques of analysis, evaluation, and modification of human behavior [6, 7].

In recent years, the use of social networks shows the growing interrelation of information technologies and the psychological effect on people. This trend allows specialists in the psychological field to study human behavior and mental health concerning the use of technologies. As a case study in [8], the authors compiled research related to the use of social networks and the interaction between individuals in them; however, they point out that the data is still limited. This compendium of research analyzes and identifies latent meta-groups of online communities with and without mental health-related conditions where aspects such as depression and autism are part of the analysis of results.

With these antecedents, it is observed that psychology makes use of information technologies to develop new knowledge. However, the field of psychology has also made it possible to strengthen concepts, criteria, and theories that have been born from computer science. Within this field, there is much digital information related to social behavior. This information requires formal processes that allow the establishment of common patterns aligned with the field of psychology. For this reason, for several years, various proposals have been analyzed to standardize the behavior of online bullies (grooming), supporting their results with psychological concepts [2, 3].

2.1.4. Subjectivity in the scientific results

In studies related to online bullying, the psychological component is observed. This component allows the establishment of the different phases that are part of the operation of the attack; however, it is believed that this human component has not allowed establishing a standard procedure that allows modeling other attacks with similar characteristics due to its high level of subjectivity. In [9] this aspect is already evidenced as an obstacle to the research validation since the peer review is based on objective and not subjective data. On the other hand, it is stated that there are research areas where the lack of availability of adequate research tools and the difficulties associated with quantifying subjective data do not allow establishing conclusive data [10]. In our case study, it was established that there is a set of tools based on statistical algorithms that allow reducing this subjectivity [2, 3]. With the use of topic modeling, relevant information

can be extracted through categories or topics from large volumes of digital information, which in our case are instant messaging chats.

2.1.5. Advanced persistent threat modeling - APT

In the technical description, [2], advanced persistent attacks have a wide range of proposals. Models such as Lockheed, LogRhythm, Mandiant, Dell Secureworks, SDAPT, BSI, and Lancaster vary their phases. This difference in criteria is due to the lack of agreements or standardization to model APT attacks. These models base their phases on experiential aspects during the analysis of the various attacks. On the other hand, from the social engineering perspective, alternative models to APTs have been proposed, such as the case of Kevin Mitnick's 4 phases (information gathering, relationship development, relationship exploitation, and execution to achieve the objective) [11], and Mounton (attack formulation, information gathering, preparation, relationship development, relationship exploitation and debrief) [12]. To standardize the criteria for modeling attacks related to social engineering related to APT concepts, Zambrano et al. [2], in their proposals, relates the phases of the models with the lexical categories determined in each evaluated attack, grooming, and bullying cases.

2.2. Review of Topic Modeling Applications

Topic modeling is an unsupervised machine learning technique, which is part of the concept of artificial intelligence and specifically natural language processing (NLP). This has the particularity of analyzing large sets of documents, detecting common patterns between the words that make them up, and grouping them into topics representing them. Factors such as coherence and perplexity allow us to study these groupings of words. In the field of text analysis, inspired by artificial intelligence, there is a wide range of methods or algorithms that process information, such as topic analysis.

Currently, several articles make comparisons and describe the applications of different models that are part of the modeling of topics, see Table 1. These surveys analyze the current usage trends and structural aspects of the models [13]. In [14] the behavior of several models with specific data is analyzed, called by the authors "short texts." The results discuss the reliability of grouping this type of poorly structured information in the resulting topics. On the other hand, the authors in [15] classify the existing topic models into two categories: topic models (e.g., LSA - LDA) and topic evolution models, which model topics considering a time factor (e.g., TOT - DTM).

In [16] a variation to LDA is proposed, called Online LDA (OLDA), whose approach provides an efficient means to track topics over time; it is essential to highlight that it has detection characteristics of emerging topics in time. This model was evaluated qualitatively and quantitatively. Another application example is given in [17] where they use the OLDA features to group scientific articles published over time.

One of the limitations of the results of the modeling of topics (word bags) is the lack of description of a lexical meaning. These models have an algorithmic mechanism that groups words according to the inherent grammatical aspects of each document. In previous research (Bullying and Grooming), the authors propose using the EMPATH system or the LIWC dictionary to establish meanings or lexical descriptions for each bag of words resulting from the applied model.

2.3. Research questions

Formalizing the cyber-attack modeling process with concepts related to traditional information security will allow researchers to support future research related to social engineering, topic modeling, and analysis of social patterns. To achieve these goals, we formulated the following research questions:

RQ1: Is the proposed modeling process applicable to other cyber-attacks that evidence social engineering tactics?

RQ2: Which topic model presents better results in the evaluation of short texts?

RQ3: Can the models determined by the proposed process be implemented in functional systems today?

3. Description of the modeling process of cyber-attacks related to Social Engineering

Figure 1 describes the process to model the cyber-attacks. It is based on the following four stages:

1. Attack selection, see Figure 2- This sub-process establishes that: any investigation related to cyber-attacks must go through a process of analysis of the literature. This analysis verifies the existence of previous

Table 1: Topic Modeling - Literature review

Literature review	*Reference*
Topic models using semantic context to improve document classification	[18], [19], [20], [21], [22], [23], [24], [25], [26]
Topic modeling in short texts based on word co-occurrence networks	[27]
Topic models that identify topics as they emerge over time	[28], [16], [29], [17], [30]
Incorporation of topic features to improve document grouping accuracy	[31]
Multi-part topic model improves information retrieval and document classification performance	[32]
Topic models that automatically detect recurring patterns of expressions	[33]
Modeling of topics using word co-occurrence patterns.	[34]
Text classification using LDA or LDA variants	[35], [36], [37], [38], [39], [40], [41], [42], [43] [44], [45], [46], [47], [48], [49], [50], [51]
Word embedding for improved topic modeling	[52], [53], [54] [55], [56], [57]
Incorporation of phrases into topics models to add coherence	[58], [59]
Parallel computing as topic modeling	[60]
Topic modeling with artificial neural networks.	[61]
Topic modeling for short texts using word patterns	[40], [62]
LDA model applying distances with probability distributions	[63], [64], [65]
LDA model applying weights to words in the sample	[66], [67]
Topic models that obtain contextualized summaries	[68]
Topic models that perform an analysis taking into account human attention	[69]
Topic models that filter and reduce generality and impurities in documents	[70], [71], [72]
Topic model adding a category distribution parameter to LDA	[73]
Topic model that takes into account entities that are present in documents	[74]
Topic modeling that exploits the sequential order of sentences and the relationship between successive sentences	[75]
Topic model focused on solving data scarcity and improving data security	[76]
Topic modeling with focus on citations and document titles	[77]
Topic models implementing hierarchical topic trees	[78]
Topic model linking short texts to long documents	[79]

studies or proposals that propose data download procedures or specific

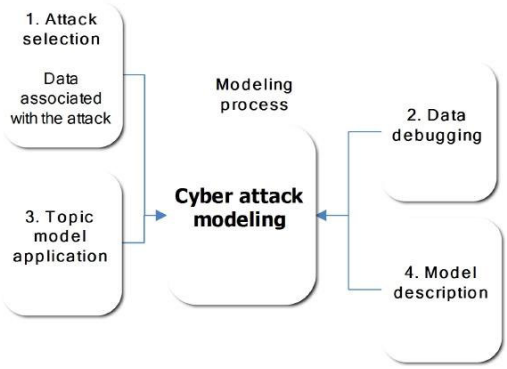


Figure 1: Cyber-attack modeling processing

modeling. In this sub-process, the researchers must collect texts, chats, or any information related to instant messaging. These data must demonstrate the use of psychological manipulation techniques or social pressure by the attackers for the investigation to be viable. The data collected must go through previous debugging processes.

2. Data debugging, see Figure 3.- In this sub-process, researchers must rely on standardized data mining models. These have recommendations for the treatment and analysis of large information bases. Models such as CRISP-DM (Cross Industry Standard Process for Data Mining) or SEMMA (Sample, Explore, Modify, Model, and Assess) specify the tasks to be carried out in each phase described by the information processing process.
3. Topic model application, see Figure 4.- In the area of natural language processing, within the Artificial Intelligence (AI) models, various algorithms are described that allow modeling of topics in such a way that unsupervised large volumes of information. Models such as LSI (latent semantic indexing), LDA (Latent Dirichlet Allocation), HDP (Hierarchical Dirichlet Process), and NMF (Non-negative matrix factorization), among others, are available for use and application in Python and Matlab. It is worth mentioning that any investigation

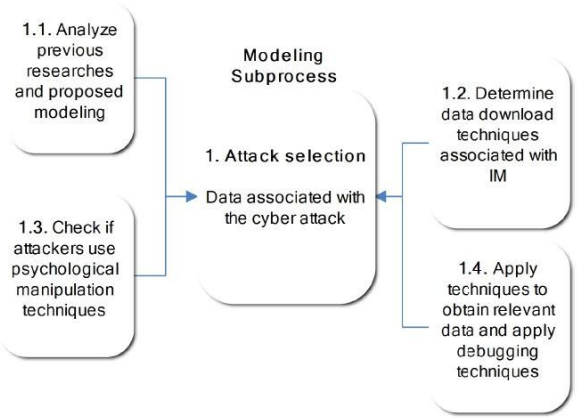


Figure 2: Subprocess - Attack selection

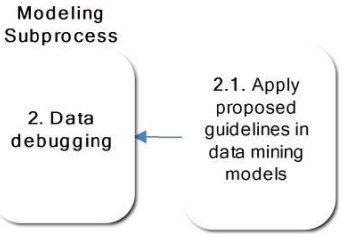


Figure 3: Subprocess - Data debugging

must go through a rigorous analysis of these models, in contrast to the data, so that the results obtained have a higher level of accuracy and a lower number of word groupings (topics). It is important to take into account that these models have certain peculiarities that associate them. On the one hand, there is the type of data that must be entered into these models. These must previously be refined and lemmatized for modeling. This modeling will be associated with the assignment

of a standard number of topics that the researcher must define for each cyber attack. Variables such as perplexity, computational cost, or coherence will determine a specific number of word groupings to refine the final models.

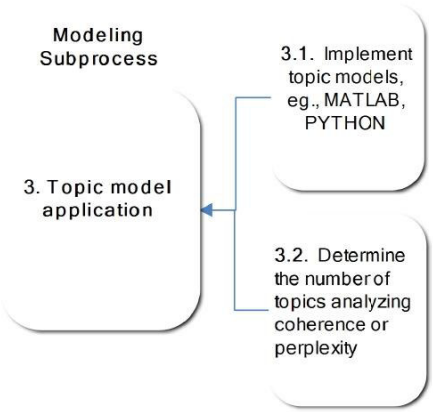


Figure 4: Subprocess - Topic model application

4. Model description, see Figure 5 - Once the word groupings are obtained, they do not have an explicit meaning. With the use of software tools or manual dictionary-based categorization, lexical categories and communication intentions are defined. This information, in the future, will be related to APT model stages and their concepts.

4. Application of the modeling process to cyber attacks

Currently, some cyber-attacks make use of psychological manipulation. Attacks such as grooming and bullying have evolved, from a technical point of view, in recent years. On the other hand, it has been observed that the research associated with studying these phenomena has not grown in the same way for their identification and mitigation. The causes that prevent this growth may be due to the confidentiality to publish the results of

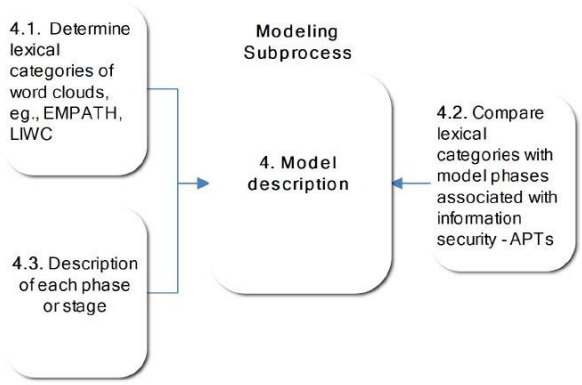


Figure 5: Subprocess - Model description

the investigations, access to data related to the attacks, legal regulations of each country, and even lack of agreements to standardize the modeling to be applied in future proposals. In [3], the authors consider that these aspects can be addressed to increase the proposals in the area. Regarding standardization, it has been determined that cyber-attacks can be modeled from the point of view of information security. Data can be collected from various freely accessible sources without this violating legal regulations or threatening the privacy of minors. Under these premises, the applicability of the proposed modeling process to the attacks of Bullying and Grooming will be justified.

4.1. Attack selection

Grooming and Bullying cyber-attacks were analyzed and developed from the perspective of Social Engineering, since the use of psychological manipulation techniques by the attackers was evidenced. The detailed description of these attacks can be found in [2, 3]. This section describes the procedural application of the modeling process to the case studies, specific details can be seen in Table 2.

Table 2: Application of CRISP-DM to Grooming and Bullying

	Grooming	Bullying
Previous analysis of the literature	Yes	Yes
Download mechanisms were taken into account to obtain the data	No	Yes (partially)
Data source	perverted-justice.com	twitter.com hatebase.org pacerteensagainstbullying.org pacerkidsagainstbullying.org
Type of data	Chats - HTML	Chats - HTML Experiences - HTML
Amount of data	128171 chat lines 100 conversations	250000 attacker-related tweets 3035 victim experiences
Download mechanism	Manual downloads Script developed in python	Scrapy Script developed in python
Analysis of data	Only from attackers (tweets) Only victims (experiences)	

4.2. Data debugging

In this section, the CRISP-DM model is adapted (see Figure 6) to Grooming and Bullying case studies, respectively.

4.3. Topic model application

In the Grooming and Bullying case studies, LDA topic modeling was applied, see Table 3. Various studies, in their comparisons, showed that this model provides better results in the analysis of short texts. This supported the application of this model to the cases described [34, 54, 64].

Regarding the scope of the applicability of the topical models, the data behavior with three different models to LDA will be demonstrated below. The implementation of these models was carried out with Python, and the databases analyzed in predecessor investigations were used for this purpose.

1. LSI is a topic model proposed by Deerwester et al. [80]. It is part of a set of natural language processing techniques, particularly distributional semantics. LSI is an indexing and retrieval method that uses a mathematical technique called SVD singular value decomposition, making it possible to determine patterns in a set of words in a collection of texts. SVD supports its operation in classical techniques

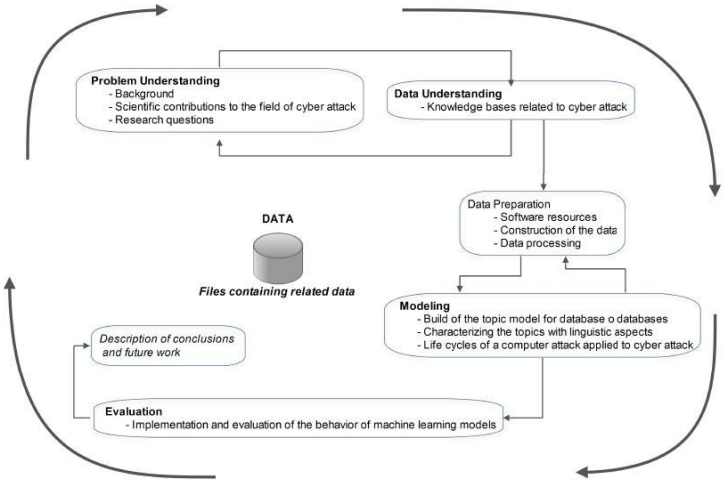


Figure 6: CRISP-DM model applied to cyber attacks

Table 3: Topics defined in the analysis tools

	Grooming	Bullying
Type of data	Short Text - Chats	Medium Text - Experiences
Applied model	LDA	LDA
Tools	Matlab Text Analytics	Python Gensim
Aspect used to define the number of topics	Perplexity	Perplexity Coherence
Number of topics defined	6	4

of second-order eigenvectors and aims to analyze large volumes of information and look for components that are not correlated. LSI

has a probabilistic variant developed by Thomas Hofmann known as probabilistic latent semantic analysis, PLSI, or PLSA. This model started the development of LDA, which is the Bayesian version of PLSI.

2. LDA model developed by Blei et al. [81]. It classifies texts from various documents into topics. LDA models its topics with Dirichlet distributions. In this process, word-topic data arrays are established using the Gibbs sampling algorithm of Dirichlet distributions. Each sample will give the probability of each word by topic.
3. NMF is a model proposed by Lee et al. [82]. This model is made up of 2 methods, one performs the dimension reduction functions, and the other performs a factor analysis. The factoring process allows a weighting based on the semantics between the words. Applying an optimization process, the model monitors that within its data, there are no negative values. Finally, a matrix of weighted terms is obtained and grouped in their respective topics.
4. HDP is a mixed model that performs an unsupervised analysis of pooled data. This model was proposed by Teh [83]. Unlike LDA, HDP infers the number of subjects from the data. With the use of variational Bayes coordinate ascending algorithms, HDP manages to optimize the processed data stochastically. Unlike its predecessor LDA, HDP is unlimited in defining the number of topics and learns from your data without the need to pre-specify the number of topics.

Figure 7 schematizes the behavior of the data (related to grooming) applied to the proposed models LSA, LDA, NMF, and HDP. The models applied the six topics, defined, and justified by the predecessor investigations. The results show that the LSI and HDP models generate greater dispersion between the coherence values obtained in each topic. Discarding these two models from the analysis, the LDA model and its HDP variant concentrate the processed data better.

Regarding Bullying, see Figure 8, the behavior of the models is similar to the case of Grooming with the difference that it has four topics. Similarly, the LSI and NMF models are discarded since the coherence values are very dispersed concerning LDA results and its predecessor HDP. The HDP model, when applied in the 2 case studies, offers better results in the classification of words by topic. It is worth mentioning that HPD is an improved model to the LDA. With this premise, it is concluded that the LDA model or its variants (HDP) are adequately adapted to short texts, as shown by the

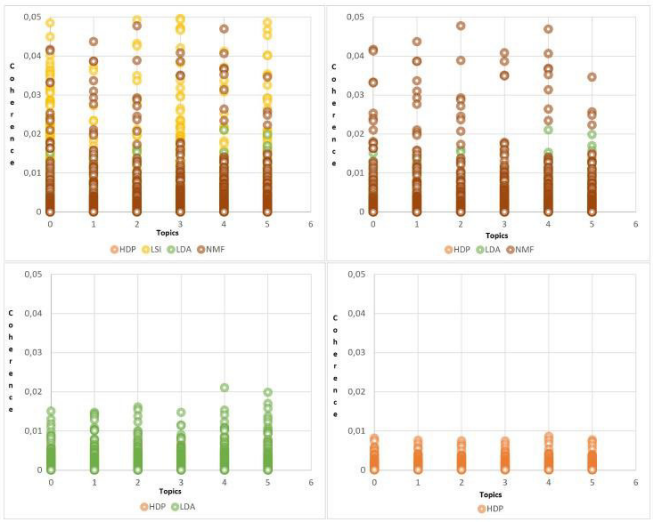


Figure 7: Graphical comparison of the behavior of the coherence values in 4 topic models - Grooming

research described above.

4.4. Model description

With the results obtained when applying LDA, we proceeded to define the lexical meaning of each group, considering the communicational intention of the attacker. Table 4, concerning Grooming, indicates that the process to determine the lexical meaning of each topic had several stages.

In the process of describing each topic lexically, the categories proposed in LIWC were used to describe the communicational intentions of Grooming. It should be noted that this tool was developed for the cognitive and emotional evaluation of texts through a series of psychological and structural categories. This program, used in the field of psychiatry, analyzes texts, word by word, in a classification of different linguistic variables, which include standard language categories (articles, prepositions, pronouns, among others), psychological processes (categories of positive emotions and negative,

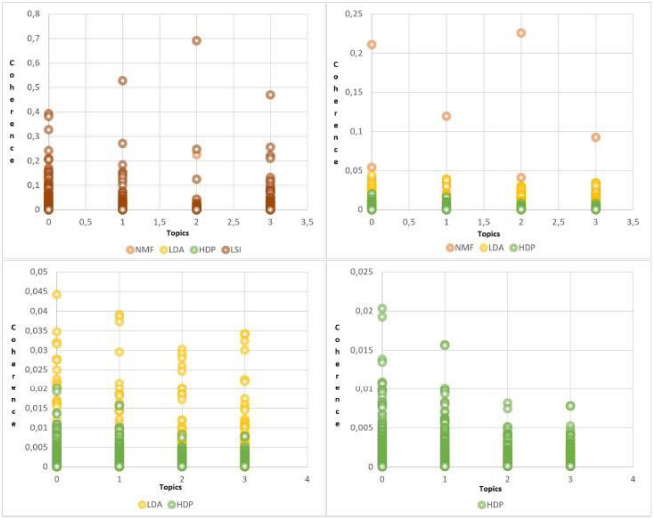


Figure 8: Graphical comparison of the behavior of the coherence values in 4 topic models - Bullying

Table 4: Linguistic tools and security models categorization

	Grooming	*Bullying*
Number of topics	6	4
Lexical Categories	LIWC Manual classification	Empath Automatic classification
Analysis of communicational intentions	Manual	Manual
Selected topics	Gethering informatión (SDAPT) Gaining Access (SDAPT) Lateral Movement (LogRhythm) Escalating Privileges (Mandiant) Execution (Mitnick) Debrief (Mounton)	Development of relationship (Mitnick) Preparing/Distracting Attack (BSI) Exploit the Relationship (Mounton) Debrief (Mounton)

cognitive variables), words related to space-temporal relativity, verb tenses, and traditional dimensions of content organized hierarchically [84].

On the other hand, EMPATH is a tool that can generate and validate new lexical categories on demand from a small set of seed terms. EMPATH sketches connotations between words and phrases using deep learning and neural integration in more than 1.8 billion words of modern fiction. EMPATH can analyze text in 200 pre-validated and built-in categories. In [85] they demonstrate that the categories validated by humans and based on data from EMPATH are highly correlated ($r = 0.906$) with similar categories from LIWC.

With the use of the tools described, we proceeded to define a lexical characteristic for each word grouping and compare these in terms of communicational intentions with the stages of the models proposed in information security as follows:

1. Manual comparison of the words of each topic with the dictionary of the LIWC tool.
2. The dictionary classifies certain words into categories, and these were selected based on the data.
3. Once the categories that most contextualized the type of grouped data had been selected, we proceeded to analyze the communicational intentions that defined these categories.
4. With this analysis, the communicational intentions of each topic were compared with the proposed phases of the different models applied to APTs.

Unlike the Grooming study, the Bullying data was analyzed by the WEB EMPATH tool, which automatically classified the words into lexical categories. With these categories, item 3 of the Grooming study was continued to determine the phases from the perspective of information security.

5. Implementation of a parental control prototype

The application of AI algorithms in the development of software systems has made it possible to create systems with extensive capabilities for identifying patterns with high precision, repetitions of activities, and in some instances, they can make decisions intelligently.

Regarding the protection of minors, parental control systems are developed to block unauthorized access to minors. However, they lack

real-time data analysis related to online bullying. For this reason, an innovative parental control prototype has been developed that analyzes instant messaging data and categorizes it according to how aggressive these texts contain. For the development of this prototype, Scrum was used; it is a flexible and orderly framework in the development of this system.

5.1. Tools for prototype development

Table 5 describes each of the software tools used to develop the prototype. It is worth mentioning that the general platform was developed under the Python language. Its design was conceptualized in a modular way and each component is susceptible to modification without there being a general change to the system. Applying best practice recommendations in software development, the prototype is developed in collaborative applications, versioning and with security approaches.

5.2. Prototype architecture

The prototype is based on the theoretical aspects described in the research (number of topics and application of learning models for classification). The implementation of this system includes two architectures, a WEB type under the REST figure and another Client-server.

Figure 9 outlines the modular operation of the system, which has the following components:

1. WEB server
2. Learning module
3. Classification module
4. Client software
5. Text message capture API
6. Setting mechanism for permissiveness level
7. Communications blocking mechanism
8. Parental alert mechanism

5.3. Component operation

The WEB parent server requires pre-processed and labeled data (in CSV format) to create a machine learning mechanism according to the topic to which they belong. With this mechanism, the system will classify and label future data coming from the client software.

Table 5: Software development tools applied to the parental control prototype

Tool	*Description*	*Prototype's component*
Git	Code version control system	Planning, development, and versioning
Gitlab	Web service for code versioning, centralized digital repository manager	Version repository
Python	Programming language	Server and desktop application
LucidChart	Real-time collaborative diagram learning tool	Diagram of architecture and system flows
RESTful services	Interface that allows communication between systems over HTTP	Communication between message interceptor and server
Google Chrome API	Set of functions, objects and libraries that allow access to information from websites opened by Google Chrome	Facebook message interceptor, implemented in the client
Flask	Micro framework developed in Python that allows the rapid development of web applications	Client app
Scikit	Library specialized in algorithms and artificial intelligence modules developed in Python	Server
RabbitMQ	Message Broker that enables asynchronous communication through message queue	Message management (FIFO queues) between client/server
Visual Studio Code	Code Editor, developed by Microsoft	Message interceptor between the server and the client app
Postman	Tool that allows you to test web services	General tests

The client software has a mechanism for extracting chats from the Facebook social network in real-time (API). This will oversee sending the chats in real-time to the parent server to classify them in the programmed topics employing labels automatically, see Figure 10. The client module will receive the tagged text and will compare the tag number with the level of the permissiveness of the system (topic configured by the user), see Figures 11 and 12. If this number exceeds the allowed topic, the system will immediately block communication with the origin WEB site (Facebook).

After a comparative analysis of labels, the texts considered malicious will

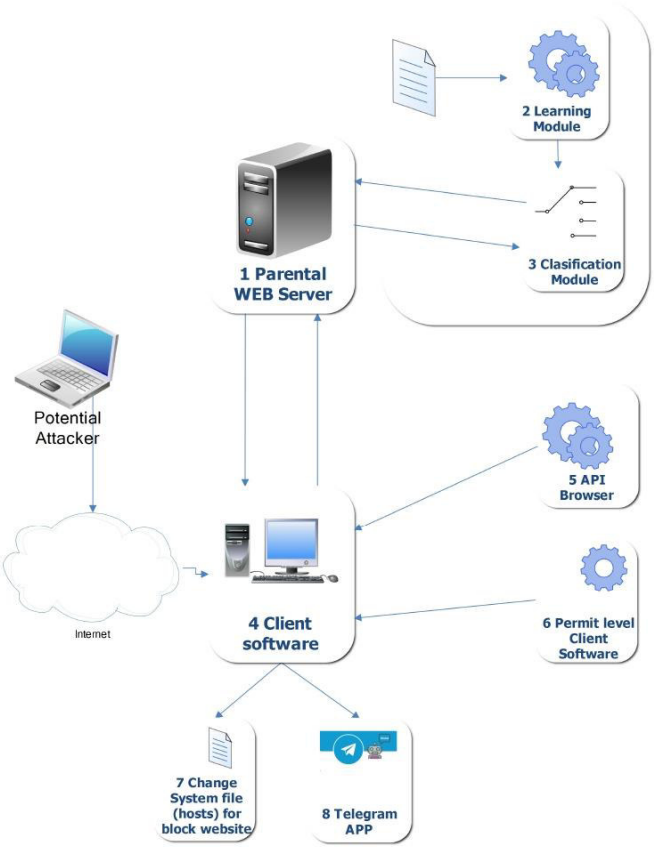


Figure 9: Parental control prototype architecture

be sent via text message (telegram) to the minor's parents, see Figure 13, notifying the level of aggressiveness in which the attack is located. This message will contain the attacker's username, time, and date of the event.

Figure 10: Modeling process

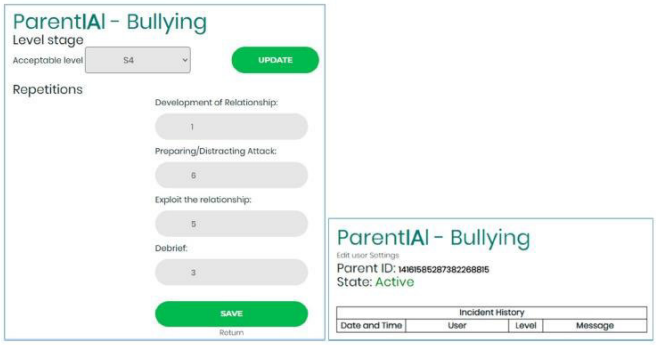


Figure 11: Modeling process

6. Answering the research questions

6.0.1. Is the proposed modeling process applicable to other cyber-attacks that evidence social engineering tactics?

The basic process started with the Grooming research. In this attack, criteria and modeling parameters were defined that were applied to a new Bullying case study. This attack was analyzed with two types of databases, unlike Bullying, and implemented in Python. The results of this study strengthened the modeling process of these attacks. For this reason, it is considered that the proposed process can model new attacks that are related to psychological manipulation. It should be noted that the process does not standardize models and tools.

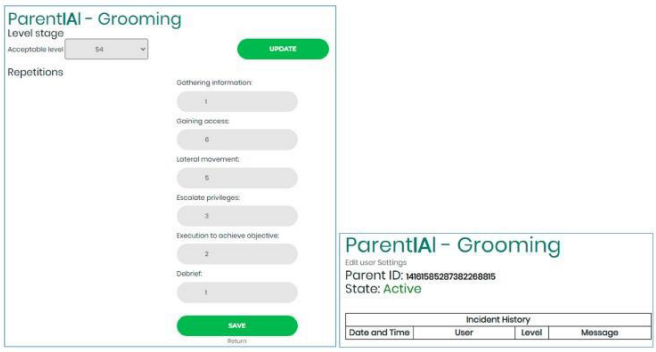


Figure 12: Modeling process

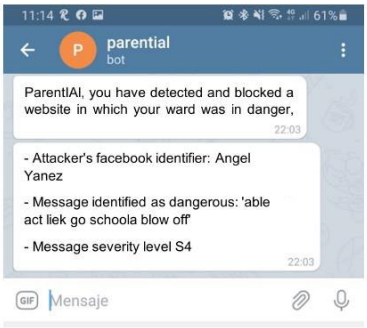


Figure 13: Modeling process

6.0.2. Which topic model presents better results in the evaluation of short texts?

As evidenced in Section 4, the LDA model and its HDP variant currently offer better short text processing results. These texts come from instant messengers and comment blogs. It should be noted that the databases used were purified under the same cleaning and normalization process.

6.0.3. Can the models determined by the proposed process be implemented in functional systems today?

In Section 5, a system based on the proposed models was developed. This modular prototype allows the modification of the databases, learning models, and classification. The prototype represents the theoretical models proposed, and its operation indicates that it applies to new cyber-attacks. For this reason, currently, there are computational resources for the development of these AI-based tools.

7. Conclusions and Future Work

In real-time, the detection and mitigation mechanisms that counteract cyber-attacks, where the objective is human destabilization, are still under development. As evidenced in the literature review conducted, efforts have been made to eliminate the effects of these phenomena; however, they persist and manifest themselves incrementally as they improve their attack and evasion techniques. Studying these attacks directed at the human psyche from the point of view of information security, such as social engineering, makes it possible to link proposed techniques, methodologies, and architectures to future cybersecurity and conventional security projects. With this connection, it would be possible to standardize knowledge and processes on cybersecurity, thus avoiding incomplete and scattered proposals. The results obtained in this research consolidate the modeling process carried out on Grooming and Bullying and enable the possibility of applying it to future social engineering attacks not yet defined.

We have combined the concepts of cybersecurity, social engineering, and traditional security to understand that these cyber-attacks are part of the same line of study. With theme modeling, different stages or topics that model the attacks were determined; however, these topics require human intervention to define a lexical concept or communicational intention. This will support investigations related to the identification of patterns of malicious behavior online.

In the experimentation phase, a typical attackers' pattern was determined in the processing of information related to the experiences of the victims. The statistical algorithm LDA and its predecessor HDP presented the best results in analyzing and distributing the information, delivering four groupings of words for Bullying and 6 for Grooming. By themselves, these classifications do not describe linguistic aspects; therefore, linguistic software was used to

define the communicational intentions of each stage. With this knowledge, the stages of the models assigned to information security were correlated, and the definitive model of our case study was defined.

References

[1] R. Von Solms, J. Van Niekerk, From information security to cyber security, computers & security 38 (2013) 97–102.

[2] P. Zambrano, J. Torres, L. Tello-Oquendo, R. Jácome, M. E. Benalcazar, R. Andrade, W. Fuertes, Technical mapping of the grooming anatomy using machine learning paradigms: An information security approach, IEEE Access 7 (2019) 142129–142146.

[3] P. Zambrano, J. Torres, Á. Yánez, A. Macas, L. Tello-Oquendo, Understanding cyberbullying as an information security attack—life cycle modeling, Annals of Telecommunications (2020) 1–19.

[4] J. M. Hatfield, Social engineering in cybersecurity: The evolution of a concept, Computers & Security 73 (2018) 102–113.

[5] P. Zambrano, J. Torres, P. Flores, How does grooming fit into social engineering?, in: Advances in Computer Communication and Computational Sciences, Springer, 2019, pp. 629–639.

[6] N. Morelli, D. Potosky, W. Arthur Jr, N. Tippins, A call for conceptual models of technology in io psychology: An example from technology-based talent assessment, Industrial and Organizational Psychology 10 (4) (2017) 634.

[7] R. F. Muñoz, Harnessing psychology and technology to contribute to making health care a universal human right, Cognitive and Behavioral Practice (2019).

[8] V.-L. Dao, C. Bothorel, P. Lenca, Community detection methods can discover better structural clusters than ground-truth communities, in: 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2017, pp. 395–400.

[9] Z. Amin, Q methodology: A journey into the subjectivity of human mind, Singapore medical journal 41 (8) (2000) 410–414.

[10] A. F. Fırat, N. Dholakia, From consumer to construer: Travels in human subjectivity, Journal of Consumer Culture 17 (3) (2017) 504–522.

[11] K. D. Mitnick, W. L. Simon, S. Wozniak, The art of deception: Controlling the human element of security. 2002, Paperback ISBN 0-471-23712-4 (2006).

[12] F. Mouton, L. Leenen, H. S. Venter, Social engineering attack examples, templates and scenarios, Computers and Security 59 (2016) 186–209. doi:10.1016/j.cose.2016.03.004.
URL http://dx.doi.org/10.1016/j.cose.2016.03.004

[13] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, Multimedia Tools and Applications 78 (11) (2019) 15169–15211.

[14] R. Albalawi, T. H. Yeap, M. Benyoucef, Using topic modeling methods for short-text data: A comparative analysis, Frontiers in Artificial Intelligence 3 (2020) 42.

[15] R. Alghamdi, K. Alfalqi, A survey of topic modeling in text mining, Int. J. Adv. Comput. Sci. Appl.(IJACSA) 6 (1) (2015).

[16] L. AlSumait, D. Barbará, C. Domeniconi, On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking, in: 2008 eighth IEEE international conference on data mining, IEEE, 2008, pp. 3–12.

[17] S. Qiao, A. Han, A way to construct evolution model of scientific papers based on the seed document and olda models, in: Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), IEEE, 2013, pp. 900–903.

[18] S. Jameel, W. Lam, L. Bing, Supervised topic models with word order structure for document classification and retrieval learning, Information Retrieval Journal 18 (4) (2015) 283–330.

[19] L. Li, Y. Sun, C. Wang, Semantic augmented topic model over short text, in: 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), IEEE, 2018, pp. 652–656.

[20] D. Peng, D. Guilan, Z. Yong, Contextual-lda: a context coherent latent topic model for mining large corpora, in: 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), IEEE, 2016, pp. 420–425.

[21] D. Liu, Y. Zeng, Y. Luo, H. Pang, X.-H. Wu, Window-based topic model for hdp, in: 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, IEEE, 2019, pp. 70–75.

[22] M. Allahyari, K. Kochut, Automatic topic labeling using ontology-based topic models, in: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), IEEE, 2015, pp. 259–264.

[23] H. T. Le, L. N. Pham, D. D. Nguyen, S. V. Nguyen, A. N. Nguyen, Semantic text alignment based on topic modeling, in: 2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), IEEE, 2016, pp. 67–72.

[24] H. Liu, R. He, H. Wang, B. Wang, Fusing parallel social contexts within flexible-order proximity for microblog topic detection, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 875–884.

[25] T. Shi, K. Kang, J. Choo, C. K. Reddy, Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1105–1114.

[26] J. Bai, L. Li, D. Zeng, Activating topic models from a cognitive perspective, in: 2016 IEEE Conference on Intelligence and Security Informatics (ISI), IEEE, 2016, pp. 55–60.

[27] Y. Zuo, J. Zhao, K. Xu, Word network topic model: a simple but general solution for short and imbalanced texts, Knowledge and Information Systems 48 (2) (2016) 379–398.

[28] R. Churchill, L. Singh, C. Kirov, A temporal topic model for noisy mediums, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 42–53.

[29] Q. Wu, X. Deng, C. Zhang, C. Jiang, Lda-based model for topic evolution mining on text, in: 2011 6th International Conference on Computer Science & Education (ICCSE), IEEE, 2011, pp. 946–949.

[30] S. A. Bahrainian, I. Mele, F. Crestani, Modeling discrete dynamic topics, in: Proceedings of the Symposium on Applied Computing, 2017, pp. 858–865.

[31] B. Liao, W. Wang, C. Jia, Clustering and recommendation of scientific documentation based on the topic model, in: Proceedings of the 2012 International Conference on Information Technology and Software Engineering, Springer, 2013, pp. 629–637.

[32] Z. Xie, L. Jiang, T. Ye, Z. He, Mptm: A topic model for multi-part documents, in: International Conference on Database Systems for Advanced Applications, Springer, 2015, pp. 154–168.

[33] A. Trabelsi, O. R. Zaïane, A joint topic viewpoint model for contention analysis, in: International Conference on Applications of Natural Language to Data Bases/Information Systems, Springer, 2014, pp. 114–125.

[34] X. Cheng, X. Yan, Y. Lan, J. Guo, Btm: Topic modeling over short texts, IEEE Transactions on Knowledge and Data Engineering 26 (12) (2014) 2928–2941.

[35] Z. Li, W. Shang, M. Yan, News text classification model based on topic model, in: 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), IEEE, 2016, pp. 1–5.

[36] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, H. Yao, Research on topic detection and tracking for online news texts, IEEE Access 7 (2019) 58407–58418.

[37] S. Sendhilkumar, M. Srivani, G. Mahalakshmi, Generation of word clouds using document topic models, in: 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), IEEE, 2017, pp. 306–308.

[38] A. U. Rehman, Z. Rehman, J. Akram, W. Ali, M. A. Shah, M. Salman, Statistical topic modeling for urdu text articles, in: 2018 24th

International Conference on Automation and Computing (ICAC), IEEE, 2018, pp. 1–6.

[39] M. Hasan, M. M. Hossain, A. Ahmed, M. S. Rahman, Topic modelling: A comparison of the performance of latent dirichlet allocation and lda2vec model on bangla newspaper, in: 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), IEEE, 2019, pp. 1–5.

[40] X. Wu, C. Li, Short text topic modeling with flexible word patterns, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–7.

[41] N. Sukhija, M. Tatineni, N. Brown, M. Van Moer, P. Rodriguez, S. Callicott, Topic modeling and visualization for big data in social sciences, in: 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), IEEE, 2016, pp. 1198–1205.

[42] P. Ahmadi, M. Tabandeh, I. Gholampour, Persian text classification based on topic models, in: 2016 24th Iranian Conference on Electrical Engineering (ICEE), IEEE, 2016, pp. 86–91.

[43] R. Pandey, G. O. Mohler, Evaluation of crime topic models: topic coherence vs spatial crime concentration, in: 2018 IEEE International Conference on Intelligence and Security Informatics (ISI), IEEE, 2018, pp. 76–78.

[44] E. Laoh, I. Surjandari, L. R. Febirautami, Indonesians' song lyrics topic modelling using latent dirichlet allocation, in: 2018 5th International Conference on Information Science and Control Engineering (ICISCE), IEEE, 2018, pp. 270–274.

[45] S. ElShal, M. Mathad, J. Simm, J. Davis, Y. Moreau, Topic modeling of biomedical text, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2016, pp. 712–716.

[46] Y. Luo, H. Shi, Using lda2vec topic modeling to identify latent topics in aviation safety reports, in: 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), IEEE, 2019, pp. 518–523.

[47] S. Mifrah, B. L. El Habib, Semantic relationship study between citing and cited scientific articles using topic modeling, in: Proceedings of the 4th International Conference on Big Data and Internet of Things, 2019, pp. 1–8.

[48] C. Zhai, C. Geigle, A tutorial on probabilistic topic models for text data retrieval and analysis, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1395–1398.

[49] B. S. Kumar, V. Ravi, Lda based feature selection for document clustering, in: Proceedings of the 10th Annual ACM India Compute Conference, 2017, pp. 125–130.

[50] O. Mitrofanova, A. Sedova, Topic modelling in parallel and comparable fiction texts (the case study of english and russian prose), in: Proceedings of the International Conference IMS-2017, 2017, pp. 175–180.

[51] L. E. George, L. Birla, A study of topic modeling methods, in: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2018, pp. 109–113.

[52] A. Onan, Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering, IEEE Access 7 (2019) 145614–145633.

[53] L. Sun, J. Chen, J. Li, Y. Peng, Joint topic-opinion model for implicit feature extracting, in: 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), IEEE, 2015, pp. 208–213.

[54] F. Zhang, W. Gao, Y. Fang, B. Zhang, Enhancing short text topic modeling with fasttext embeddings, in: 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), IEEE, 2020, pp. 255–259.

[55] Z. Liu, T. Qin, K.-J. Chen, Y. Li, Collaboratively modeling and embedding of latent topics for short texts, IEEE Access 8 (2020) 99141–99153.

[56] W. Liang, R. Feng, X. Liu, Y. Li, X. Zhang, Gltm: A global and local word embedding-based topic model for short texts, IEEE access 6 (2018) 43612–43621.

[57] L. Li, Y. Sun, X. Han, C. Wang, Research on improve topic representation over short text, in: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), IEEE, 2018, pp. 848–853.

[58] M. Xu, R. Yang, S. Ranshous, S. Li, N. F. Samatova, Leveraging external knowledge for phrase-based topic modeling, in: 2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI), IEEE, 2017, pp. 29–32.

[59] X. Li, C. Li, J. Chi, J. Ouyang, Short text topic modeling by exploring original documents, Knowledge and Information Systems 56 (2) (2018) 443–462.

[60] C. Dai, Y. Wang, Q. Wang, Topic model and similarity calculation of text on spark, in: 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, 2017, pp. 15–19.

[61] S. Subramani, V. Sridhar, K. Shetty, A novel approach of neural topic modelling for document clustering, in: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2018, pp. 2169–2173.

[62] B. Jadhav, D. Bhosale, D. Jadhav, Pattern based topic model for data mining, in: 2016 International Conference on Inventive Computation Technologies (ICICT), Vol. 2, IEEE, 2016, pp. 1–6.

[63] W. Hong, X. Zheng, J. Qi, W. Wang, Y. Weng, Project rank: An internet topic evaluation model based on latent dirichlet allocation, in: 2018 13th International Conference on Computer Science & Education (ICCSE), IEEE, 2018, pp. 1–4.

[64] Q. Chen, L. Yao, J. Yang, Short text classification based on lda topic model, in: 2016 International Conference on Audio, Language and Image Processing (ICALIP), IEEE, 2016, pp. 749–753.

[65] T. T. Dao, T. D. Thanh, T. N. Hai, V. H. Ngoc, Building vietnamese topic modeling based on core terms and applying in text classification, in: 2015 Fifth International Conference on Communication Systems and Network Technologies, IEEE, 2015, pp. 1284–1288.

[66] S. Lee, J. Kim, S.-H. Myaeng, An extension of topic models for text classification: A term weighting approach, in: 2015 International Conference on Big Data and Smart Computing (BIGCOMP), IEEE, 2015, pp. 217–224.

[67] H. Guo, Q. Liang, Z. Li, An improved ad-lda topic model based on weighted gibbs sampling, in: 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE, 2016, pp. 1978–1982.

[68] P. Yang, W. Li, G. Zhao, Language model-driven topic clustering and summarization for news articles, IEEE Access 7 (2019) 185506–185519.

[69] J. Wang, L. Chen, L. Qin, X. Wu, Astm: An attentional segmentation based topic model for short texts, in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018, pp. 577–586.

[70] H. Azarbonyad, M. Dehghani, T. Kenter, M. Marx, J. Kamps, M. De Rijke, Hitr: Hierarchical topic model re-estimation for measuring topical diversity of documents, IEEE Transactions on Knowledge and Data Engineering 31 (11) (2018) 2124–2137.

[71] F. Wang, R. Liu, Y. Zuo, H. Zhang, H. Zhang, J. Wu, Robust word-network topic model for short texts, in: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2016, pp. 852–856.

[72] T. T. Wai, S. S. Aung, Enhanced frequent itemsets based on topic modeling in information filtering, International Journal of Software Innovation (IJSI) 5 (4) (2017) 33–43.

[73] D. Zhao, J. He, J. Liu, An improved lda algorithm for text classification, in: 2014 International Conference on Information Science, Electronics and Electrical Engineering, Vol. 1, IEEE, 2014, pp. 217–221.

[74] H. Kim, Y. Sun, J. Hockenmaier, J. Han, Etm: Entity topic models for mining documents associated with entities, in: 2012 IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 349–358.

[75] S. Li, Y. Zhang, R. Pan, Bi-directional recurrent attentional topic model, ACM Transactions on Knowledge Discovery from Data (TKDD) 14 (6) (2020) 1–30.

[76] D. Jiang, Y. Song, Y. Tong, X. Wu, W. Zhao, Q. Xu, Q. Yang, Federated topic modeling, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 1071–1080.

[77] T. Nguyen, P. Do, Citationlda++ an extension of lda for discovering topics in document network, in: Proceedings of the Ninth International Symposium on Information and Communication Technology, 2018, pp. 31–37.

[78] N. Kawamae, Topic chronicle forest for topic discovery and tracking, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 315–323.

[79] Y. Yang, F. Wang, F. Jiang, S. Jin, J. Xu, A topic model for hierarchical documents, in: 2016 IEEE First International Conference on Data Science in Cyberspace (DSC), IEEE, 2016, pp. 118–126.

[80] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American society for information science 41 (6) (1990) 391–407.

[81] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.

[82] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.

[83] Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei, Hierarchical dirichlet processes, Journal of the american statistical association 101 (476) (2006) 1566–1581.

[84] J. R. Araújo, M. F. Cabana, I. P. Rivera, Aplicación de la herramienta liwc al análisis del discurso político: Los mítines de los candidatos en las elecciones al parlamento de galicia de 2012, in: Investigar la Comunicación Hoy. Revisión de políticas científicas y aportaciones metodológicas: Simposio Internacional sobre Política Científica en Comunicación, Facultad de Ciencias Sociales, Juridicas y de la Comunicación, 2013, pp. 47–64.

[85] E. Fast, B. Chen, M. S. Bernstein, Empath: Understanding topic signals in large-scale text, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 4647–4657.

Patricio Zambrano currently participates in the PhD program of Systems of Escuela Politécnica Nacional Quito-Ecuador, as a research student. In the scientific field he has made significant contributions to the privacy of information with the help of data analytics and Machine Learning. He also works as a full professor in the Department of Computing and Computer Science of the same institution. He received the degree in Electronic Engineering in telecommunications in 2006 at the Polytechnic School of the Army of Sangolquí-Ecuador. Then he reached the master's degree in information and connectivity networks in 2012. His research interests include the security and privacy of the network, research, the application of artificial intelligence to new topics of unconventional research.

Jenny Torres is professor and researcher at the Faculty of Engineering Systems at the Escuela Politécninca Nacional (EPN). She received her PhD in Computer Science at Sorbonne University Campus Pierre and Marie Curie in France in 2013. In 2009 she obtained her M.Sc in Computer Science Security at the University Paris-Est Créteil. Before obtaining a SENESCYT scholarship, she completed a master's degree in Management of Networks and Telecommunications at the Polytechnic School of the Army and in 2008. In 2006 she got her Computer Systems engineer degree at the EPN. Her research focuses on computer security, network management, identity management, wireless networks and open infrastructures. She was an invited researcher at the University of Paraná, Curitiba, Brazil and is part of the research teams Phare and NR2 in France and Brazil, respectively.

Luis Tello-Oquendo received the electronic and computer engineering degree (Hons.) from Escuela Superior Politécnica de Chimborazo (ESPOCH), Ecuador, in 2010; the M.Sc. degree in telecommunication technologies, systems, and networks, and the Ph.D. degree (Cum Laude) in telecommunications from Universitat Politècnica de València (UPV), Spain, in 2013 and 2018, respectively. In 2011, he was a Lecturer with the Facultad de Ingeniería Electrónica, ESPOCH. From 2013 to 2018 he was Graduate Research Assistant with the Broadband

Internetworking Research Group, UPV. From 2016 to 2017 he was a Research Scholar with the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta, GA, USA. He is currently an Associate Professor with the College of Engineering, National University of Chimborazo, Ecuador and Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, Ecuador. His research interest include machine type communications, wireless software-defined networks, 5G and beyond cellular systems, Internet of Things, machine learning. He is a member of the IEEE and ACM. He received the Best Academic Record Award from the Escuela Técnica Superior de Ingenieros de Telecomunicación, UPV, in 2013, and the IEEE ComSoc Award for attending the IEEE ComSoc Summer School at The University of New Mexico, Albuquerque, NM, USA, in 2017.

Ángel Yánez received the computer engineering degree from Escuela Polit́ecnica Nacional (EPN). He worked in the PMO area at the Akros corp company and performed functions in the area of operations. Currently he is linked to projects related to programming and information management.

Luis Velásquez graduate student of the Faculty of Computer Science of the Escuela Politécnica Nacional (EPN). He is currently working as a research assistant. His interest is focused on areas related to software development, ICT management and IT security. He has specialized in topics modeling and data analysis.