Prototipo de motor de búsqueda para la detección de tweets con semántica violenta

Trabajo Terminal No. 2019-A038

*Alumnos: Maldonado Ledo Diana Guadalupe**, Tule Uscanga Carlos Enrique**
Directores: Ferrer Tenorio Jorge, Acosta Bermejo Raúl.
e-mail: carlos.enrique.tule@gmail.com, dianagml24@gmail.com***

Resumen – Este protocolo plantea el desarrollo de un sistema de búsqueda que filtre de manera efectiva los tweets con tendencias violentas, involucradas en la semántica de las palabras, mediante el entrenamiento de un algoritmo, para que detecte de manera eficaz los usuarios con estas tendencias. Con el propósito de otorgar una herramienta de apoyo a las instituciones de seguridad pública u otras instancias dentro del mismo rubro, para la captura de presuntos criminales.

Palabras clave – *Detección de perfiles, Red social, Semántica de las palabras, Tendencia violenta, Twitter.*

1. Introducción

Hoy en día se mandan más de 500 millones de tweets alrededor del mundo, y el 71% de los usuarios se informa sobre las noticias más recientes mediante esta medio, por su parte en México aproximadamente 5.77 millones de personas tienen una cuenta en esta red social. [1]
Las redes sociales se crearon con el fin de poner interconectar el mundo, y crear una infinidad de relaciones y, es por ello que han ido adaptándose a las necesidades tecnológicas de la sociedad.

La importancia de las redes sociales radica principalmente en la difusión, la libertad, velocidad y versatilidad, pues es un puente entre espacio tiempo. Sin embargo, las redes sociales pueden desarrollar en los usuarios una adicción que se caracteriza por:

1. Es dominado en sus pensamientos, sentimientos y conducta por su uso (saliencia).
2. Invierte grandes cantidades de tiempo y esfuerzo en la actividad e incremento de ésta.
3. Altera sus estados emocionales (ansiedad, enojo) como consecuencia de implicarse en la actividad (modificación del humor).
4. Se perturba cuando es interrumpido en la actividad o se le reduce el acceso (abstinencia).
5. Reanuda la actividad de manera persistente una vez que, aparentemente, la ha dejado o la ha reducido (recaída).
6. Niega tener consecuencias por la actividad.[2]

La identidad de los usuarios en redes sociales no siempre es en su totalidad verificable, la identidad en medios sociales tiende a ser abierta y múltiple, pues el contacto directo no existe y el manejo de las palabras es factor fundamental para obtener nociones con respecto a la veracidad

Ahora bien, las teorías que establecen esta clase de vínculo entre léxico y sintaxis suelen declarar siempre su validez psicológica, en el sentido en que adjudican estatus de realidad a las descripciones y explicaciones: los modos de operar del lenguaje tendrían correspondencia con modos de operación de la mente humana. [3]
Es por esta manera en que la mente interpreta las palabras y el significado de acuerdo a la selección semántica que se muestra dentro de las redes sociales, afectando directamente a los usuarios,

SOFTWARE	Búsqueda de usuarios	Motor Gratuito	Filtración de mensajes	Respuesta Fiable	Uso de Machine Learning
Sherlock	✓	✓	X	✓	X
Motor de Búsqueda Propuesta	✓	✓	✓	✓	✓

Tabla 1. Resumen de productos similares.

2. Objetivo

General

Desarrollar un prototipo de motor de búsqueda que permita detectar los tweets con semántica violenta por medio de un análisis, que permita ubicar a los usuarios que utilizan un lenguaje violento en sus publicaciones.

Específicos

1. Crear el motor de búsqueda de para la red social Twitter, que permita el análisis de la información pública, así como la contemplación de diferentes tipos de perfiles.
2. Integrar al motor de búsqueda la parte de análisis y clasificación de tweets haciendo uso de Machine Learning
3. Analizar la semántica de los tweets para verificar si los mensajes tienen una semántica violenta.

3. Justificación

El hombre siempre ha tenido la necesidad de hablar y expresarse; buscar, analizar y expresar la información, es un rasgo de la naturaleza muy importante para la vida en sociedad.
Las palabras que usamos transmiten una gran cantidad de información acerca de quiénes somos, a quién nos dirigimos, y las situaciones en las que estamos. [4]

Es gracias a las palabras que se pueden transmitir ideas y reflejar la forma de pensar de cada individuo Tienen un impacto en nuestro estado de ánimo y, por la conexión que existe entre lenguaje y pensamiento, se ven reflejados en la forma de expresarse dentro del entorno.

Actualmente, el ambiente de las redes sociales se ha vuelto parte fundamental de la comunicación, pues además de acercarnos a nuevos lugares y/o personas, la libertad de expresión es más notoria y, desde luego más influyente, es por ellos que se puede decir que las palabras que usamos se filtran en nuestra mente y afectan nuestras acciones.

Las redes sociales representan un espacio de libertad donde las personas pasan a mostrar rasgos que no expresan de manera presencial, pero por ser un entorno totalmente público, el uso de las palabras, fotos, etc. se vuelve parte del mundo virtual en el cual ya no hay un control, y en el cual fluye la información para aspectos tanto negativos, como la confusión entre lo público y lo privado, así también como positivos.

La comunicación digital es entonces, una forma en que las relaciones y la comunicación van cambiando y se van adoptando nuevas formas de pensamiento, provocando entonces un cambio en las acciones y en la percepción del mundo que nos rodea. Y dichas formas de pensamiento se van aplicando en acciones negativas, que afectan tanto a los niños, como adolescentes y gente adulta, que están involucradas en las redes sociales, ya sea por medio de ataques como lo son el bullying, acoso, violencia psicológica, entre otros; afectando directamente el autoestima y la seguridad de las víctimas.

4. Productos o Resultados esperados

Los resultados esperados del desarrollo del proyecto son:

1. Prototipo del motor de búsqueda (software):

1.1. Módulo de búsqueda.
1.2. Módulo de análisis de tweets. (Machine Learning)
1.3. Módulo de clasificación de tweets.
1.4. Base de datos
1.5. Interfaz de usuario

2. Manual Técnico
3. Manual de Usuario

El prototipo de motor de búsqueda a desarrollar permitirá ingresar un nombre de usuario, lo cual arrojará los usuarios que se relacionen a la búsqueda y así obtener los tweets que han hecho. Posteriormente se hará un análisis de la semántica que tienen los tweets y se hará una clasificación de ellos para detectar cuales tienen contenido violento y saber que usuarios son los que lo crean.

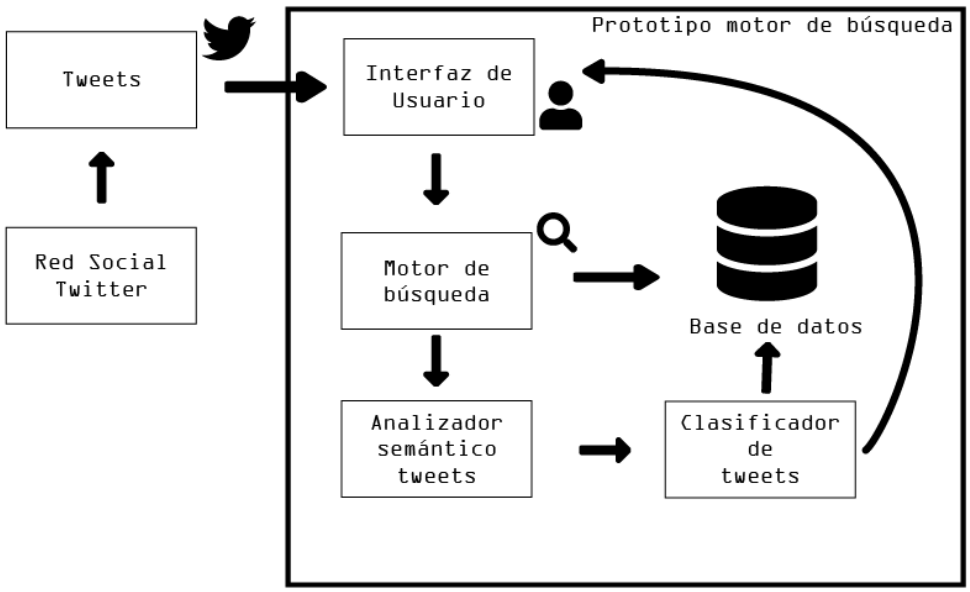


Figura 1. Arquitectura de prototipo de motor de búsqueda.
Fuente: Creación Propia

5. Metodología

El proyecto requiere de un desarrollo en donde cada módulo se realice de manera paralela y de esta manera integrar cada parte al sistema final. Por tal motivo la metodología que se escogió fue la incremental, ya que nos permite hacer desarrollos independientes y así obtener incrementos funcionales de forma gradual para obtener un producto funcional. La metodología tiene 4 etapas:

I. **Análisis:** Fase más importante de esta metodología, en esta etapa se definirán todos los requerimientos funcionales y no funcionales junto con los casos de uso del proyecto y las reglas del negocio,
II. **Diseño:** En esta etapa, se diseñará el sistema con la respectiva documentación de cada incremento que especifica y describe la estructura del software del sistema.
III. **Código:** En esta etapa, se codificaron todas las funcionalidades correspondientes al incremento en curso.
IV. **Pruebas:** durante esta etapa se realizarán una serie de ensayos que permitan verificar el correcto funcionamiento del incremento realizado

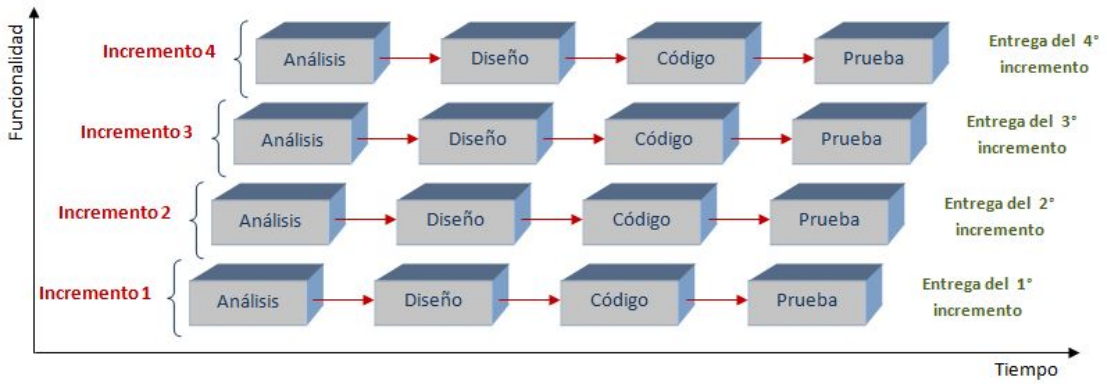


Figura 2. Metodología Incremental.

El proyecto utilizará el modelo cliente servidor, ya que el usuario del sistema hará peticiones al servidor en el momento que realice la búsqueda, y el servidor realizará el análisis correspondiente para arrojar los usuarios que tengan tweets con contenido violento.

El prototipo se realizará en 4 incrementos:

1. Módulo de búsqueda de usuario
2. Módulo de análisis de tweets
3. Módulo de clasificación de tweets
4. Interfaz de usuario

6. Cronograma

Anexo 1.

7. Referencias

[1] STATISTA, Leading Countries based of number of Twitter users as of Janueay 2019, , [Online] Disponible en: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[2] Fernández Sánchez, Néstor. (2013), Trastornos de conducta y redes sociales en Internet. *Salud Ment* [Online] Vol.36, Núm.6, Págs. 521-527, Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-33252013000600010...Pág. 229, [Online] Disponible en: https://www.ijf.cjf.gob.mx/publicaciones/revista/28/Delitos_inform%C3%A1ticos.pdf

[3] Cárdenas, Viviana. (2010), *La Relación entre semántica y sintaxis desde la perspectiva de la producción del lenguaje escrito*, [Online] Universidad Nacional de Salta, Argentina, Pág. 246, Disponible en: http://www.scielo.org.mx/pdf/tods/n23/n23a8.pdf.

[4] Ramírez-Esparza, Nairan. (2007, jun), "La Psicología del uso de las palabras: Un programa de computadora que analiza textos en español", *Revista Mexicana de Psicología*, [Internet], Núm.24, [Disponible en] https://www.redalyc.org/pdf/2430/243020635010.pdf .

Nombre del alumno(a): ____Tule Uscanga Carlos Enrique_____ TT No.:
Título del TT: Prototipo de motor de búsqueda para la detección de tweets con semántica violenta.

Actividad	JUL	AGO	SEP	OCT	NOV	DIC	ENE	FEB	MAR	ABR	MAY
Investigación Machine Learning.											
Desarrollo del Módulo 1 (Búsqueda).											
Toma de requerimientos no funcionales.											
Desarrollo de Diagramas de casos de uso.											
Implementación del Módulo 1.											
Corrección de errores.											
Desarrollo del Módulo 2 (Análisis).											
Desarrollo de Machine Learning para análisis.											
Pruebas Modulo 2.											
Evaluación TT II.											
Desarrollo del Módulo 3 (Clasificación).											
Implementación del módulo 3.											
Corrección de errores.											
Desarrollo del Módulo 4 (Interfaz de usuario).											
Integración con módulos 1, 2 y 3.											
Corrección de errores.											
Evaluación TT II.											

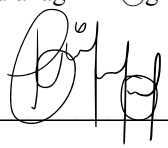
Nombre del alumno(a): __Maldonado Ledo Diana Guadalupe_____ TT No.:
Título del TT: Prototipo de motor de búsqueda para la detección de tweets con semántica violenta.

Actividad	JUL	AGO	SEP	OCT	NOV	DIC	ENE	FEB	MAR	ABR	MAY
Investigación semántica y uso de las palabras											
Desarrollo del Módulo 1 (Búsqueda)											
Toma de requerimientos funcionales											
Desarrollo de casos de uso y reglas de negocio											
Desarrollo de la base de datos											
Pruebas del Módulo 1											
Desarrollo del Módulo 2 (Análisis)											
Definición de las reglas semánticas para análisis											
Implementación Módulo 2											
Evaluación TT I											
Desarrollo del Módulo 3 (Clasificación)											
Definición de la clasificación de tweets mediante semántica											
Pruebas de Módulo 3											
Desarrollo del Módulo 4 (Interfaz de usuario)											
Diseño de Interfaz de usuario.											
Implementación de la interfaz de usuario											
Pruebas											
Manuales (Usuario y Técnico)											
Evaluación TT II.											

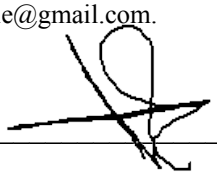
8. Alumnos y Directores

CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Art. 3, fracc. II, Art. 18, fracc. II y Art. 21, lineamiento 32, fracc. XVII de la L.F.T.A.I.P.G.
PARTES CONFIDENCIALES: No. de boleta y Teléfono.

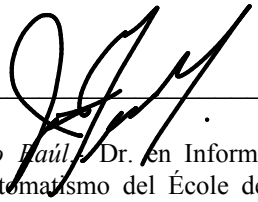
Maldonado Ledo Diana Guadalupe.- Alumno de la carrera de Ing. en Sistemas Computacionales en ESCOM, Especialidad Sistemas, Boleta: 2014630280, Tel. 5518127783, email dianagml24@gmail.com.

Firma: 

Tule Uscanga Carlos Enrique.- Alumno de la carrera de Ing. en Sistemas Computacionales en ESCOM, Especialidad Sistemas, Boleta: 2012630449, Tel. 5563408198, email carlos.enrique.tule@gmail.com.

Firma: 

Ferrer Tenorio Jorge.- M en C. Estudios Latinoamericanos por parte de la UNAM- FFL, Tel. 5729 6000 Ext. 52070 Profesor de ESCOM/IPN (Dpto de Formación Integral e Institucional) desde 1999, Áreas de Interés: MRS email: jorgeferrert@gmail.com

Firma: 

Acosta Bermejo Raúl.- Dr. en Informática, Tiempo Real, Robótica y Automatismo del École de Mines de Pais en 2003, M. en C. en Computación del CINVESTAV en 1997, Ing. en Electrónica de la UAM en 1993, Profesor de ESCOM/IPN (Dpto de Posgrado) desde 1993, Áreas de Interés: MRS, Redes. Ext. 52028, email racostab@ipn.mx.

Firma: