INSTITUTO POLITÉCNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO

Trabajo Terminal:

“Sistema de identificación de artículos utilizando reconocimiento de patrones.”
2015B-032

Presentan:
Calvillo Juárez Rogelio
Castañeda Vite Carlos Rodrigo
Mejía Hernández Miguel

En el presente documento se encuentran los resultados correspondientes al desarrollo del Trabajo Terminal cuyo objetivo es la implementación de un sistema que analice imágenes mediante el procesamiento de las mismas . Este sistema pretende servir como herramienta capaz

Palabras Clave: Reconocimiento de patrones, procesamiento de imágenes, inteligencia artificial, aprendizaje máquina, aplicación web.

Directores:
M. en C. Ramírez Morales Mario Augusto, M. en E. Silva Sánchez Carlos

Índice general

1. Antecedentes **1**
- 1.1. Introducción . . . 1
- 1.2. Descripción del problema . . . 2
- 1.3. Solución propuesta . . . 2
- 1.4. Estado del arte . . . 2
- 1.5. Justificación . . . 2

2. Marco Teórico **4**
- 2.1. Minería de texto . . . 4
- 2.2. Lenguaje natural . . . 6
- 2.3. Procesamiento del lenguaje natural . . . 6
- 2.4. Análisis morfológico . . . 7
- 2.5. *Stemming* . . . 7
- 2.6. Lematización . . . 9
- 2.7. *Stopwords* . . . 9

3. Especificaciones del proyecto **11**
- 3.1. Objetivo . . . 11
  - 3.1.1. Objetivo general . . . 11
  - 3.1.2. Objetivos específicos . . . 11
- 3.2. Metodología . . . 11
- 3.3. Descripción de prototipos . . . 12
- 3.4. Arquitectura del sistema . . . 12

Índice de figuras

2.1. Proceso de la Minería de texto 4

3.1. Metodología 12
3.2. Arquitectura general del sistema 13

Índice de tablas

1.1. Comparativa de Sistemas que realizan análisis de textos 2
1.2. Trabajos Teminales relacionados con el procesamiento de lenguaje natural 3

2.1. Categorías de palabras 7

3.1. Prototipos y Funcionalidades Generales 14

CAPÍTULO 1

Antecedentes

1.1. Introducción

Con la llegada de las computadoras personales en el año 1970, surgió la demanda de aplicaciones prácticas que utilizaran reconocimiento de patrones.

El reconocimiento de patrones es una disciplina científica cuyo objetivo es la clasificación de objetos dentro de categorías o clases. Dependiendo de la aplicación, estos objetos pueden ser imágenes, señales o cualquier tipo de entidades medibles que necesiten ser clasificadas.[1] El reconocimiento de patrones es una parte integral de la mayoría de los sistemas de inteligencia artificial construidos para la toma de decisiones.

Una subdisciplina de la inteligencia artificial es el procesamiento de imágenes cuyo término es utilizado para describir a la función implementada por software o hardware que permite analizar y clasificar imágenes la cual emplea teoría del reconocimiento de patrones. [2]

El presente documento contiene el análisis y la documentación del Trabajo Terminal 2015-B032 "Sistema de identificación de artículos utilizando reconocimiento de patrones.", el cual pretende implementar técnicas de las dos disciplinas descritas anteriormente.

1.2. Descripción del problema

1.3. Solución propuesta

El presente Trabajo Terminal implementa un sistema que mediante reconocimiento de patrones, aprendizaje máquina y minería de datos; sea capaz de analizar y clasificar texto en el idioma español a partir de un conjunto de datos estadísticos. Aplicado al caso de estudio *online grooming*.

1.4. Estado del arte

El interés por desarrollar sistemas que apoyan el análisis de textos se ha incrementado. En la tabla 1 se mencionan algunos sistemas que hacen uso de estos aplicado a diferentes problemáticas sociales.

Nombre	Descripción	Tipo	Año	Lugar de desarrollo
ChildDefence	La aplicación ChildDefence permite recolectar conversaciones (escritas en inglés) que un niño ha tenido con una persona, provenientes de mensajes de texto o chats en línea, y luego hacer que se analicen en el propio teléfono. La aplicación, ha sido entrenada para identificar si la persona es un niño o un adulto.	Gratuita	2011	Isis Forensics
LIWC	Es un programa de computadora analizador de textos, que determina el porcentaje de emociones expresadas en un texto, con base en diversas categorías.	Comercial	2001	Austin Texas University, EUA
KidsWatch	Es un programa de control parental. Funciona de tal manera que cuando en la conversación de un chat encuentra palabras referentes al sexo, drogadicción, armas, suicidio entre otros se les avisa a los padres acerca del comportamiento sospechoso.	Comercial	2002	Computer Business Solutions

Tabla 1.1: Comparativa de Sistemas que realizan análisis de textos

Dentro de la Escuela Superior de Cómputo, también se encontró un Trabajo Terminal que implementa procesamiento de lenguaje natural y minería de datos para la toma de decisiones. La descripción se encuentra en la tabla 1.

1.5. Justificación

La llegada de internet abrió las puertas a grandes posibilidades de comunicación: redes sociales, foros, chats, etcétera. El término *grooming* hace referencia a las acciones que lleva acabo un adulto para establecer amistad con un menor por medio de internet, con el objetivo de obtener una satisfacción sexual mediante la obtención de imágenes con contenido erótico o sexual del menor. A pesar de que estas situaciones tienen su origen dentro de la red, muy frecuentemente terminan en abuso físico a menores o tráfico de pornografía infantil.

Nombre	Descripción	Tipo	Año	Lugar de desarrollo
Prototipo de sistema de información con minería de datos para la toma de decisiones. (20060106)	Sistema con una adaptación de minería de datos dirigida hacia las PyMES, obteniendo información relevante de sus bases de datos mostrando interpretaciones de los resultados de la minería en enunciados de lenguaje natural.	Trabajo Terminal	2006	ESCOM

Tabla 1.2: Trabajos Teminales relacionados con el procesamiento de lenguaje natural

Los problemas que se presentan para atacar el grooming son principalmente: La inocencia de los menores, el anonimato en el que se mantienen los adultos implicados y la facilidad con la que se puede acceder a internet hoy en día.

La detección de una posible amenaza hacia un infante podría realizarse de manera manual, donde un padre de familia tenga acceso a todos los mensajes que se comparten en un canal. Sin embargo, esta detección es poco factible ya que el análisis de los mensajes podría ser tardado dependiendo del volumen de la información.

Es por ello que este sistema pretende actuar como una herramienta capaz de automatizar el análisis de dichas conversaciones y con ello hacer uso del sistema implementado.

CAPÍTULO 2

Marco Teórico

2.1. Minería de texto

Definimos minería de texto como un proceso mediante el cual se buscan patrones destacados y nuevos conocimientos en un conjunto de textos, es decir, es el proceso encargado de descubrir conocimientos que no existen tal cual en ningún texto del conjunto seleccionado pero que destacan al relacionar el contenido de varios de ellos [16]. Es decir, búsqueda de regularidades o patrones que se encuentran en un texto, a partir de técnicas de aprendizaje automático.

Existen dos etapas principales dentro de este proceso: preprocesamiento y descubrimiento. Ver 2

Todas las etapas están muy interrelacionadas: la primera etapa condiciona el descubrimiento de los patrones que la minería de texto puede realizar.

El preprocesamiento consiste en transformar el texto en algún tipo de estructura que nos facilite su análisis.El proprocesamiento de textos implica la eliminación de información textual que no será relevante para resolver la finalidad de nuestro proyecto [17]. Incluye la eliminación de palabras que no aportan información, es decir, stop words, así como la unificación de los términos restantes mediante técnicas de stemming. Por otro lado en el descubrimiento se analiza la estructura ateriormente mencionada teniendo como objetivo descubrir patrones interesantes o conocimientos nuevos.

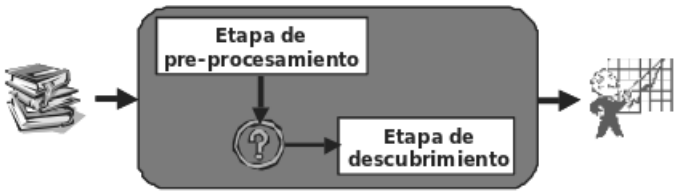

Figura 2.1: Proceso de la Minería de texto

La minería de textos se utiliza principalmente para: extraer información relevante de un documento, agregar y compartir información automáticamente, clasificar y organizar documentos según su contenido, organizar depósitos para búsqueda y recuperación, clasificar textos e indexarlos en la web [18].

En el caso del presente trabajo terminal, se utilizó la minería de datos con el afán de identificar patrones dentro de conversaciones que nos indiquen cuando estas pueden llegar a tener contenido sexual.

2.2. Lenguaje natural

El lenguaje se considera como un mecanismo que nos permite hablar y entender. Los lenguajes naturales, es decir, el inglés, el francés, el español, etc. son una herramienta genuina para la comunicación entre los seres humanos, ya sea en forma oral o escrita. Actualmente, el avance tecnológico en los medios de comunicación impresos y electrónicos nos permite obtener grandes volúmenes de información en forma escrita. La mayoría de esta información se presenta en forma de textos en lenguajes naturales. Toda esa información contenida en los textos es muy importante ya que permite analizar, comparar, entender el entorno en el que vive el ser humano. Sin embargo, se presentan dificultades por la imposibilidad humana de manejar esa enorme cantidad de textos. Entre las herramientas que ayudan en las tareas diarias, la computadora es, hoy en día, una herramienta indispensable para el procesamiento de grandes volúmenes de datos. Pero todavía no se logra que una máquina al capturar una colección de textos los comprenda suficientemente bien; por ejemplo, para que pueda aconsejar qué hacer en determinado momento basándose en toda la información proporcionada, para que pueda responder a preguntas acerca de los temas contenidos en esa información pero no explícitamente descritos, o para que pueda elaborar un resumen de la información. Para lograr esta enorme tarea de procesamiento de lenguaje natural por computadora, analizando oración por oración para obtener el sentido de los textos, es necesario conocer las reglas y los principios bajo los cuales funciona el lenguaje, a fin de reproducirlos y adecuarlos a la computadora, incluyendo posteriormente el procesamiento de lenguaje natural en el proceso general del conocimiento y el razonamiento. El estudio del lenguaje, está relacionado con diversas disciplinas. De entre ellas, la Lingüística General es el estudio teórico que se ocupa de los métodos de investigación y de las cuestiones comunes a las diversas lenguas. Esta disciplina a su vez comprende una multitud de aspectos (temporales, metodológicos, sociales, culturales, de aprendizaje, etc.). Los aspectos metodológicos y de aplicación brindan los principios y las reglas necesarios en el procesamiento de textos. Los principios y las reglas de la lingüística general, aunados a los métodos de la computación forman la Lingüística Computacional. Esta es la área dentro de la cuál se han desarrollado y discutido muchos formalismos adecuados para la computadora a fin de reproducir el funcionamiento del lenguaje con la finalidad de extraer sentido a partir de textos y viceversa, transformando los conceptos de sentidos específicos a los correspondientes textos correctos. [5]

2.3. Procesamiento del lenguaje natural

El término "Procesamiento del lenguaje natural"(PLN) hace referencia a las técnicas de tratamiento del lenguaje y su aplicación en las diversas áreas por medio de métodos computacionales. [3]

El PLN es un área de investigación en continuo desarrollo, es aplicado en la actualidad para la realización de diversas actividades como son: traducción automática, sistemas de recuperación de información, elaboración automática de resúmenes, etcétera. [3]

Si bien es cierto que la evolución de dicha área la posiciona para liderar una nueva dimensión en las aplicaciones informáticas del futuro, la complejidad implícita en el tratamiento del lenguaje tiene sus limitaciones en los resultados.

Podemos definir el PLN como el reconocimiento y utilización de la información expresada en lenguaje humano a través del uso de sistemas informáticos. En su estudio intervienen diferentes disciplinas: la lingüística, ingeniería informática, filosofía, matemáticas y psicología. [3]

Es importante el estudio del lenguaje para determinar el uso que implica éste en diferentes tareas y de esta manera poder moldear el conocimiento de una manera adecuada. [3]

Es necesario tener en cuenta dos puntos: El primero tener en cuenta el problema de representación lingüística para el cual se desea aplicar dicha técnica. Y en segunda el problema de tratamiento mediante recursos informáticos.

2.4. Análisis morfológico

En el español existe una enorme cantidad de palabras que pueden desempeñar diferentes funciones gramaticales. El análisis de un texto producirá una gran multiplicidad de combinaciones posibles.

El análisis morfológico consiste en la detección de la relación que se establece entre las unidades mínimas que forman una palabra: reconocimiento de sufijos o prefijos. Este nivel de análisis mantiene una estrecha relación con el léxico. [3]

El análisis morfológico consiste en determinar la forma, clase o categoría gramatical de cada palabra de una oración. [6]

Sustantivo	Masculino o Femenino, Singular o Plural, Común o Propio, Concreto o Abstracto, Individual o Colectivo. Si lleva Sufijo, puede dar idea de aumento, de disminución o de desprecio (Aumentativo, Diminutivo o Despectivo).
Adjetivo	Masculino o Femenino, Singular o Plural, Determinativo o Calificativo; Demostrativo, Posesivo, Numeral, Relativo o Indefinido. Grado Positivo, Grado Comparativo o Grado Superlativo. Cuando portan Sufijo pueden considerarse como Aumentativos, Diminutivos o Despectivos.
Verbo	Persona, Singular o Plural, Tiempo, Modo, Voz, Conjugación, Forma Personal o No Personal.
Adverbio	Lugar, Modo, Tiempo, Cantidad, Intensidad, etc. Si modifica al verbo, o a otro adverbio.
Pronombre	Masculino o Femenino, Singular o Plural, Determinado o Indeterminado.
Artículo	Lugar, Modo, Tiempo, Cantidad, Intensidad, etc. Si modifica al verbo, o a otro adverbio.
Conjunción	Coordinada o Subordinada; Copulativa, Adversativa, Distributiva o Disyuntiva; Causal, Final, Comparativa, Concesiva, Consecutiva o Condicional.

Tabla 2.1: Categorías de palabras

2.5. Stemming

Frecuentemente se usa una palabra para realizar alguna consulta, pero únicamente una variable de esa palabra está presente en el texto ya sea en singular, plural, gerundio, etcétera. El problema puede ser resulto con la sustitución de una palabra por todas sus formas.

En la mayoría de los casos, las variantes morfológicas de las palabras tienen interpretaciones semánticas similares y se pueden considerar como equivalentes. Por esta razón, se han desarrollado algoritmos de derivación, o analizadores lingüísticos, que tratan de reducir una palabra a su *stem* o raíz. Por lo tanto, los términos clave de una consulta o documento están representados por las raíces en lugar de las palabras originales. Esto no sólo significa que las diferentes variantes de un término pueden confundir a una sola forma representativa, sino que también reduce el tamaño del diccionario, es decir, el número de términos distintos necesario para la

7

representación de un conjunto de documentos. A pequeños resultados del tamaño del diccionario en un ahorro de espacio de almacenamiento y tiempo de procesamiento.

El *Stemming* es un método utilizado para reducir una palabra a su raíz. [8]

2.6. Lematización

Uno de los procesos fundamentales para tratar los textos lingüísticos es la lematización, que consiste en asignar una forma representativa a distintas formas concretas variables: formas conjugadas del verbo, cambios según el género y número de adjetivos y sustantivos, etc. Es necesario a la hora de realizar unos estudios estadísticos, redactar un vocabulario o diccionario, intentar una búsqueda de información por palabras clave y analizar la combinación de elementos, entre otros objetivos de investigación.

Este proceso puede involucrar tareas complejas como entender el contexto de una categoría gramatical de una palabra dentro de una oración.

En muchos idiomas, las palabras aparecen flexionadas de difentes maneras. Por ejemplo en español el verbo 'caminar' puede aparecer como 'camino', 'caminó', 'caminá' 'caminando'. La forma base 'caminar', que puede ser buscada y encontrada en un diccionario, es llamada lema de la palabra. La combinación de la foma base con la flexión es llamada lexema de la palabra.

La importancia de la lematización radica en el hecho que, para acceso por contenido a bases de datos textuales, permite superar las limitaciones de una búsqueda simple de strings, haciendo que relaciones ocultas por la variabilidad morfológica de las palabras queden manifiestas. La lematización mejora por lo tanto el recubrimiento (recall) aunque pueda ser a expensas de la precisión cuando diferentes conjugaciones morfológicas de una misma raiz están asociadas a conceptos distintos.

La lematización está muy relacionada con el etiquetado automático de textos (POS tagging), que consiste en atribuir a cada palabra su categoría gramatical, ya que la categoría puede determinarse por las flexiones o derivaciones (ej: en castellano -ar indica un infinitivo, -ado un participio pasado masculino singular, etc.). Muchos esquemas de procesamiento de textos, aplicados a lenguas flexivas europeas, plantean un etiquetado automático previo a la lematización, de manera que al lematizar se cuente con la información de la categoría gramatical de las palabras. Sin embargo, la atribución de etiquetas correctas depende en general de una lematización implícita basada en un análisis de sufijos y prefijos, lo que permite una primera predicción que se corrige, en una segunda etapa, en función del contexto immediato de la palabra analizada (Brill). Esta manera de proceder presenta algunos problemas: (i) requiere de un corpus manualmente etiquetado de gran dimensión para derivar reglas de etiquetado automático adecuadas, (ii) no aprovecha la existencia de paradigmas de conjugación o derivación, (iii) sólo considera raíces libres. [9]

2.7. Stopwords

Las palabras que aparecen con frecuencia entre los documentos no son buenas para la recuperación de información. Así palabras que aparecen en más del 80 porciento de documentos no son consideradas y se les llama *stopwords* [10] :

- Los artículos, los pronombres, las preposiciones, y las conjunciones son candidatos naturales.
- Algunos verbos, adverbios, y adjetivos se podían tratar como *stopwords*.
- Los términos específicos de un dominio se podían tratar como *stopwords*. Se suele tener una lista de palabras que no son buenos términos de indexación llamada STOPLIST, Lista de Palabras Vacías o Diccionario Negativo. La salida del analizador léxico es comprobada con la STOPLIST y se eliminan los términos que aparecen en ella. También se puede realizar la comprobación durante la etapa del análisis léxico (esto para mejorar el rendimiento) pero no suele ser muy usado en muchos casos.

Ventajas:

- Las palabras vacías aparecen mucho y su lista de referencias es muy grande:
- Si las quitamos el archivo invertido será más pequeño.
- El archivo invertido se reduce en un 30 ó 40 porciento.
- Mejora la eficiencia, porque hay una mejor selección de palabras claves.

- La indexación es más rápida.

Desventajas:

- Por otro lado, la eliminación de *stopwords* puede reducir el recall, lo que hace que sea interesante la indexación del texto completo.

CAPÍTULO 3

Especificaciones del proyecto

3.1. Objetivo

3.1.1. Objetivo general

Desarrollar un sistema que sea capaz de analizar y clasificar texto en diferentes categorías a partir de un conjunto de datos estadísticos mediante: reconocimiento de patrones, aprendizaje máquina y minería de datos. A fin de poder clasificar conversaciones como peligrosas o no peligrosas.

3.1.2. Objetivos específicos

Desarrollar:

- Aplicación para el intercambio de conversaciones.
- API para el análisis de conversaciones.
- Aplicación para el procesamiento de texto.
- Aplicación que funcione, como conexión entre los módulos del sistema.
- Un set de pruebas de desempeño del sistema para verificar su eficiencia.

3.2. Metodología

La metodología usar es desarrollo evolutivo, se basa en la idea de desarrollar una implementación inicial, exponiénndola a los comentarios del usuario y refinándola a través de las diferentes versiones hasta que se desarrolle un sistema adecuado. Las actividades deespecificación, desarrollo y validacińn se entrelazan en vez de separarse, con una rápida retroalimentación entre estas.Existen dos tipos de desarrollo evolutivo: desarrolloexploratorio y prototipos desechables, utilizaremos el desarrollo exploratorio,donde el objetivo del

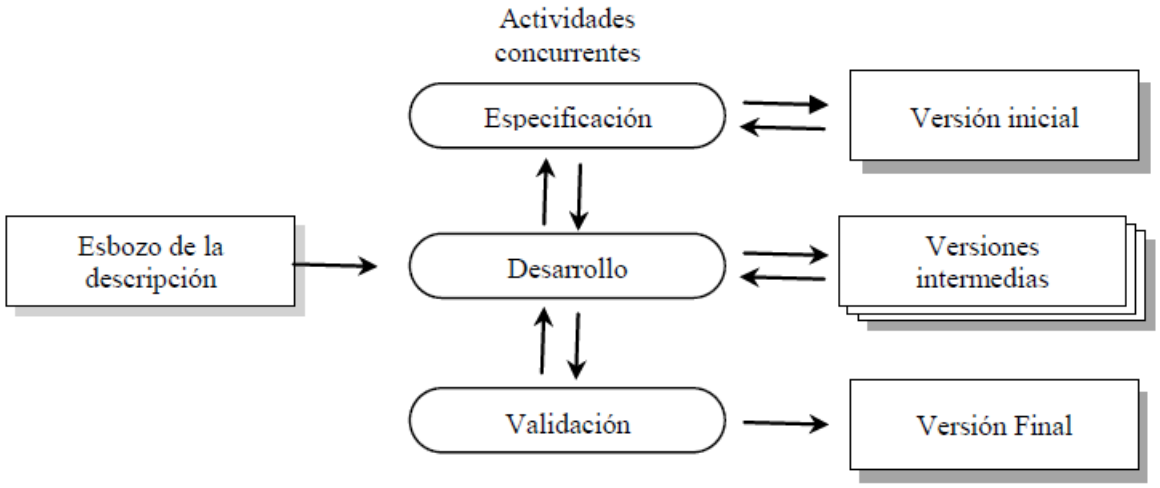

Figura 3.1: Metodología

proceso es trabajar con el usuario (cliente)para explorar sus requerimientos y entregar un sistema final. El sistema evoluciona agregando nuevos atributos propuestos. Ver 3.

3.3. Descripción de prototipos

La tabla 3.1 muestra el nombre de los prototipos y sus características generales.

3.4. Arquitectura del sistema

El sistema contendrá los siguientes modulos:

Sistema de Mensajería Este modulo del sistema será el encargado de generar información de pruebas para posteriormente ser procesada.

Procesamiento de texto Este módulo tiene como objetivo el recibir un archivo de texto con la conversación a analizar el cual fue generado en el módulo de mensajería. La salida de éste será una estructura invariante para poder ser analizada.

API de análisis Será el módulo que mediante reconocimiento de patrones, hará una revisión de los textos para poder clasificar a una fuente en un comportamiento.

Módulo integrador Este módulo es el encargado de presentar de forma visual los resultados obtenidos del proceso de clasificación de conversaciones.

Base de conocimiento Contendrá un conjunto de datos referentes al caso de estudio, que son necesarios para la clasificación de las fuentes.

La figura 3.2 muestra la arquitectura general del sistema de análisis de texto.

12

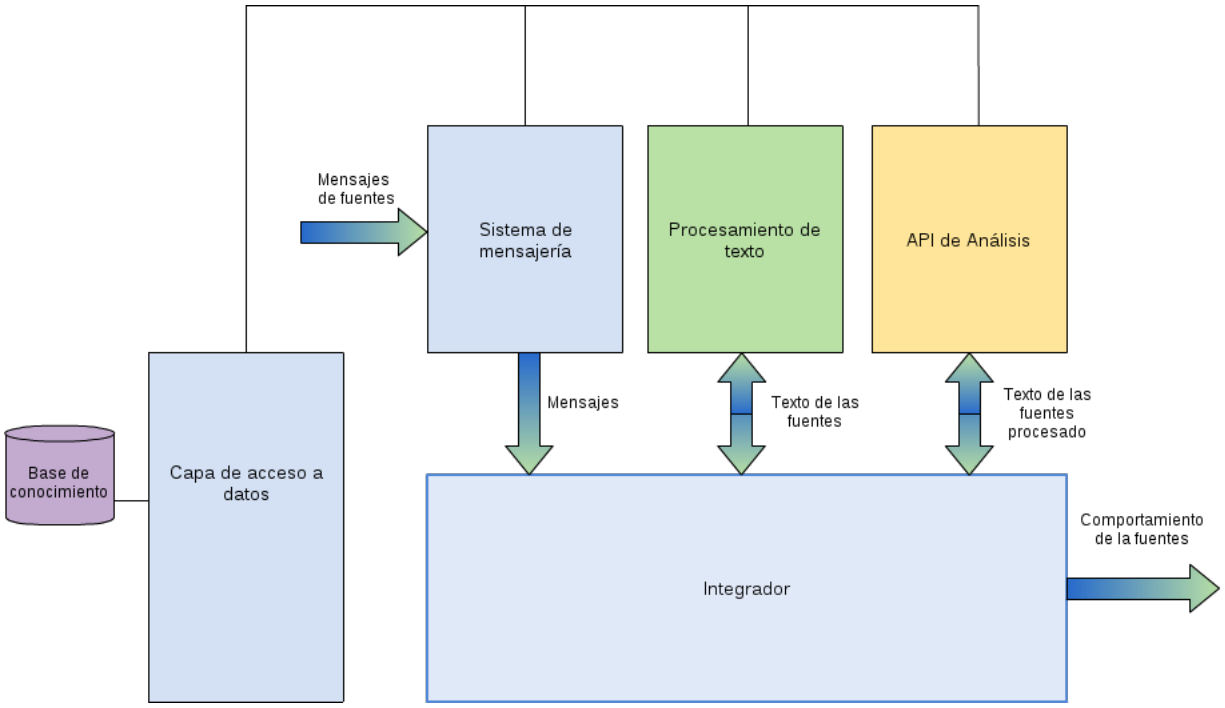

Figura 3.2: Arquitectura general del sistema

13

N	Nombre	Descripción
P1	Identificador y clasificador de frases del Online grooming	• Hacer una clasificación de frases. • Generar vector de clases de frases. • Almacenamiento de vectores.
P2	Generador de Vectores de palabras de carácter sexual	• Extracción de características de carácter sexual. • Generar vectores de frecuencias de nivel 5. • Almacenamiento de vectores.
P3	Clasificador de conversaciones con incidencia en Nivel 5	• Implementación de algoritmos de clasificación. • Entrenamiento de clasificador. • Pruebas del clasificador.
P4	Clasificador de comportamiento de conversaciones	• Análisis de vector de incidencias de Nivel 1, 2, 3, 4 y 6. • Clasificador: Red Neuronal.
P5	API de Análisis	• Integración de clasificadores. • Módulo de decisión.
P6	Sistema de Mensajería	• Sistema que simula conversaciones. • Sistema de pruebas fuera de línea • Integración con la API de análisis.

Tabla 3.1: Prototipos y Funcionalidades Generales

14

Bibliografía

[1] J. Kittler. Reconocimiento de Patrones". Instituto de Ingeniería Eléctrica, notas del curso del Prof. J. Kittler en la Univ. de Surrey N.p., Sep. 2002. Web. 6 Apr. 2014.

[2] Juan Jesús Romero, Carlos Dafonte, Ángel Gómez, Fernando Jorge Penousal. Ïnteligencia Artificial y Computación Avanzada". Colección informática No. 13 N.p., Sep. 2007. Web. 6 Apr. 2014.

[3] Sosa, Eduardo. "Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones (Parte I)". El Profesional de la Información. N.p., Jan. 1997. Web. 6 Apr. 2014.

[4] Flores, Jorge. "Grooming, acoso a menores en la red". Observatorio de la Seguridad de la Información del INTECO N.p., Oct. 2008. Web. 6 Apr. 2014.

[5] "Qué es Procesamiento del Lenguaje Natural". Centro de Investigación en Computación N.p. Web. 6 Apr. 2014.

[6] UNED .Análisis morfológico". Apuntes de grado en informática N.p. Web. 6 Apr. 2014.

[7] .Análisis Sintáctico". Facultad de ingeniería UDB. N.p., Jan. 2013. Web. 6 Apr. 2014.

[8] Ramírez, Kryscia. "Stemming-Lematización". Recuperación de información. N.p., Oct. 2000. Web. 6 Apr. 2014.

[9] Villegas, D. Fernández, E. "Lematizador". N.p., Jan. 2012. Web. 6 Apr. 2014.

[10] .Eliminación de Stopwords". Procesamiento Sobre Texto N.p., Jan. 2012. Web. 6 Apr. 2014.

[11] "Spanish stop word list". Snowball. N.p. Web. 6 Apr. 2014.

[12] Çorpus de Referencia del Español Actual (CREA) - Listado de frecuencias". Web. 6 Apr. 2014.

[13] Internet Grooming, http://www.internet-grooming.net/index.html#quees

[14] Internet Grooming, http://www.perverted-justice.com

[15] Aditi Gupta, Ponnurangam Kumaraguru, Ashish Sureka. Characterizing pedophile conversations on the Internet using Online Grooming Web. Delhi: Indraprastha Institute of Information Technology, 17 Agosto 2012. Disponible en: http://arxiv.org/abs/1208.4324.

[16] Kodratoff (1999), Knowledge Discovery in Texts: A Definition and Applications, Proc. of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99), 1999

[17] Carmen Gálvez,PhD, TEXT-MINING: THE NEW GENERATION OF SCIENTIFIC LITERATURE ANALYSIS IN MOLECULAR BIOLOGY AND GENOMICS

[18] Bressán G. 2003. Almacenes de datos y minería de datos. 2003.