Plataforma de trabajo colaborativo para etiquetar conjuntos de datos utilizados en el aprendizaje automático en la clasificación de imágenes y opiniones

Trabajo Terminal No. _ _ _ _ _ _ _ _ _
*Alumnos: *Pacheco Castillo Isaías*
Directores: Juárez Gambino Omar, García Mendoza Consuelo Varinia
e-mail: isaiaspachecoc@gmail.com

Resumen ─ Se propone crear una plataforma web de trabajo colaborativo que utilice inteligencia humana para realizar etiquetado de datos. El etiquetado permite identificar elementos de interés en un conjunto de datos, por ejemplo, objetos en una imagen u opiniones en un texto. Estos datos son utilizados por algoritmos de aprendizaje automático durante la etapa de entrenamiento para generar un modelo que pueda realizar tareas de clasificación. La plataforma permitirá definir tareas de etiquetación de imágenes y texto. Posteriormente mediante el trabajo colaborativo múltiples personas (*Workers*) realizarán dichas tareas. Además, se contempla remunerar económicamente a los etiquetadores y mostrar estadísticas relevantes de los datos etiquetados.

Palabras clave – Etiquetado de datos, Trabajo colaborativo, Clasificación, Aprendizaje automático.

1. Introducción

Hoy en día se tiene una gran demanda de análisis de datos avanzados que conduzcan al uso del aprendizaje automático (machine learning) [1] para el desarrollo de tecnologías como big data, bussiness intelligence y automatización. Como explica Sandhu [2], el aprendizaje automático es un subconjunto de la inteligencia artificial, que utiliza técnicas computarizadas para resolver problemas basados en datos históricos e información. Algunos de los algoritmos utilizados en el aprendizaje automático son; aprendizaje supervisado y no supervisado.

En el aprendizaje supervisado se aplica un algoritmo y se crea un modelo para asignar una entrada a una salida. Para que este funcione se necesita de un conjunto de datos etiquetados que el modelo pueda aprender y de esta forma tomar decisiones correctas.

El proceso para crear estos conjuntos de datos etiquetados, necesarios para construir los modelos, suele ser costoso, complicado y requiere mucho tiempo, por lo que una plataforma web basada en el trabajo colaborativo(crowdsourcing) y que aproveche la inteligencia colectiva puede ser de gran ayuda para facilitar la obtención de estos datos.

Crowdsourcing [3] es un tipo de actividad participativa en línea en la cual una persona u organización, denominados crowdsourcers, propone a un grupo de individuos con ciertas características, a través de una convocatoria abierta, la realización voluntaria de una tarea. Completar esta tarea implica que los participantes obtengan la satisfacción de algún tipo de necesidad, en este caso económica, mientras que el crowdsourcer obtendrá y utilizará a su favor lo que el usuario ha aportado.

En [4] se puede observar que existe inteligencia colectiva en grupos de personas que realizan tareas cognitivas y esto implica que se tenga un mayor desempeño en la realización de estas tareas de forma grupal que de forma individual lo cual será aprovechado en la propuesta planteada.

Existen varias soluciones en el mercado que utilizan el crowdsourcing como Amazon Mechanical Turk [5], Clickworker [6], Upwork [7] y Lionbridge [8]. A continuación, se muestra una tabla comparativa entre estas soluciones y la propuesta planteada.

Plataforma	Descripción	Facilidad de creación de tareas	Muestra estadísticas del conjunto de datos etiquetado	Presencia en México	Interfaz intuitiva
Amazon Mechanical Turk	Permite a las empresas aprovechar la inteligencia colectiva, las habilidades y los conocimientos de una fuerza de trabajo global para acelerar el aprendizaje automático.	✗	✗	✗	✗
Clickworker	Utiliza el poder de la multitud global de clickworkers para generar, validad y etiquetar datos.	✓	✗	✓	✓
Upwork	Es una plataforma gratuita que conecta profesionales freelance y agencias con empresas	✓	✗	✓	✗
Lionbridge	Proporciona datos de alta calidad para el aprendizaje automático a escala.	✗	✗	✓	✗
Propuesta	Aprovecha la inteligencia humana y el trabajo colaborativo para etiquetar, validar y mostrar estadísticas del conjunto de datos.	✓	✓	✓	✓

Tabla 1: Tabla comparativa entre plataformas similares

Como se puede ver en la Tabla 1, la propuesta de este trabajo terminal cuenta con todas las características de las demás plataformas, pero además muestra algunas estadísticas sobre el conjunto de datos etiquetados. Estas estadísticas permitirán conocer, por ejemplo, el acuerdo alcanzado entre etiquetadores (acuerdo inter-anotador) para la tarea, así como la frecuencia de las etiquetas.

Para mantener la misma terminología que las plataformas similares se nombrara *Requesters* a los usuarios que establecen las tareas y *Workers* a los usuarios que resuelven las tareas.

2. Objetivo

Crear una plataforma web de crowdsourcing que utilice inteligencia humana y aproveche la inteligencia colectiva para etiquetar conjuntos de datos utilizados en el aprendizaje automático para la clasificación de imágenes y opiniones, así como mostrar estadísticas acerca del conjunto de datos.

Objetivos específicos.

1. Crear el módulo para la creación de tareas
2. Crear el módulo para resolver las tareas
3. Crear el módulo para manejar pagos
4. Crear el módulo de análisis y estadísticas

3. Justificación

La tarea de etiquetar conjuntos de datos utilizados en el aprendizaje automático suele ser costosa y la que más trabajo humano y tiempo consume. Por lo tanto, los resultados de este trabajo terminal permitirán:

1) Definir tareas de etiquetación de datos de manera fácil
2) Etiquetar conjuntos de datos de manera rápida y eficiente

3) Mostrar información y estadísticas acerca del conjunto de datos etiquetados

El trabajo propuesto es novedoso ya que este, además de etiquetar de forma rápida y eficiente los conjuntos de datos, mostrará ciertas estadísticas y análisis sobre el conjunto de datos etiquetado, también se implementará una interfaz intuitiva y fácil de usar.

4. Productos o Resultados esperados

Los productos esperados son los siguientes:

- Una plataforma web para la creación y resolución de tareas de etiquetado
- Documentación del análisis y diseño
- Manual técnico y de usuario

En la Figura 1 se muestra la arquitectura del sistema donde se puede observar el funcionamiento básico de cada módulo que se va a implementar y la interacción con los dos tipos de usuarios.

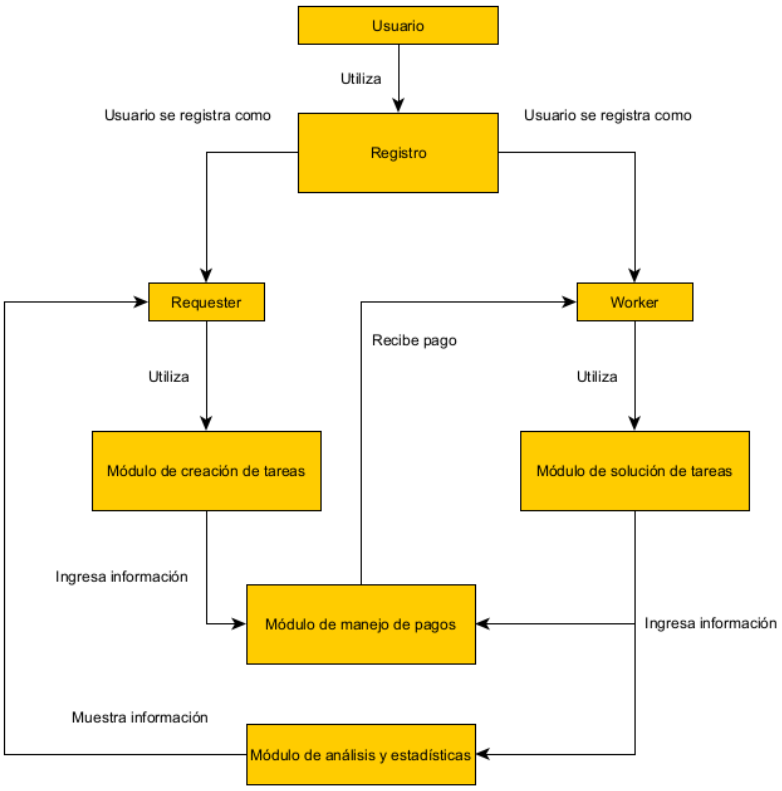

Figura 1. Arquitectura de la plataforma

5. Metodología

La metodología empleada será la incremental [9] ya que una de las bondades de esta metodología es la entrega de resultados en un periodo corto de tiempo, esto con el fin de visualizar los avances gradualmente y si es necesario hacer ajustes en cada paso del desarrollo.

Las fases de la metodología incremental son:

- Análisis
- Diseño
- Desarrollo
- Pruebas

Después de cada incremento es posible entregar un producto funcional que será evaluado para ver si cumple los requerimientos necesarios y en caso de ser necesario se pueden realizar correcciones o seguir con otro incremento.

6. Cronograma

Actividad	AGO	SEP	OCT	NOV - DIC	ENE	FEB	MAR	ABR	MAY -JUN
Investigación de las tecnologías a utilizar para el cliente y servidor	■								
Diseño e implementación de la base de datos		■							
Diseño e implementación de las operaciones cliente - servidor			■						
Creación del registro para los Requesters y los Workers				■					
Pruebas del registro de usuarios				■					
Documentación del trabajo terminal	■	■	■	■	■	■	■	■	■
Creación del módulo de creación de tareas para los Requesters					■				
Creación del módulo de selección y solución de tareas para los Workers						■			
Pruebas iniciales para el etiquetado de un conjunto de datos							■		
Creación del módulo de análisis y estadísticas del conjunto de datos etiquetado								■	
Pruebas de etiquetado y análisis del conjunto de datos								■	
Creación y pruebas del módulo de pagos									■
Pruebas de etiquetado, estadísticas y pagos									■
Documentación del trabajo terminal	■	■	■	■	■	■	■	■	■

7. Referencias

[1] M. W.Berry, A. Mohamed y B. Wah Yap, Supervised and Unsupervised Learning for Data Science, USA: Springer, 2020.

[2] S. Tejenderkaur H., «MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING,» *International Journal of Advanced Research in Computer Science*, vol. 9, nº 2, pp. 1-3, 2018.

[3] E. E.-A. y. F. González-Ladrón-de-Guevara, «Towards an integrated crowdsourcing definition,» *Journal of Information Science*, pp. 9-13, 2012.

[4] A. Williams Woolley, C. F. Chabris, A. Pentland, N. Hashmi y T. W. Malone|, «Evidence for a Collective Intelligence Factor in the Performance of Human Groups,» *Science*, p. 4, 2010.

[5] Amazon, «Amazon Mechanical Turk,» Amazon, [En línea]. Available: https://www.mturk.com/. [Último acceso: 08 04 2021].

[6] Clickworker, «Clickworker,» Clickworker, [En línea]. Available: https://www.clickworker.com/. [Último acceso: 08 04 2021].

[7] Upwork, «Upwork,» Upwork, [En línea]. Available: https://www.upwork.com/. [Último acceso: 08 04 2021].

[8] LionBridge, «LionBridge,» LionBridge, [En línea]. Available: https://lionbridge.ai/. [Último acceso: 08 04 2021].

[9] I. SOMMERVILLE, Ingeniería del software, Madrid: PEARSON EDUCACIÓN, 2005.

8. Alumnos y Directores

Pacheco Castillo Isaias. – Alumno de la carrera de Ingeniería en Sistemas Computacionales en la Escuela Superior de Cómputo del Instituto Politécnico Nacional, Boleta: 2017361951, Tel: 5519804576, email: isaiaspachecoc@gmail.com

Firma: _____

Joel Omar Juárez Gambino. – Licenciado en Informática por la Facultad de Informática, UAS. Doctor en Ciencias de la Computación por el CIC, IPN. Sus áreas de estudio son: Inteligencia Artificial, Lenguaje Natural y Representación de Conocimiento. Departamento de Ciencias e Ingeniería de la Computación, ESCOM, Tel. 57296000 Ext. 52022, email: jjuarezg@ipn.mx

Firma: _____

Consuelo Varinia García Mendoza. – Ingeniera en Sistemas Computacionales por la ESCOM, IPN, Doctora en Ciencias en Tecnología Avanzada por el CICATA-Legaria, IPN. Sus áreas de estudio son: Análisis de algoritmos y Optimización. Departamento de Ciencias e Ingeniería de la Computación, ESCOM, Tel. 57296000 Ext. 52022, email: cvgarcia@ipn.mx

Firma: _____

Pacheco Isaias protocolo

3 mensajes

Isaias Pacheco Castillo <isaiaspachecoc@gmail.com> 3 de junio de 2021, 18:07
Para: Omar Juárez Gambino <omarjg82@gmail.com>, consuelo.varinia@gmail.com

Buen día profesores, espero se encuentren muy bien.

Les adjunto la versión final del protocolo que se va a entregar.
Saludos.

--
Isaías P.

Pacheco_Isaias_Protocolo.pdf
215K

Consuelo Varinia <consuelo.varinia@gmail.com> 3 de junio de 2021, 18:46
Para: Isaias Pacheco Castillo <isaiaspachecoc@gmail.com>

Gracias Isaias, confirmo de recibido.

Saludos

Consuelo Varinia García Mendoza
[El texto citado está oculto]

Omar Juárez Gambino <omarjg82@gmail.com> 3 de junio de 2021, 19:03
Para: Isaias Pacheco Castillo <isaiaspachecoc@gmail.com>

Hola. Acuso de recibido.
[El texto citado está oculto]