

**DEPARTAMENTO DE CIENCIAS**

**DE LA COMPUTACIÓN**

**CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN (SISTEMAS E INFORMATICA)**

**TEMA:**

**“Social media como medio de análisis y prevención del acoso cibernético, aplicado a los perfiles de Facebook, utilizando técnicas de procesamiento de lenguaje natural con base en** **Levenshtein”**

**PLAN DE PROYECTO DE INVESTIGACION**

**Previo a la obtención del Título de Ingeniero en Sistemas e Informática**

**PROPUESTO:**

**DANNY ALEXANDER JÁCOME ANDINO**

**Sangolquí, 09 de mayo de 2019**

**INDICE**

I. TÍTULO DEL PROYECTO 3

II. ÁREA DE CONOCIMIENTO 3

III. ANTECEDENTES 3

IV. PLANTEAMIENTO DEL PROBLEMA 4

V. ESTADO DEL ARTE 4

VI. OBJETIVOS 8

VII. JUSTIFICACIÓN 9

VIII. ALCANCE 9

IX. DEFINICIÓN DE LA INVESTIGACIÓN 10

X. HIPÓTESIS DE TRABAJO 11

XI. HERRAMIENTAS 11

XII. FACTIBILIDAD DEL PROYECTO 12

XIII. TEMARIO 12

XIV. CRONOGRAMA 13

XV. BIBLIOGRAFIA PRELIMINAR 14

# CAPITULO I: INTRODUCCIÓN

# TÍTULO DEL PROYECTO

Social media como medio de análisis y prevención del acoso cibernético, aplicado a los perfiles de Facebook, utilizando técnicas de procesamiento de lenguaje natural con base en Levenshtein.

# ÁREA DE CONOCIMIENTO

* Procesamiento de lenguaje natural (PLN).
* Minería de datos.
* Comparación de cadenas de caracteres Algoritmos de distancia (Levenshtein)

# ANTECEDENTES

Facebook, desde su creación en el 2004 ha tenido un crecimiento exponencial, ahora cuenta con más de seis millones de cuentas solo de estudiantes universitarios en USA, diariamente en todo el mundo se crean alrededor de 20.000 nuevas cuentas. (Samuel D. Gosling, 2007)

En el Ecuador, cerca de un 97.9% de personas que están sobre la franja de edad de los 12 años tiene una cuenta en Facebook, convirtiendo a la red social más utilizada del país comparando con otras redes sociales como Twitter, YouTube y otras. (INEC, 2014)

Al ver desde una perspectiva mundial, Facebook también se encuentra en el primer lugar dentro de las redes sociales con mayor número de usuarios, aunque no es tan marcada la diferencia como en el caso de un referente como Google de acuerdo con los motores de búsqueda en muchos países. Las redes sociales se han convertido y siguen siendo cada vez más populares entre las personas de todo el mundo ya que ayudan a mantenerse en contacto. (J. Haucap y U. Heimeshoff, 2013)

El acoso digital o ciberbullying surge por el rápido avance de las nuevas tecnologías de la comunicación como internet, telefonía móvil, videojuegos, etc. según (Willar) el ciberbullying puede ser definido, en pocas palabras, como “el envío y acción de colgar ‘sending’ y ‘posting’ de textos o imágenes dañinas o crueles en internet u otros medios digitales de comunicación”.

La violencia penetra en todos los ámbitos sociales, tradicionales (escuela, familia, iglesia, política, etc.) y emergentes, como es el caso de la sociedad virtual promovida por las TIC’s. (Solarte Lindo, 2002) y (Tokunaga, 2010). Un amplio estudio realizado entre alumnos de la ESO muestra que el ciberacoso a través de las redes sociales influye significativamente en las posibilidades de fracaso escolar. (Avilés, 2010)

Las redes sociales almacenan una gran cantidad de información “Social media big data”, la misma que se explotado y es clave de las ideas cruciales sobre el manejo y comportamiento humano dentro de la sociedad, esto ha sido estudiado y analizado por académicos, corporaciones, políticos, periodistas y gobiernos (Boyd y Crawford 2012, Lazer et al, 2009).

Según (Samuel D. Gosling, 2007). Los comentarios, notificaciones y otro tipo de expresión escrita o visual, son formas en las cuales un sujeto intenta transmitir como le gustaría ser visto o tratado por los demás, las características van desde pistas sutiles que se encuentran en la elección de la ropa de un individuo a demandas más directas como declaraciones verbales. Otro tipo, es el residuo de la conducta, esta se refiere a las pistas dadas por el comportamiento de una persona. Por ejemplo, una colección de películas bien organizadas refleja la tendencia de un individuo a ser organizado.

Los comentarios que son expresados como texto escrito por los usuarios de las redes sociales en forma de lenguaje humano pueden ser procesadas y analizadas para obtener información de gran valides, utilizando para esto procesamiento de lenguaje natural (PLN), el mismo que se entiende como la habilidad de la máquina para procesar información, y no solo identificar letras y/o sonidos del lenguaje humano. (Vásquez, A. C., Quispe, J. P., & Huayna, A. M. 2009).

Dentro del procesamiento de lenguaje natural (PLN) una de las técnicas que se pueden utilizar es el análisis morfológico o léxico el cual consiste dividir el texto en oraciones y estas oraciones en palabras para luego realizar un análisis interno de las palabras para extraer lemas, rasgos flexivos, unidades léxicas compuestas, para realizar este proceso se puede utilizar el algoritmo de Levenshtein que se basa en técnicas de comparación alfabética y utiliza métricas de distancia de edición. Para lo cual se asigna un peso a cada comparación entre dos cadenas, en términos del número de operaciones de inserción, eliminación y sustitución necesarias para editar y convertir la primera cadena en la segunda.

# PLANTEAMIENTO DEL PROBLEMA

La evolución en los estudios sobre el acoso (bullying), se han ido introduciendo cambios en la realidad del maltrato, y han ido apareciendo nuevas formas de agredir a los iguales, nuevas modalidades de acoso. Una de ellas es el acoso cibernético (ciberbullying) que consiste en utilizar las nuevas tecnologías de la información y la comunicación, principalmente Internet y el teléfono móvil, para hostigar y acosar.

Es claro que el acoso cibernético (ciberbullying) está presente en Ecuador y que cada vez es más común encontrar este tipo de comportamientos en los estudiantes de colegio, universidades, profesionales, etc. las afecciones de este tipo de acoso puede provocar diferentes problemas dentro del desenvolvimiento de las personas, los cuales se pueden ser reflejadas en el aprovechamiento académico, disciplinario, laboral y/o en el comportamiento dentro de la sociedad.

Las consecuencias que ocasionan el bullying y ciberbullying son dañinas para todos los implicados con diferentes señales y niveles de sufrimiento, los efectos mas notables son los que presentan las víctimas. Los agresores y observadores de maltrato también son influenciados por los hábitos negativos y pueden ser parte de su comportamiento actual y futuro. Todas las personas implicadas en actos de Bullying o Ciberbullying tienen riesgo a sufrir desajustes psicosociales y trastornos psicopatológicos en la adolescencia y/o en la vida adulta, una de las consecuencias más extrema es el suicidio o la muerte de la víctima Olweus (1973). Las consecuencias, aunque no sean tan extremas sí afectan a la salud, a la calidad de vida, al bienestar y al correcto desarrollo de la persona.

Las víctimas de Ciberbullying sufren mayor daño psicológico debido a que la información lesiva está disponible para todo el mundo las 24 horas del día, los acosadores con frecuencia son anónimos, el proceso de victimización es continuo e inevitable, la mayoría de las veces es muy difícil eliminar el material publicado y suele estar accesible de forma pública durante largos períodos de tiempo. (Feinberg y Robey, 2009).

# OBJETIVOS

* 1. **Objetivo General**

Utilizar el social media como medio de análisis y prevención del acoso cibernético, aplicado a los perfiles de Facebook, utilizando técnicas de procesamiento de lenguaje natural con base en Levenshtein para obtener un perfil de personalidad de apoyo hacia los psicólogos.

* 1. **Objetivos Específicos**
  2. Identificar perfiles de usuario en Facebook donde el ciberbullying este presente, usuarios con una franja de edad mayor a los 12 años y menor a los 18 años.
  3. Realizar una investigación exploratoria para identificar y plantear un diccionario de datos de acuerdo con el ciberbullying, con las palabras más influyentes de acuerdo con las directrices de un psicólogo.
  4. Identificar la interfaz de programación de aplicaciones (API) que permita extraer la información de Facebook.
  5. Realizar un aplicativo de escritorio para el procesamiento de los datos obtenidos de las redes sociales.
  6. Analizar y evaluar los patrones de ciberbullying identificados mediante matrices de probabilidad e impacto y el apoyo de un psicólogo experimentado en el ámbito del ciber acoso para sugerir un perfil de personalidad.

# JUSTIFICACIÓN

En la actualidad el maltrato entre iguales a través de los dispositivos móviles o Internet acoso cibernético (ciberbullying) se ha convertido en un problema complejo, Las bases comunes entre bullying y ciberbullying son como caras de una moneda que comparten componentes básicos del maltrato entre iguales (Olweus, 1998; Ortega,1992).

Son el mismo fenómeno, el ciberbullying muestra el ejercicio del abuso con tecnologías como los dispositivos móviles e Internet, a través de sus múltiples modalidades. Ello no debe impedir descubrir los componentes comunes y detallar los específicos, que sin duda existen y marcan las consecuencias que provocan.

Las tareas de prevención e intervención en la comunidad habrán de dirigirse a analizar esos componentes comunes como los identitarios del maltrato entre iguales y a encarar los propios del ciberbullying con medidas efectivas en el contexto en que sucede.

Las actuaciones de acoso contra los iguales, ya sea de forma directa y presencial como de forma distanciada a través de dispositivos móviles y espacios virtuales, pretenden socavar la dignidad de sus destinatarios, las víctimas.

La mayoría participan de intención y de recurrencia y visualizan la distancia de poder que hay entre quienes agreden y quienes son maltratados. En su justificación están desprovistas de argumentación admisible y cargadas de gratuidad y de juicios morales inaceptables. La manifestación de estas conductas es exponente de una carencia de orden moral que las oriente. Quien las practica no recuerda, ni posee, ni construye referentes morales adecuados para tomar conciencia, y termina echando mano de otros. Suele ignorar los que le ajustan a la norma social y moral. No cumple ni la premisa de ‘deber hacer’ –moral- ni la de ‘querer hacer’ ética (Tognetta, 2009), y orienta su conducta hacia lo negativo e indeseado, hacia la agresión y el abuso.

Según un estudio realizado por (Bachrach, 2015) la información que se puede obtener de Facebook es de gran ayuda para identificar este tipo de conductas, debido a que en este se incluyen datos como pensamientos, opiniones y sentimientos, lo que permite determinar la personalidad del usuario de Facebook. Lo cual hace mucho más fácil identificar niveles o patrones de ciberbullying dentro de los usuarios de Facebook, mediante el análisis del contenido que se publica dentro del mismo.

# ALCANCE

El proyecto se llevará a cabo hasta la implementación de un prototipo el cual tendrá las siguientes funcionalidades:

1. Planteamiento de un diccionario de datos personalizado con relación al ciberbullying.
2. Recolección de datos del usuario de Facebook por medio de una API que permita la extracción de los datos.
3. Almacenamiento de los datos en una base.
4. Limpieza de datos y eliminación de las palabras de conexión dentro de los comentarios extraídos como “el, la, los, las, de, y, con, etc.”.
5. Realizar un aplicativo de escritorio para el procesamiento de los datos obtenidos.
6. Analizar, evaluar y sugerir un perfil de personalidad.

El estudio se realizará con personas de la ciudad de Quito que estén dentro del rango de edad de los 12 a los 18 años, de preferencia que sean estudiantes de colegio, para obtener una muestra lo más homogénea posible y numéricamente representativa al momento de recolectar los datos, para de esta manera tener un menor rango de error o sesgo al identificar patrones de comportamiento, frente al acoso cibernético (ciberbullying).

Una vez conocido los objetivos específicos se proponen las siguientes preguntas de investigación que se detallan por cada objetivo en la Tabla 2.

|  |  |
| --- | --- |
| **Objetivos Específicos** | **Pregunta De Investigación** |
| 1. **Identificar perfiles de usuario en Facebook donde el ciberbullying este presente, usuarios con una franja de edad mayor a los 12 años y menor a los 18 años.** | 1. ¿En qué rango de edades está presente el ciberbullying con mayor frecuencia? 2. ¿Las redes sociales han permitido que los usuarios realicen actos de agresión en contra de sus iguales? |
| 1. **Realizar una investigación exploratoria para identificar y plantear un diccionario de datos de acuerdo con el ciberbullying, con las palabras más influyentes de acuerdo con las directrices de un psicólogo.** | 1. ¿Existe alguna metodología o tipo de investigación que por medio de la observación permita obtener información? 2. ¿Cómo se pueden plantear características lógicas que identifiquen el ciberbullying? 3. ¿Qué tipo de profesional puede identificar el comportamiento de un sujeto? |
| 1. **Identificar la interfaz de programación de aplicaciones (API) que permita extraer la información de Facebook.** | 1. ¿Se puede acceder a la información de Facebook? 2. ¿Qué aplicación permite la comunicación de dos componentes de software para extraer información desde una red social? |
| 1. **Realizar un aplicativo de escritorio para el procesamiento de los datos obtenidos de las redes sociales.** | 1. ¿Cómo se podría facilitar el procesamiento de la información obtenida desde Facebook? 2. ¿Se ha utilizado antes PLN basado en Levenshtein para procesar información? 3. ¿Es fácil implementar un aplicativo escritorio con PLN? 4. ¿Qué metodología de desarrollo de software se usará para la aplicación? |
| 1. **Analizar y evaluar los patrones de ciberbullying identificados mediante matrices de probabilidad e impacto y el apoyo de un psicólogo experimentado en el ámbito del ciber acoso para sugerir un perfil de personalidad.** | 1. ¿Son claros los patrones identificados? 2. ¿Es fácil evaluar a un individuo basándose en estos patrones? 3. ¿Son satisfactorios los resultados obtenidos? 4. ¿El perfil de personalidad sugerido es congruente con la de una prueba de un psicólogo? |

Tabla 2, Preguntas de investigación

La ilustración 1 muestra la topología experimental planteada, así como las principales herramientas que se utilizarán para la extracción, procesamiento y análisis de los datos obtenidos.

Por medio de la Api de Facebook se podrán extraer los datos desde los perfiles de Facebook a NetBeans, para almacenarlos en una base de datos MySQL, para luego llevar cabo el proceso de limpieza y procesamiento de los datos por medio de (PLN) y finalmente evaluar el conocimiento obtenido.

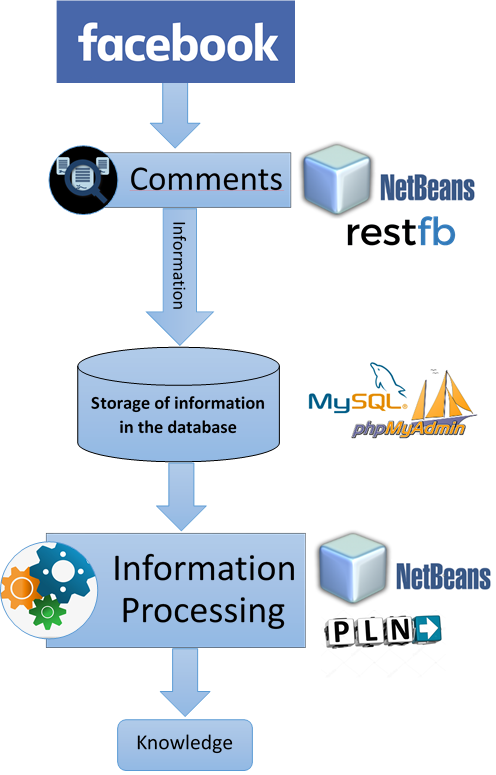


Ilustración 1: Topología experimental del sistema.

# CAPITULO II: ESTADO DEL ARTE

Con el fin de identificar y analizar el estado del arte actual se estructuro en cuatro secciones que están relacionadas con el tema que se estudiara durante el presente trabajo de investigación: En la primera sección se expone características de ciberbullying que se han identificado. En la segunda sección, se exponen artículos que tratan a las redes sociales y el ciberbullying. En la tercera sección se presentan trabajos de investigación que tratan a la social media data mining y ciberbullying. En la última sección se presentan artículos que especifican las formas de analizar el contexto por medio de procesamiento de lenguaje natural y algoritmos de comparación de cadenas de caracteres (análisis morfológicos).

El proceso de estudio de antecedentes realizado contempla las siguientes fases: (1) planteamiento del estudio sistemático de literatura, (2) obtención del grupo de control y palabras clave para la investigación, (3) creación y afinación de la cadena de búsqueda y, (4) elaboración del estado del arte. Cada actividad de este proceso se describe a continuación.

1. **Planteamiento del estudio preliminar sistemático de literatura**

Como fase inicial del proceso de estudio de antecedentes, se realizó la descripción del problema en el cual está enfocada la investigación, a fin de definir el marco de búsqueda de estudios científicos, seguido del planteamiento de las preguntas de investigación y finalmente el establecimiento de los criterios de inclusión y exclusión.

1. **Obtención del grupo de control y palabras clave para la investigación**

De acuerdo con (Levac & Colquhoun, 2010) esta fase implica identificar los estudios relevantes y desarrollar un plan de decisión para dónde buscar, qué términos usar y qué fuentes se deben consultar. Las fuentes incluyen bases de datos electrónicas, listas de referencia, búsquedas manuales de revistas clave, organizaciones y conferencias.

Tras el análisis de varios artículos científicos propuestos por los investigadores, se obtuvo el grupo de control (GC) del estudio de antecedentes, este grupo está conformado por 5 artículos científicos que se consideraron como relevantes e influyentes en la investigación. Esta información se encuentra detallada en la Tabla 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Código** | **Título** | **Cita** | **Palabras clave** |
| **EC1** | **A Comparison of Common Users across Instagram and Ask.fm to Better Understand Cyberbullying** | *Li, H. H. S., Yang, Z., Lv, Q., Han, R. I. R. R., & Mishra, S. (2014, December)* | Correlation, Media, Twitter, Educational institutions, Facebook, Computer science |
| **EC2** | **Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología** | *Mariñelarena-Dondena, L., Errecalde, M. L., & Castro Solano, A. (2017).* | Técnicas de Minería de Textos, Ciencias de la Computación, Evaluación, psicología |
| **EC3** | **Determinación de niveles de agresividad en comentarios de la red social Facebook por medio de Minería de Texto** | *Martel, W., Carranco, D., & Cevallos, D. (2016).* | Text Mining, distancia de Levenshtein, API de consultas de Facebook. |
| **EC4** | **Detección de ciberbullying a través de un sistema de reconocimiento óptico de caracteres en imágenes de páginas web** | *BARRAGÁN, R. M* | Ciberbullying, cyberbullying, internet, reconocimiento óptico. |
| **EC5** | **Arquitectura (advi) para la detección de vocabulario de ciberbullying en internet combinando técnicas de big data analytics y web semántica** | *Zuñiga, i. C., rosas, f. L., arteaga, j. M., & veyna, j. I. L. (2016).* | Cyberbullying, Big Data, PLN, Semantic Web, Supervised Learning, Genetic Algorithms, Web Semántica, aprendizaje supervisado, algoritmos Genéticos. |

Tabla 1, Grupo de control y palabras clave

La selección de los artículos científicos pertenecientes al grupo de control (GC), permitió determinar las palabras clave que se encuentren alineadas al objetivo de la investigación, las cuales fueron: cyberbullying, ciberbullying, big data, PLN, aprendizaje supervisado, minería de texto, text mining, Facebook, redes sociales, Levenshtein, Matching techniques.

1. **Creación y afinación de la cadena de búsqueda**

Una vez que las palabras clave han sido identificadas, se procede a crear y probar diversas cadenas de búsqueda, en este caso, se lo realizó en el buscador científico-académico Google Scholar. La primera cadena de búsqueda creada fue la siguiente:

***ALL*** *({cyber bullying}* ***OR*** *{cyberbullying})* ***AND******ALL*** *({big data}* ***OR*** *{text mining}) AND* ***ALL*** *({redes sociales}* ***OR*** *{Facebook})* ***AND ALL*** *({natural language processing})*

Esta cadena obtuvo muy pocos artículos científicos como resultado de búsqueda, por lo que se tuvo que seguir con la afinación de esta y tras varias cadenas de búsqueda generadas con la combinación de las palabras clave definidas y los conectores específicos de la base digital, se determinó la cadena de búsqueda final:

***ALL*** *({cyber bullying}* ***OR*** *{cyberbullying}* ***OR*** *{cyber-bullying})* ***AND******ALL*** *({big data}* ***OR*** *{social media}* ***OR*** *{text mining})* ***AND******ALL*** *({social networks}* ***OR*** *{Facebook}* ***AND******ALL****({PLN}* ***OR*** *{NLP}* ***OR*** *{natural language processing})* ***AND ALL****({Levenshtein}* ***OR*** *{*Matching *techniques})*

A fin de detallar con mayor precisión el tipo de artículo científico que se consideraría como válido para la realización del estudio de antecedentes se determinó aplicar dos filtros de búsqueda adicionales:

1. **Año:** Los artículos candidatos deberían haber sido publicados a partir del año 2010. Se decidió aplicar este filtro con el objetivo de que las soluciones propuestas a la problemática investigada, se encuentren lo más apegadas a la realidad actual de los avances tecnológicos.
2. **Tipo de documento:** Se consideraron únicamente documentos de tipo: Conference Paper y Article, debido a que este tipo de documentos son reflejo de una investigación que genera un impacto elevado en el mundo científico.
3. **Elaboración del estado del arte**

Según (Menay-López & Fuente-Mella, 2014) el ciberbullying, es un modo disimulado de acoso verbal y escrito. Willard (2004) identifica siete categorías de violencia verbal y escrita a través de las nuevas tecnologías:

1. **Flaming:** envío de mensajes vulgares o que muestran enfado sobre una persona a un grupo online o a esa persona vía email o SMS.
2. **Acoso online:** envío repetido de mensajes ofensivos vía email o SMS a una persona.
3. **Ciberstalking:** acoso online que incluye amenazas de daño o intimidación excesiva.
4. **Denigración:** envíos perjudiciales, falsas y crueles afirmaciones sobre una persona a otras o comentarios en lugares online.
5. **Suplantación de la persona:** Hacerse pasarse por la víctima y enviar o colgar archivos de texto, video o imagen que hagan quedar mal al agredido**.**
6. **Outing:** enviar o colgar material sobre una persona que contenga información sensible, privada o embarazosa, incluido respuestas de mensajes privados o imágenes.
7. **Exclusión**: Cruel expulsión de alguien de un grupo online.

El artículo de (Shally Bhardwaj, 2013) se explica que la conducta social en línea y el uso de las redes sociales están relacionadas con la personalidad y que las personas se comportan de la misma manera que lo hacen en Internet, debido a que en las redes sociales están asociadas por la interacción humana. Este estudio determinó que las personas no son conscientes a la hora de usar las redes sociales, por ejemplo, dar like a una página de su agrado, esta función se la hace inconscientemente de acuerdo al gusto que define la personalidad del usuario. Según los estudios (Seyed Morteza Ghavami, 2015) la personalidad de un usuario puede predecirse con exactitud a través de la información disponible públicamente. Para esto, se construyó un sistema con características lingüísticas y estructurales utilizadas para predecir la personalidad de las personas en las redes sociales. Este sistema puede predecir puntuaciones de rasgos de personalidad dentro del 11% -18% de sus valores reales.

En el artiulo planteado por (Martel, W., Carranco, D., & Cevallos, D. 2016). hacen un estudio sobre la agresividad en las redes sociales tomando como base de datos los comentarios de Facebook y utilizan la herramienta R que tiene incorporado el algoritmo de Levenshtein cuyo nombre de paquete es “stringdist”. Este paquete tiene muchos algoritmos entre ellos el “stringsim” de la misma familia. Este último algoritmo, retorna la probabilidad de similitud entre dos palabras el mismo que es utilizado en este trabajo, lo que no se especifica en este trabajo es el número de arreglos máximos para una palabra y el valor que tiene cada uno de ellos dentro del análisis.

En el trabajo que realizaron por (Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. 2003) con un grupo de técnicas de coincidencia fonéticas y de caracteres, basadas en tokens e hibridas, fueron sometidas a pruebas y luego comparadas. Este trabajo no analizo el problema con la tasa de error, el tipo de errores tipográficos y el tamaño de un conjunto de datos. Además, en el trabajo solo se menciona el valor del umbral con el que se trabaja sin mencionar como fue calculado dicho valor para cada técnica, al finalizar la evaluación ellos afirman que Monge-Elkan tiene mejor promedio, pero Soft-IDF tuvo mejor desempeño que el resto de las técnicas desde una mirada general.

(Hassanzadeh, O., Sadoghi, M., & Miller, R. J. 2007, September). Realiza una un estudio general de algunas técnicas de coincidencia de palabras donde evalúa la precisión con diferentes tipos de datos. Para este trabajo se tomó solo en cuenta técnicas de coincidencia de caracteres basadas en tokens, los erros que se tomaron en cuenta fueron los de edición y reemplazo. Se menciona que el valor umbral utilizado si afectara al rendimiento de cada una de las técnicas. Las cosas que no se toman en cuenta para este estudio son los errores tipográficos y la cantidad de datos en cada grupo para las pruebas.

Al igual que Hassanzadeh, (Peng, T., Li, L., & Kennedy, J. 2012) presenta un trabajo comparativo sobre el rendimiento de técnicas de coincidencias de nombres, en el cual consideraron algunos factores que Hassanzadeh no tomo en cuenta, como son la tasa de errores, el tamaño del conjunto de datos, los cuales tienen un efecto sobre los resultados de las diferentes técnicas. Los resultados experimentales confirman que no hay una técnica mejor, en general, Jaro-Winkler y Jaro funcionan mejor que los demás en la coincidencia de nombres. El trabajo también afirmó que la tasa de error en el conjunto de datos tiene efecto sobre los valores umbral. Sin embargo, al igual que Hassanzadeh no consideraron los errores tipográficos y nombres de pila.

Con base en una revisión exhaustiva de las investigaciones existentes, estamos seguros de que no existe una técnica única y exclusiva que pueda resolver todas las tareas de coincidencia de caracteres, particularmente cuando los datos están en idioma español y no solo se utilizan palabras pertenecientes al diccionario, si no que se utilizan los neologismos y modismos de las regiones del Ecuador los cuales pueden cambiar de sentido al contexto dependiendo de la región.

# CAPITULO III: METODOLOGÍA Y MARCO TEÓRICO

# DEFINICIÓN DE LA INVESTIGACIÓN

La investigación para el trabajo propuesto es de tipo exploratorio ya que consiste en proveer una referencia general de la temática, se investiga y analiza para plantear de forma concreta una propuesta que ayude con la problemática planteada. Y que llegará a las conclusiones basándose en experimentos donde se reflejarán los hechos de una forma cuantitativa, esto experimentos se llevarán a cabo con varios grupos homogéneos de estudiantes pertenecientes a colegios de Quito, lo mismo que nos ayudará a tener resultados con un menor sesgo y concluir si los resultados son fiables, este proceso se apoyara en la opinión de un psicólogo experimentado y test de conducta.

Según (Letelier, 2006) las metodologías ágiles permiten tener un control en el desarrollo del proyecto, tanto sobre los recursos humanos, como las herramientas a usar. Las metodologías ágiles buscan desarrollar el software más rápido sin tanta documentación, esto se debe a que el cliente o el experto deben estar en constante participación en el proyecto, por lo que el mismo puede sufrir varios cambios según los requerimientos del cliente.

En el presente proyecto se utilizará la metodología ágil de Programación Extrema (XP), ya que esta permite avanzar el proyecto de una manera rápida sin tener que realizar tanta documentación y con un enfoque más al desarrollo y el manejo de las herramientas a utilizar. (Letelier, 2006)

# HIPÓTESIS DE TRABAJO

El estudio del social media como análisis y prevención del acoso cibernético, es una investigación cualitativa, se propone la siguiente hipótesis:

**La social media aplicado a perfiles de Facebook permitirá analizar y prevenir el acoso cibernético basándose en PLN con base en Levenshtein.**

# HERRAMIENTAS

En el proyecto se pretende utilizar las siguientes herramientas:

**IDE de desarrollo NetBeans**

NetBeans (Oracle, NetBeans IDE, 2017) es una herramienta de desarrollo integrado libre, su principal lenguaje de programación es Java, con esta herramienta es posible conectarse a bases de datos. Por lo cual se utilizará para el desarrollo de la aplicación que realizará el procesamiento del lenguaje natural.

**MYSQL**

MySQL (Oracle, MySQL, 2017) es un sistema de gestión de bases de datos relacional y que se la puede conseguir de manera open sourse. Esta herramienta permite guardar los datos a obtener de los perfiles de Facebook para posteriormente ser analizados con procesamiento de lenguaje natural. Los datos obtenidos sobre un perfil de Facebook se pueden guardar de una forma ordenada y con relación a diferentes tablas.

**Librería RestFB**

RestFB es un cliente de API de Facebook Graph simple y flexible escrito en Java. Se trata de software de código abierto liberado bajo los términos de la Licencia MIT.

La API de RestFB es realmente mínima y sólo necesita usar un método para obtener información y publicar nuevos elementos en Facebook. Proporcionamos implementaciones predeterminadas para todos los componentes principales. Por medio de esta librería obtendremos la conexión directa al API Graph de Facebook y a la información “comentarios” existente dentro de los perfiles de Facebook.

# 

# FACTIBILIDAD DEL PROYECTO

**Factibilidad Técnica**

Para realizar el proyecto se cuenta con el equipo computacional del estudiante a cargo del desarrollo del trabajo de investigación. El estudiante cuenta por su parte con un computador con el rendimiento requerido para soportar las herramientas que se van a utilizar, descritas en la sección anterior.

**Factibilidad Operativa**

El estudiante y el docente tutor de tesis deben poseer conocimientos de uso de técnicas de procesamiento de lenguaje natural, y en los algoritmos de camparacion de caracteres basados en distancias, que serán evaluados en el presente trabajo. Adicional, es necesario mencionar que los involucrados en el desarrollo del presente proyecto poseen conocimientos de herramientas de desarrollo de software y utilidades de la API de Facebook.

**Factibilidad Económica**

Todo el proyecto será autofinanciado por el estudiante.

|  |  |
| --- | --- |
| **Descripción** | **Costos** |
| Servicios Básicos | $ 120.00 |
| Papelería e Impresiones | $ 150.00 |
| Transporte | $ 80.00 |
| Licencias De Software | $ 0.00 |
| Laptop | $ 1100.00 |
| Horas De Desarrollo | $ 1500.00 |
| **Total** | **$ 2950.00** |

# TEMARIO

**Capítulo I:** INTRODUCCIÓN

* 1. Antecedentes.
  2. Planteamiento del problema.
  3. Justificación.
  4. Objetivos.
  5. Alcance.

**Capitulo II:** ESTADO DEL ARTE

1. Bullying y Ciberbullying
2. Clasificación Ciberbullying.
   1. Flaming.
   2. Acoso online.
   3. Ciberstalking.
   4. Denigración.
   5. Suplantación de la persona.
   6. Outing.
   7. Exclusión.
3. Técnicas de Procesamiento de Lenguaje Natural (PLN)
   1. Técnicas de deletreo y distancia
   2. Técnicas de sonido y fonética
   3. Técnicas compuestas
4. Algoritmos de comparación de caracteres
5. Social media data mining

**Capítulo III:** METODOLOGÍA Y MARCO TEÓRICO

1. Metodología de desarrollo
2. Herramientas.

**Capítulo IV:** DISEÑO Y DESARROLLO

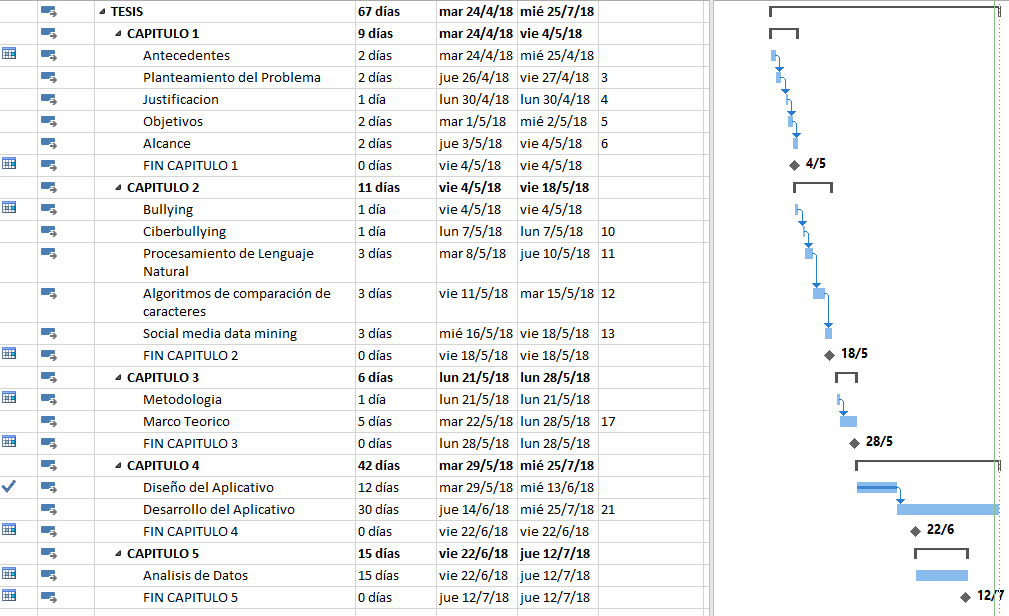
1. Diseño del aplicativo
2. Implementación del aplicativo
3. Procesamiento de lenguaje natural
4. Algoritmo de comparación de caracteres
5. Planteamiento de patrones

**Capítulo V:** ANÁLISIS DE DATOS

1. Análisis de patrones
2. Evaluación de resultados
3. Pruebas psicológicas (test personalidad)
4. Discusión de resultados

**Conclusiones y recomendaciones**

# CRONOGRAMA



# 

# BIBLIOGRAFIA PRELIMINAR

*INEC, «Encuesta de Condiciones de vida realizada por el Instituto Nacional de Estadísticas y Censos,» Quito, 2014.*

*J. Haucap y U. Heimeshoff, «Google, Facebook, Amazon, eBay: Is the Internet driving competition or market monopolization?,» 2013.*

*Delgado, J., Galarraga, F., Fuertes, W., Toulkeridis, T., Villacís, C., & Castro, F. (2016, March). A proposal of an entity name recognition algorithm to integrate governmental databases. In eDemocracy & eGovernment (ICEDEG), 2016 Third International Conference on (pp. 26-33). IEEE.*

*Garaigordobil, M. (2011). Prevalencia y consecuencias del cyberbullying: una revisión. International Journal of Psychology and Psychological Therapy, 11(2).*

*Bachrach, Y. (2015). Human Judgments In Hiring Decisions Based On Online Social Network Proﬁles. IEEE.*

*Samuel D. Gosling, S. G. (2007). Personality Impressions Based on Facebook Profiles. ICWSM’2007 Boulder, Colorado, USA.*

*Menay-López, L., & de la Fuente-Mella, H. (2014). Plataformas comunicacionales del ciberbullying: Una aplicación empírica en dos colegios de la quinta región, Chile. Estudios pedagógicos (Valdivia), 40(2), 117-133.*

*Levac, D., & Colquhoun, H. (2010). Scoping studies: advancing the methodology. Implementation science.*

*Río-Pérez, J. D., Sádaba, C., & Bringué, X. (2010). Menores y redes¿ sociales?: de la amistad al cyberbullying.*

*Mariñelarena-Dondena, L., Errecalde, M. L., & Castro Solano, A. (2017). Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. Revista Argentina de Ciencias del Comportamiento, 9(2).*

*Martel, W., Carranco, D., & Cevallos, D. (2016). Determinación de niveles de agresividad en comentarios de la red social Facebook por medio de Minería de Texto. GEEKS DECC-REPORTS, 6(1).*

*Zuñiga, i. C., rosas, f. L., arteaga, j. M., & veyna, j. I. L. (2016). Arquitectura (advi) para la detección de vocabulario de ciberbullying en internet combinando técnicas de big data analytics y web semántica. Dyna new technologies, 3(1).*

*BARRAGÁN, R. M. Detección de ciberbullying a través de un sistema de reconocimiento óptico de caracteres en imágenes de páginas web.*

*Constante Portero, D. D. (2018). Las redes sociales como plataforma para el grooming en los jóvenes de 12 a 17 años de la ciudad de Ambato (Bachelor's thesis, Universidad Técnica de Ambato. Facultad de Diseño, Artes y Arquitectura. Carrera de Diseño Gráfico Publicitario.).*

*Li, H. H. S., Yang, Z., Lv, Q., Han, R. I. R. R., & Mishra, S. (2014, December). A comparison of common users across instagram and ask. fm to better understand cyberbullying. In Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on (pp. 355-362). IEEE.*

*Hassanzadeh, O., Sadoghi, M., & Miller, R. J. (2007, September). Accuracy of Approximate String Joins Using Grams. In QDB (pp. 11-18).*

*Peng, T., Li, L., & Kennedy, J. (2012). A Comparison of Techniques for Name Matching. GSTF journal on Computing, 2(1).*

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., & Fienberg, S. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, *18*(5), 16-23.

Cohen, W., Ravikumar, P., & Fienberg, S. (2003, August). A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation* (Vol. 3, pp. 73-78).

Samuel D. Gosling, S. G. (2007). Personality Impressions Based on Facebook Profiles. *ICWSM’2007 Boulder, Colorado, USA*.

Olweus, D. (1973). Hackkycklingar och översittare [Bullies and whipping boys]. *Kungälv, Sweden: Almqvist & Wiksell*.