Detector de texto ofensivo durante la navegación Web
Trabajo Terminal No. — — — — - — — —

*Alumnos: *Esquivel Salvatti José Luis, López Morales Miguel Angel, López Felipe Eliel*
Directores: Luna Benoso Bejamin, Edgardo Adrian Franco Martinez
**jesquivels1400@alumno.ipn.mx*

Resumen - La cantidad de información disponible en internet crece constantemente lo que algunas veces expone a los usuarios información que contiene lenguaje ofensivo que es el mismo que se usa para humillar o herir la integridad de las personas por lo que en el proyecto actual se desarrollará una extensión web y un software alojado en un servidor que utilice procesamiento de lenguaje natural y técnicas de inteligencia artificial con el objetivo de detectar lenguaje ofensivo en la navegación web de los jóvenes de entre 15 y 17 años de edad y notificar al tutor encargado de dichos jóvenes.

Palabras clave – Análisis de texto, control parental, lenguaje ofensivo.

1. Introducción

La creciente relación humana con los medios digitales nos ha llevado a una dependencia de esta tal que su uso no se ha limitado por edades dado que en la actualidad es necesario en algunos casos que los niños hagan uso de estas herramientas, dejándolos expuestos a un posible encuentro con lenguaje ofensivo en modo texto o en su defecto manejar lenguaje ofensivo [1].

Para entender qué es el lenguaje ofensivo se debe revisar antes la definición de sus componentes, por lo tanto, según definiciones de la RAE el lenguaje es la facultad del ser humano de expresarse y comunicarse con los demás a través del sonido articulado o de otros sistemas de signos, es un sistema de comunicación que tiene como característica un estilo y modo de hablar y escribir de cada persona en particular [2]. Mientras que la definición de un acto ofensivo sería todo aquel que tenga como objetivo humillar o herir el amor propio o la dignidad de alguien, puede considerarse también actos donde se dañen físicamente o maltraten a otra persona [3]. Por lo tanto, en este trabajo el lenguaje ofensivo se definirá como “todas aquellas expresiones, palabras o texto que sean discriminatorias, despectivas y que tengan como propósito dañar a otra persona o grupo” [2,3].

Autores como Marc Prensky sostienen que los cambios presentados en la sociedad debido al uso de medios digitales hace que sea necesario tener un mejor entendimiento de las tecnologías pues sus mayores usuarios son los llamados nativos digitales, una parte de la población que ha nacido en la era digital o en otras palabras a los jóvenes y niños. Si bien existe un debate por los años en los que podemos considerar a una persona como un nativo digital o inmigrante digital (adulto nacido antes de los 1980) el hecho es que no existe ningún nativo digital que naciera “digital” sino que se hacen por el uso de las tecnologías, por lo tanto aunque sabemos que los jóvenes actualmente establecen relaciones y amistades en un espacio digital, que se manejan de manera competente en la navegación web y que los métodos educativos deben adaptarse a ellos y a la tecnología para aprovechar al máximo sus cualidades no podemos negar que nadie nace con criterios y habilidades para la selección y filtro de la búsqueda ni el procesamiento de la información, no nace entendiendo el uso ético y seguro de las tecnologías. De esto sabe más el adulto, que tiene mayor capacidad y experiencia para seleccionar y discriminar entre lo válido, lo inútil y lo pernicioso, aunque este adulto (docentes, tutores o padres, por ejemplo) carezca de ciertas destrezas técnicas [4].

Prensky también sostiene que en la educación actual a los jóvenes de 15 a 17 se les exige asumir un rol de investigador, usuario y experto de tecnología para completar su desarrollo formativo, lo que significa que se les pide un uso constante y responsable de herramientas tecnológicas como el internet donde por el exceso de información es difícil mantener una verificación de la información. Esto significa que los jóvenes al vivir gran parte de su vida en línea están expuestos a una gran cantidad de información que se podría considerar perjudicial para ellos por lo que se decidió enfocarse a este sector de la población [5].

Las extensiones web (también conocidos como add-ons) son aquellos aditamentos que se pueden agregar en los navegadores web (Google Chrome, Mozilla Firefox, entre otros) y que permiten realizar actividades en específico por lo que es posible interactuar con el contenido del HTML[6], de esta forma se puede saber el contenido de la página y hace posible el análisis para detectar el uso de lenguaje ofensivo. Para realizar este análisis se desarrollará un software que estará alojado en un servidor, el cual se comunicará con la extensión para obtener el contenido de las páginas visitadas durante la navegación mientras la extensión esté activa.

El sistema detecta el contenido de la navegación de los tutorados y notifica al tutor si se trata o hace uso de lenguaje ofensivo. Nuestro sistema se une a los diversos sistemas que tienen como objetivo detectar un acto ofensivo o discriminatorio los cuales se enfocan en detectar actos como el bullying, el acoso, el sexismo, racismo o algún otro acto ofensivo o denigrante. La diferencia sustancial es el método empleado para detectar su tema, en nuestro caso el lenguaje ofensivo es detectado por medio del análisis del texto de la página web, otra diferencia es el acto posterior al detectarlo pues lo que se realiza en nuestro sistema es enviar una notificación al tutor para que su criterio sea usado para juzgar el contenido visitado. De esta manera actos como el acosar o ser acosado en redes sociales, seguir discursos de odio, presenciar casos de racismo o participar en ellos, acceder a contenido para adultos, entre otras actividades podrían ser detectadas, prevenidas y discutidas por los involucrados (tutor y tutorado) para llegar a la mejor solución posible sin invadir la navegación o censurar el contenido directamente [7].

Se han encontrado los siguientes sistemas o aplicaciones con características u objetivos relacionados con nuestro proyecto :

Tipo	**Proyecto**	**Descripción**	**Características**	**Comparación con nuestra propuesta**
Tesis (Montoro Montarroso A. 2019, España).	Análisis De Sentimientos Para La Prevención De Mensajes De Odio En Las Redes Sociales.	Aplicación web que identifica y clasifica mensajes de comunicación violenta y de odio [8].	•Análisis de sentimientos. •Procesamiento de lenguaje natural. •Identificación y clasificación de mensajes de odio en las redes sociales. •Software de escritorio.	La aplicación utiliza el procesamiento de lenguaje natural para detectar los mensajes de odio en redes sociales, mientras que nuestra propuesta abarcaría toda la web que pueda ser vista por un navegador.
Boletín (Jara D., Ramírez S., Dulce M. y Reyes C. 2018, Colombia).	Termómetro de violencia digital.	Esquema de aprendizaje supervisado para clasificar todos los mensajes y usuarios según su toxicidad y su provocación [9].	•Análisis de sentimientos. •Aplicación web para análisis de modelo matemático. •Análisis estadístico de la violencia digital en redes sociales. •Machine Learning	En este trabajo se enfocan en utilizar modelos matemáticos y machine learning para clasificar los mensajes por su toxicidad, nuestra propuesta tiene como opción crear un modelo con inteligencia artificial para detectar lenguaje ofensivo.

Tesis (Guzmán Falcón, E. 2018, México).	Detección de lenguaje ofensivo en Twitter basada en expansión automática de lexicones.	Etiquetado automático y un enfoque basado en aprendizaje el cual se encarga de identificar mensajes ofensivos en función de lo aprendido con los datos etiquetados automáticamente [10].	•Detección de lenguaje ofensivo basada en expansión automática de lexicones. •Orientada en tuits. •Machine learning. •Software de escritorio.	El trabajo detecta lenguaje ofensivo con machine learning y un software de escritorio, mientras nuestra propuesta sería un servicio web.
Artículo científico (Arcila C., Blanco-Herrero, D. y Valdez M. 2020, España).	Rechazo y discurso de odio en Twitter: análisis de contenido de los tuits sobre migrantes y refugiados en español.	Se analiza el rechazo verbal al extranjero como potencial detector de discurso de odio a través de dos análisis de contenido de tuits en español recogidos con la API de Twitter [11].	•Técnicas de big data. •Detección de discurso de odio. •Orientada en tuits. •Software de escritorio.	Es un sistema que se dedica a detectar el lenguaje ofensivo, pero solo se limita a detección en relación a contenido sobre migrantes y refugiados en Twitter, mientras que nuestro sistema tendría como objetivo todo el texto en español de la web.

Tabla 1. *Resumen de proyectos similares.*

2. Objetivo

Objetivo general:

- Desarrollar un software que permita la identificación de lenguaje ofensivo de páginas web y notificar a los tutores sobre el contenido que visitan los jóvenes.

Objetivos específicos:

- Implementar técnicas de inteligencia artificial para la detección del lenguaje ofensivo.
- Desarrollar una extensión web que se agregue al navegador para la detección del lenguaje ofensivo.
- Configurar un servidor para la detección del lenguaje ofensivo.
- Realizar pruebas con diferentes navegadores web que permitan el uso de extensión.
- Documentar el proceso de desarrollo del software y los resultados obtenidos.

3. Justificación

Analizando el comportamiento, según datos del INEGI, de los distintos grupos de edad de la población total, el que concentra la mayor proporción de usuarios de internet en México, respecto al total de cada grupo de edad, es el grupo de 18 a 24 años con una participación de 90.5% del total de la población. El segundo grupo de edad donde el uso de internet está más generalizado, es el de 12 a 17 años, con un aproximado de 12.2 millones de usuarios, equivalente al 90.2% del total de la población, grupo en el que se incluye el rango de edades de nuestro proyecto [12].

De acuerdo con los resultados de la Encuesta Nacional sobre Disponibilidad y Uso de las Tecnologías de la Información en los Hogares (ENDUTIH) 2020, en México, 75% de la población de 12 años y más utilizó Internet en cualquier dispositivo en el periodo comprendido entre julio y noviembre de 2020. De esta población el 21% de ellos declaró haber vivido, entre octubre de 2019 y noviembre de 2020, alguna situación de acoso cibernético por las que se indagó, siendo mayor para mujeres (22.5%) que para los hombres (19.3%). Los adolescentes y jóvenes son los más expuestos: 23.3% de los hombres de 20 a 29 años y 29.2% de las mujeres de 12 a 19 años, señalaron haber vivido algún tipo de ciberacoso.

Las situaciones experimentadas con mayor frecuencia por parte de la población de mujeres que ha vivido ciberacoso fueron: recibir insinuaciones o propuestas sexuales (35.9%), contacto mediante identidades falsas (33.4%) y recibir mensajes ofensivos (32.8%); mientras que para la población de hombres que han vivido ciberacoso fueron: contacto mediante identidades falsas (37.1%), recibir mensajes ofensivos (36.9%) y recibir llamadas ofensivas (23.7%) [13].

Nuestra propuesta de aplicación potencialmente beneficiaría a los tutores que estén interesados en utilizar nuestro programa para evitar la exposición del lenguaje ofensivo a sus hijos, ya que nuestra aplicación ayudaría a detectar el lenguaje ofensivo en las páginas visitadas, con lo cual se pretende que este ayude a mejorar el contenido al que están expuestos, al poder notificar al tutor y que este tome las acciones pertinentes.

La complejidad que tiene desarrollar un proyecto como este radica en las diversas áreas con las cuales debemos estar familiarizados para hacer la correcta investigación y posterior implementación del proyecto, siendo algunos de los conocimientos necesarios el análisis del lenguaje, reconocimiento de patrones, probabilidad y el análisis de textos. El lenguaje tiende a expandirse cada vez más por la exposición de factores como la interacción con otros idiomas, aparición de nuevas expresiones. Así mismo, en la actualidad dichos cambios se han acelerado por la exposición cada vez más frecuente a internet, por lo tanto un entendimiento total de este es una tarea que resulta casi imposible, lo que eleva la complejidad de la óptima identificación del significado de las palabras y por lo tanto del proyecto mismo.

4. Productos o Resultados esperados

Una extensión web capaz de envíar el contenido y el código fuente de una página web a un servidor que tiene montado un software para buscar contenido con lenguaje ofensivo con la finalidad de que los padres o tutores tengan conocimiento de la navegación realizada por sus hijos. En caso de detectar lenguaje ofensivo en el contenido se notificará a ambas partes, tanto el tutor por medio de correo electrónico como al usuario navegando en la web por medio de la extensión.

A continuación se muestra el diagrama del sistema esperado.

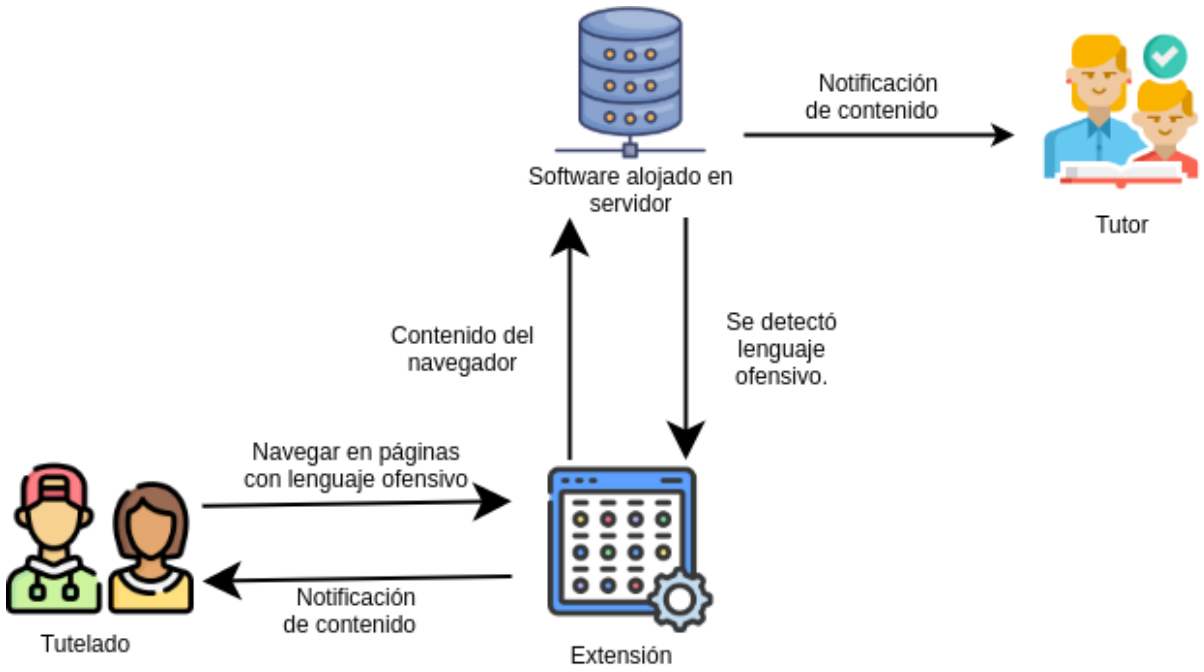


Figura 1. *Arquitectura del sistema (imágenes disponibles en flaticon.es)*

Así mismo se hará entrega de:

1. El código.
2. La documentación técnica del sistema.
3. El manual de usuario.

5. Metodología

Tomando en cuenta el número de integrantes, el plazo de tiempo que tenemos para hacer este proyecto y las distintas formas en las que podemos crearlo se tomó la decisión de utilizar la metodología ágil SCRUM, con la cual podremos empezar a trabajar con un diseño simple y separar el plan de trabajo en módulos, consiguiendo mayor flexibilidad, concretarlos requisitos del sistema y retroalimentación en distintos puntos de tiempo en la creación del proyecto.

Debido a que usando la metodología ágil el producto evoluciona basándose en bucles de retroalimentación buscando llegar al objetivo inicial, aunque esto no significa que no pueda recibir modificaciones en la dirección, ya que responde a las necesidades presentadas durante los periodos de trabajo o sprints. Con cada periodo se busca terminar una serie objetivos menores lo que nos permite entregar avances y recibir retroalimentación suficiente para marcar prioridades o introducir cambios en la aplicación [14].

6. Cronograma

Cronograma: José Luis Esquivel Salvatti
Título del TT: Detector de texto ofensivo durante la navegación Web
TT No:

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Sprint diario.											
Planificación del proyecto.											
Análisis del entorno.											
Diseño del sistema.											
Creación del product backlog											
Definición e implementación de Sprint 1											
Definición e implementación de Sprint 2											
Definición e implementación de Sprint 3											

Evaluación de TT I											
Definición e implementación de Sprint 4											
Definición e implementación de Sprint 5											
Definición e implementación de Sprint 6											
Definición e implementación de Sprint 7											
Codificación de los componentes del sistema.											
Integración y pruebas.											
Generación del reporte técnico											
Evaluación de TT II											

Cronograma: Miguel Ángel López Morales
Título del TT: Detector de texto ofensivo durante la navegación Web
TT No:

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Sprint diario.											
Planificación del proyecto.											
Análisis del entorno.											
Diseño del sistema.											
Análisis y desarrollo de la documentación.											
Creación o búsqueda de un corpus de lenguaje ofensivo											
Definición e implementación de Sprint 1											
Definición e implementación de Sprint 2											
Definición e implementación de Sprint 3											

Evaluación de TT I											
Definición e implementación de Sprint 4											
Definición e implementación de Sprint 5											
Definición e implementación de Sprint 6											
Definición e implementación de Sprint 7											
Codificación de los componentes del sistema.											
Integración y pruebas.											
Generación de manual de usuario											
Evaluación de TT II											

Cronograma: Eliel López Felipe
Título del TT: Detector de texto ofensivo durante la navegación Web
TT No:

Actividad	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
Sprint diario.											
Planificación del proyecto.											
Análisis del entorno.											
Selección y configuración de herramientas para el entorno de desarrollo.											
Análisis y desarrollo de la documentación.											
Definición e implementación de Sprint 1											
Definición e implementación de Sprint 2											
Definición e implementación de Sprint 3											

Evaluación de TT I											
Definición e implementación de Sprint 4											
Definición e implementación de Sprint 5											
Definición e implementación de Sprint 6											
Definición e implementación de Sprint 7											
Codificación de los componentes del sistema.											
Generación del reporte técnico											
Generación de manual de usuario											
Evaluación de TT II											

7. Referencias

[1] Ortiz-Ospina, E. (s. f.). Internet statics. Our World in Data. https://ourworldindata.org/internet
[2] Real Academia Española. (s. f.). *Lenguaje* | *Diccionario de la lengua española*. «Diccionario de la lengua española» - Edición del Tricentenario. https://dle.rae.es/lenguaje
[3] Real Academia Española. (s. f.-b). ofender | Diccionario de la lengua española. «Diccionario de la lengua española» - Edición del Tricentenario. https://dle.rae.es/ofender
[4] García Aretio, L. (2019). Necesidad de una educación digital en un mundo digital. RIED. Revista Iberoamericana de Educación a Distancia, 22(2), 9. doi: https://doi.org/10.5944/ried.22.2.23911
[5] Prensky, M. (2015). Enseñar a nativos digitales. Ediciones SM.
[6] ¿Qué son las extensiones? - Mozilla | MDN. (s. f.). MDN Web Docs. https://developer.mozilla.org/es/docs/Mozilla/Add-ons/WebExtensions/What_are_WebExtensions
[7] Los 7 principales riesgos de Internet para los niños. (s. f.). EMY Cursos en el Extranjero. https://www.emy.org/7-riesgos-que-tiene-internet-para-los-ninos
[8] Montoro Montarroso, A. (2019). Análisis De Sentimientos Para La Prevención De Mensajes De Odio En Las Redes Sociales [Trabajo de Grado, Universidad de Castilla]. TFG (uclm.es)
[9] Jara, D., Ramírez, S., Dulce, M. y Reyes, C. (2018). Termómetro de Violencia Digital. Quantil. https://quantil.co/wp-content/uploads/2018/09/Boletín-3.-MINTIC.-Abril-2018.pdf
[10] Guzmán Falcón, E. (2018). Detección de lenguaje ofensivo en Twitter basada en expansión automática de lexicones [Trabajo de Grado, Instituto Nacional de Astrofísica, Óptica y Electrónica]. https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/1722/1/GuzmanFE.pdf
[11] Arcila Calderón C., Blanco-Herrero D. y Valdez Apolo María Belén (2020). Rechazo y discurso de odio en Twitter: análisis de contenido de los tuits sobre migrantes y refugiados en español. Revista Española de Investigaciones Sociológicas, 172: 21-40. http://dx.doi.org/10.5477/cis/reis.172.21
[12] INEGI. (2021). En México hay 84.1 millones de usuarios de internet y 88.2 millones de usuarios de teléfonos celulares: Endutih 2020. https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/OtrTemEcon/ENDUTIH_2020.pdf
[13] INEGI. (2020, 5 de julio). Módulo sobre ciberacoso 2020. Instituto Nacional de Estadística y Geografía (INEGI). https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/EstSociodemo/MOCIBA-2020.pdf
[14] Rubin, K. S. (2012). Essential Scrum: A practical guide to the most popular agile process. Addison-Wesley.
[15] Alfonso, Pedro Luis, & Mariño, Sonia, & Godoy, María Viviana (2011). Propuesta metodológica para la gestión de proyecto de software ágil basado en la Web. Multiciencias, 11(4),395-401.[fecha de Consulta 20 de Octubre de 2021]. ISSN: 1317-2255. Disponible en: https://www.redalyc.org/articulo.oa?id=90421972009

8. Alumnos y Directores

Miguel Ángel López Morales.- Alumno de la carrera de
Ingeniería en Sistemas Computacionales en ESCOM
Especialidad Sistemas, Boleta: 2014030740,
Tel. 5567538689, email: mlopezm1305@alumno.ipn.mx
Firma:_____

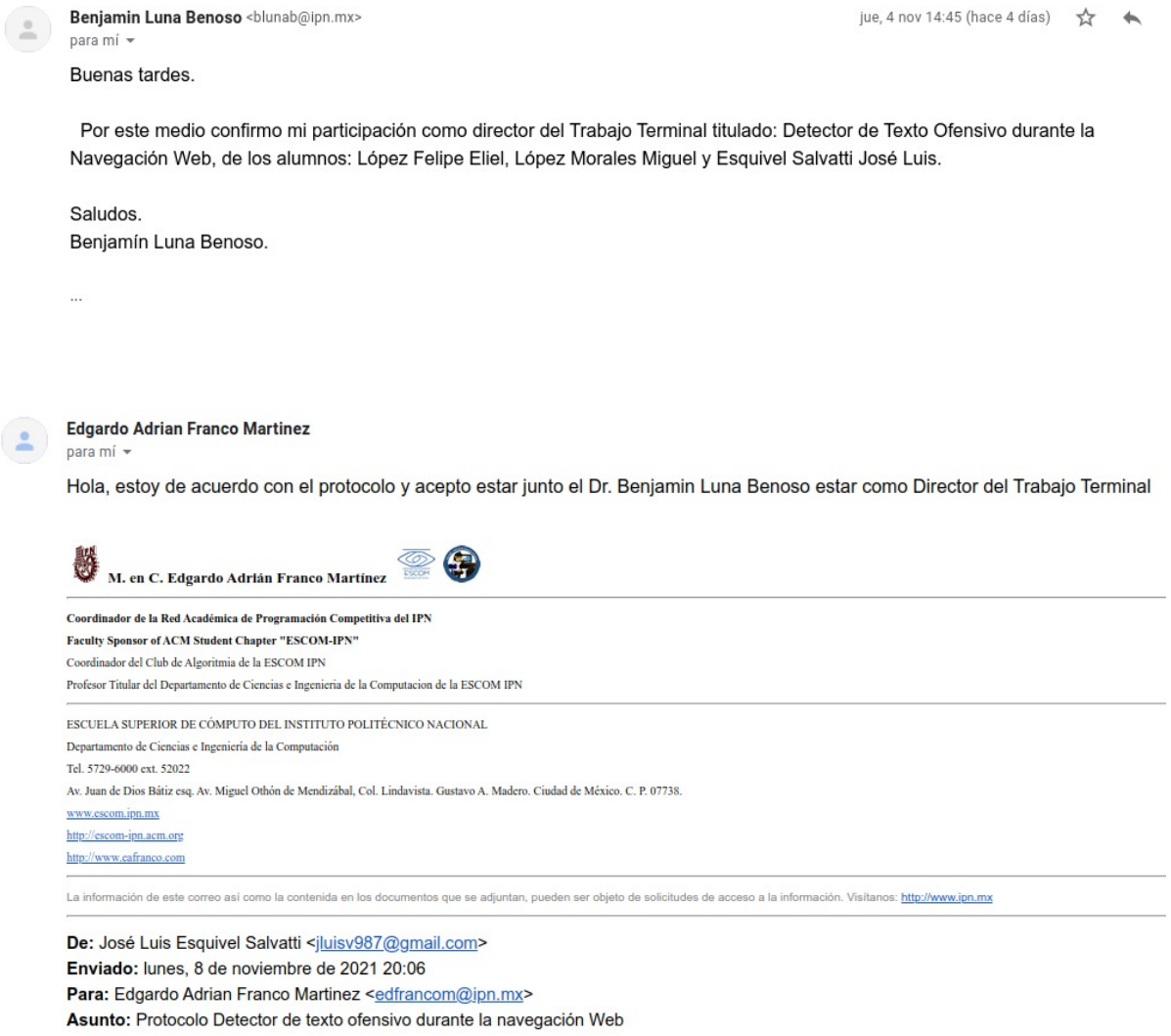
Eliel López Felipe.- Alumno de la carrera de
Ingeniería en Sistemas Computacionales en ESCOM
Especialidad Sistemas, Boleta: 2018340281,
Tel. 5573333886, email: elopezf1700@alumno.ipn.mx
Firma:_____

José Luis Esquivel Salvatti.- Alumno de la carrera de
Ingeniería en Sistemas Computacionales en ESCOM
Especialidad Sistemas, Boleta: 2015020363 ,
Tel. 5585344716, email: jesquivels1400@alumno.ipn.mx
Firma:_____

Dr. Benjamín Luna Benoso. - Licenciatura en Matemáticas
por la ESFM del IPN. Maestría y doctorado en ciencias de
la computación del CIC-IPN. Actualmente profesor de tiempo
completo de la ESCOM.
Áreas de interés: Reconocimiento de patrones, Análisis de imágenes,
Morfología matemática.
Tel: 57296000 ext: 52022, email: blunab@ipn.mx
Firma:_____

Edgardo Adrián Franco Martínez.- Maestría en Ciencias de la
Computación (CINVESTAV-IPN). Actualmente profesor de la
ESCOM del IPN. Ingeniería en Sistemas Computacionales con
especialidad en Electrónica (ESCOM-IPN)
Áreas de interés: Educación, Programación y Sistemas,
Algoritmia y Programación Competitiva
Tel: 5729 6000 ext.: 52022. edfrancom@ipn.mx
Firma:_____

Acuses de confirmación del correo electrónico de los directores e integrantes del equipo de trabajo terminal.



Benjamin Luna Benoso <blunab@ipn.mx> jue, 4 nov 14:45 (hace 4 días)
para mí

Buenas tardes.

Por este medio confirmo mi participación como director del Trabajo Terminal titulado: Detector de Texto Ofensivo durante la Navegación Web, de los alumnos: López Felipe Eliel, López Morales Miguel y Esquivel Salvatti José Luis.

Saludos.
Benjamín Luna Benoso.

...

Edgardo Adrian Franco Martinez
para mí

Hola, estoy de acuerdo con el protocolo y acepto estar junto el Dr. Benjamin Luna Benoso estar como Director del Trabajo Terminal

M. en C. Edgardo Adrián Franco Martínez

Coordinador de la Red Académica de Programación Competitiva del IPN
Faculty Sponsor of ACM Student Chapter "ESCOM-IPN"
Coordinador del Club de Algoritmia de la ESCOM IPN
Profesor Titular del Departamento de Ciencias e Ingenieria de la Computacion de la ESCOM IPN

ESCUELA SUPERIOR DE CÓMPUTO DEL INSTITUTO POLITÉCNICO NACIONAL
Departamento de Ciencias e Ingeniería de la Computación
Tel. 5729-6000 ext. 52022
Av. Juan de Dios Bátiz esq. Av. Miguel Othón de Mendizábal, Col. Lindavista. Gustavo A. Madero. Ciudad de México. C. P. 07738.
www.escom.ipn.mx
http://escom-ipn.acm.org
http://www.eafranco.com

La información de este correo así como la contenida en los documentos que se adjuntan, pueden ser objeto de solicitudes de acceso a la información. Visítanos: http://www.ipn.mx

De: José Luis Esquivel Salvatti <jluisv987@gmail.com>
Enviado: lunes, 8 de noviembre de 2021 20:06
Para: Edgardo Adrian Franco Martinez <edfrancom@ipn.mx>
Asunto: Protocolo Detector de texto ofensivo durante la navegación Web

Eliel López Felipe lun, 8 nov 22:43 (hace 17 horas)

para mí

Buenas noches

Yo Eliel López Felipe estoy de acuerdo en realizar el protocolo
(Detector de texto ofensivo durante la navegación Web) con ustedes

Muchas gracias

El lun, 8 nov 2021 a las 22:41, José Luis Esquivel Salvatti
(<jluisv987@gmail.com>) escribió:
>
>

José Luis Esquivel Salvatti <jluisv987@gmail.com> lun, 8 nov 22:46 (hace 17 horas)

para Eliel

Por este medio confirmo mi participación en el protocolo Detector de texto ofensivo durante la navegación web
-José Luis Esquivel Salvatti

Miguel Ángel López Morales 15:51 (hace 29 minutos)

para mí

Buena tarde.
Por este medio confirmo mi participación como alumno que desarrollará el Trabajo Terminal titulado: Detector de Texto Ofensivo durante la Navegación Web, junto a mis compañeros: López Felipe Eliel, y Esquivel Salvatti José Luis.

Sin más por el momento le deseo una buena tarde.

Atentamente:
Miguel Ángel López Morales.