Procesamiento de lenguaje natural y Bert para el perfilado de autores en la red social X

Natural language processing and Bert for social network author profiling X

Recibido: mayo 14 de 2025 | Revisado: mayo 28 de 2025 | Aceptado: junio 17 de 2025

Ivan Petrlik Azabache[1]
Ciro Rodríguez Rodríguez[1]
Pedro Lezama Gonzales[1]
Luz Torres-Talaverano[1]
Enma Graciela Vásquez Hurtado[2]
Karina Inés Hinojosa Pedraza[2]

Resumen

En la actualidad X se ha convertido en una de las redes sociales más importantes para expresar opiniones e intereses en la red. La gran cantidad de datos generados permite obtener sistemas automatizados para perfilar a los usuarios en función del género, nacionalidad e intereses temáticos. Hay dificultades en este proceso no solo por el contenido breve, sino también por la ambigüedad y el uso de varios idiomas.

El objetivo de esta propuesta es el de generar un modelo de aprendizaje profundo utilizando BERT que sea capaz de identificar atributos demográficos y temáticos a partir de los tweets. Se usarán modelos preentrenados del tipo de BERT y Multilingual BERT, aplicados sobre corpus del PAN Author Profiling Task (CLEF 2019) en inglés y español.

El trabajo propuesto profundizará en el análisis mediante datos de la clasificación supervisada para la clasificación de género y nacionalidad y la extracción de temas a través de técnicas no supervisadas, como LDA y BERTopic. Estas opciones incluyen técnicas de preprocesamiento, reducción dimensional (UMAP) y evaluación mediante métricas como la precisión y la exactitud.

Es previsible que los resultados del análisis puedan demostrar la aplicabilidad de BERT para conseguir perfilados automáticos aplicados en el ámbito de marketing, de análisis sociopolíticos y de personalización de contenidos.

Palabras clave: Lenguaje natural, Bert, Perfilado, Red Social X

Abstract

Today X has become one of the most important social networks for expressing opinions and interests on the web. The large amount of data generated allows automated systems to profile users based on gender, nationality and thematic interests. There are difficulties in this process not only because of the short content, but also because of the ambiguity and the use of several languages.

The goal of this proposal is to generate a deep learning model using BERT that is able to identify demographic and thematic attributes from tweets. Pre-trained models of the BERT and Multilingual BERT type will be used, applied

1 Universidad Nacional Mayor de San Marcos, Lima - Perú
2 Universidad Nacional Federico Villarreal, Lima - Perú

Autor de correspondencia:

https:

on PAN Author Profiling Task (CLEF 2019) corpora in English and Spanish. The proposed work will deepen the analysis using supervised classification data for gender and nationality classification and topic extraction through unsupervised techniques, such as LDA and BERTopic. These options include preprocessing techniques, dimensional reduction (UMAP) and evaluation using metrics such as precision and accuracy.
It is expected that the results of the analysis can demonstrate the applicability of BERT for automatic profiling in marketing, socio-political analysis and content personalization.

Keywords: Natural language, Bert , Profiling , Social Network X

Introducción

A lo largo de los últimos años, la inteligencia artificial ha desempeñado un papel relevante en multitud de sectores y ha protagonizado un acontecimiento importante en lo que respecta a su avance y su uso diario. Uno de los campos que más cambios ha sufrido ha sido el de consumo masivo y retail y, sobre todo en el ámbito, el marketing y la personalización publicitaria. Las empresas han ido muy rápido y han ajustado las diferentes estrategias en función de la IA con mucha fuerza en lo que se refiere a conocer mejor a los consumidores y analizar sus hábitos, así como ofrecer productos y servicios a medida (Davenport et al, 2021, Harvard Business Review).

El uso excesivo de las redes sociales como la gran variedad de datos que existen fue lo que propició un cambio acelerado de las técnicas de marketing. Históricamente, las empresas de consumo masivo han buscado una mayor comprensión de sus públicos como eje para la creación de valor (Mendivelso y Lobos, 2019) pero en este caso, se debía realizar a través de métodos tradicionales (encuestas, estudios de mercado, etc.), los cuales, aunque válidos presentan problemáticas asociadas (sesgos, resultados lentos y altos costes).

Las redes sociales como Facebook o Twitter han revertido la limitación de los antiguos métodos de acceso a determinadas opiniones del público siendo capaz de representar su opinión en tiempo real y sin intermediarios. Actualmente se puede realizar un análisis de un gran número de temáticas y de un gran número de públicos gracias a cómo incrementan las plataformas en contenido generado de forma constante. Este tamaño de los datos en las redes como herramientas del marketing generan oportunidades para nuevas estrategias de mercado más eficientes, aunque también el reto de procesar y extraer información para la elaboración de decisiones.

Por tanto, no debería sorprendernos que buena parte de la publicidad actual esté centrada en redes sociales en la que haciendo uso de la información se establecen estrategias personalizadas para ser más propensos a la compra. La personalización se vuelve en este momento uno de los factores que mejoran esta probabilidad, y conocer características tales como el género, la nacionalidad, el idioma o los intereses de los usuarios se vuelve muy útil.

La presente investigación trata de aplicar un modelo de Deep Learning - BERT, acrónimo en inglés de

Bidirectional Encoder Representations from Transformers, para analizar tuits que se ha recogiendo en un corpus en concreto (Rangel et al., 2016) caracterizando de este modo a los autores de los mismos en cuanto a género, nacionalidad e intereses. El objetivo de ello es usar el procesamiento del lenguaje natural (NLP) con el fin de ayudar a la personalización de la publicidad mediante el análisis de contenido en redes sociales.

La relevancia de esta investigación también se encuentra en que, tradicionalmente el perfilado de autor y la extracción de temas de interés se han llevado a cabo con modelos clásicos de machine learning y fundamentalmente en inglés. En esta dirección se encuentra marcada la posibilidad de aplicar un modelo de Deep Learning tipo BERT (Devlin et al., 2019), en análisis de contenido en español, en función de sus variantes regionales y de sus características lingüísticas, de tal forma que se amplía el alcance y la aplicabilidad de esta forma de proceder en contextos hispanohablantes.

A continuación, se van a presentar los antecedentes de la respectiva investigación: Delmondes y Paraboni (2022) argumentan que el autor profiling busca inferir datos demográficos a partir de la información contenida en los textos. La técnica obtiene los mejores resultados cuando el entrenamiento y la verificación se realizan en el mismo dominio. En ámbitos entre dominios surgen problemas como el desajuste debido a los léxicos. Su aportación con BERT y múltiples dominios produce una mejora significativa en la precisión comparándola con los modelos de tipo tradicional.

Jiménez-Villar (2020) señala que el principal desafío que describe el perfilado de autor es la escasez de datos etiquetados e indica el uso de técnicas de aumento de datos, así como de modelos SVM para mejorar el rendimiento en tareas como la detección de anorexia y depresión obteniendo en concreto mejoras entre el 1% y el 18% de la métrica F1 con respecto a modelos sin aumento. Aunque los resultados obtenidos son bastante similares a los resultados del estado del arte usando un modelo de red neuronal simple.

De acuerdo con el trabajo de investigación realizado por Mamgain, Balabantaray y Das (2019) en el que se destaca, por una parte, el crecimiento que han ido obteniendo los datos textuales en las redes sociales y la creciente importancia que se le daba a la Identificación Automática de Autoría. En segundo lugar, el perfilado de autor, que tiene como objetivo inferir características como el género o la variedad lingüística, el bajo el que, en diferentes áreas como el marketing o el análisis forense, pero que predicen en este estudio las características por determinar en función de la utilización de los datos extraídos y tratados con el formato utilizado en los términos del reto del PAN 2017 del inglés. Además, mencionan las características bajo las que dicen que resaltan los primeros retos del PAN para el desarrollo de esta área.

Chiu, Sandroni y Paraboni (2018) describen el perfilado de autores como una de las tareas más relevantes en PLN, el cual presenta aplicaciones en temas relacionados con la seguridad, las ventas, y el análisis forense. Hacen hincapié en que el uso de los textos en las redes sociales va en aumento, considerando que la mayor parte de las investigaciones se agotan en ejemplos del inglés o bien distinciones de tipo de edad o de

género. Posteriormente, la investigación que ellos mismos han llevado a cabo se centra en el perfilado de autores en portugués a partir de Facebook y también la predicción del grado de religiosidad. Examinaron diferentes enfoques de esta tarea obteniendo resultados iniciales que se podrían considerar prometedores.

Para Patra, Das y Das (2018) destaca la creciente atención que ha tenido el perfilado de autor, es decir, la inferencia de ciertos rasgos tales como sexo, edad o personalidad desde el estilo textual. Este estudio experimental sobre PAN-2018 se basa en datos multimodales de Twitter (texto e imagen) para obtener la predicción del género por parte del autor, SVM con características semánticas y estilísticas los métodos de clasificación, obteniendo así precisiones en torno al 76% en árabe, inglés y español.

De acuerdo con lo que mencionan Veenhoven y colaboradores (2018) tomaron parte en PAN utilizando un modelo bi-LSTM con atención para predecir el sexo de autores a partir de los tuits y las imágenes en inglés, español y árabe. De hecho, sugirieron traducir datos de otros lenguajes como estrategia para aumentar el entrenamiento. Esta técnica económica y eficaz aumentó la precisión, obteniendo puntuaciones de 79.3%, 80.4% y 74.9% en los tres idiomas.

Método

La metodología Design Thinking se aplicó para determinar un prototipo debidamente validado considerando las siguientes etapas de empatizar, definir, idear, prototipar y evaluar que a continuación se va a desarrollar:

Empatizar

El propósito de esta etapa radica precisamente en responder a una actual necesidad del marketing, que es la comprensión cabal del consumidor para llevar a cabo campañas de marketing ajustadas a sus necesidades y personalizadas. Para ello, el trabajo que proponemos se basa en aprovechar la cantidad de datos generada en las redes sociales, especialmente X, como fuente de información. Dado que los usuarios están compartiendo continuamente su opinión y sus intereses, la conversión de esos datos en conocimiento es más que necesaria. Nuestra propuesta consiste en construir un modelo capaz de extraer y analizar toda esa información para favorecer la toma de decisiones estratégicas en el marketing.

Definir

La propuesta sugiere la elaboración de un modelo en respuesta a las actuales necesidades de información de marketing, justo como se argumentó en la fase de empatizar. El modelo debe ser capaz de predecir el sexo del autor de tuits, una información relevante para personalizar los anuncios en esos tuits. Debe tener también la capacidad de identificar el país de residencia, que es muy útil para poder segmentar campañas de publicidad en función de algunas características culturales o regionales. Por último, el modelo debe ser capaz de detectar los temas de interés del usuario (deportes, entretenimiento, etcétera), lo que nos permitirá orientar mejor los anuncios en cuanto a contenido y, por tanto, para estos anuncios en X haber desarrollado la cualidad del conocimiento posible de las personas objetivo. De este modo, se daría la posibilidad de orientar una

estrategia publicitaria más efectiva y más personalizada.

Idear

Durante la etapa de "Idear" del método de Design Thinking se generaron casi todas las ideas posibles de la innovación y adquirir algo de relevancia, en la que cobraba importancia el preprocesado, en los modelos de aprendizaje profundo. Dado que los tuits presentan lenguaje informal, emojis, menciones y etiquetas, resulta clave el preprocesado. Los elementos sin valor semántico, como las URLs deben desaparecer y otros como menciones y etiquetas, permitir conservar elementos que sí pueden arrojar información en el preprocesado. En este sentido, las jergas o términos regionales pueden permitir, por ejemplo, indicar el país del autor y los emojis pueden ayudar a identificar variables de tipo de género y permitir a su vez, mejorar el perfilado en Twitter.

Prototipar

A partir de lo expuesto, se define una aproximación inicial al prototipo, ya en condiciones de verificar cómo dicho modelo se comporta empleando BERT y distintas maneras de preprocesamiento; por tanto, se confrontan las distintas alternativas de realizarlo o no, y en definitiva se elaboran modelos teniendo en cuenta si se eliminan menciones, hashtags o emoticones como p. ej. Se pretende averiguar qué información textual aporta mayor valor semántico y predictivo. Y lo más importante, los resultados determinarán qué modelo es el más adecuado a la hora de seleccionar, bajo criterios de evaluación especificados, el modelo prototípico.

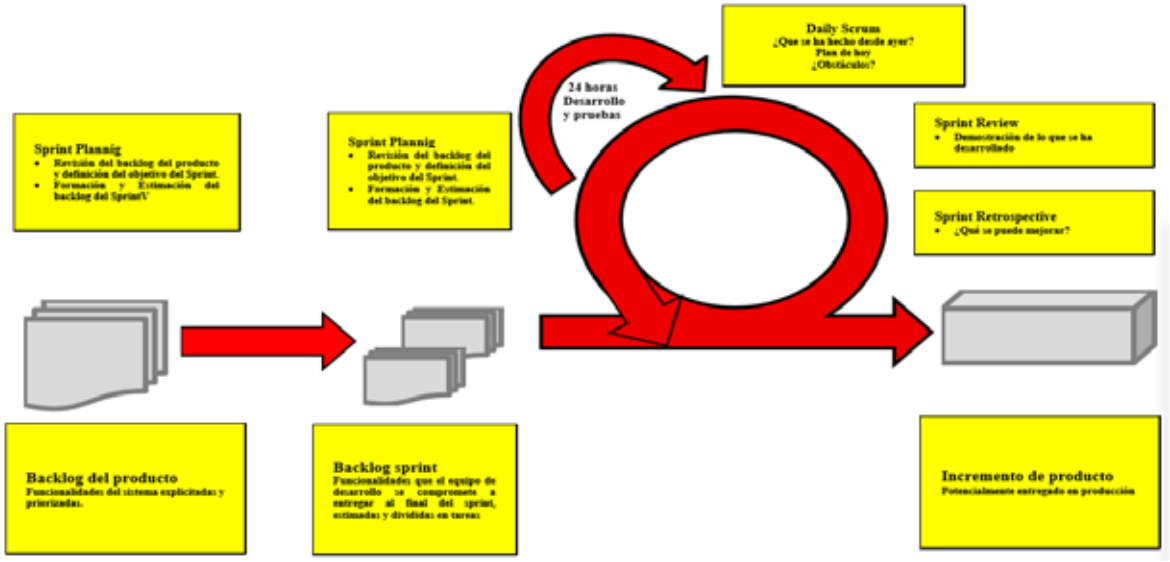
Selección de prototipos y criterio para hacerlo

Para escoger el mejor prototipo se utilizarán diferentes métricas particulares de la predicción. Para la predicción del género (predicción binaria balanceada) se usarán métricas concretas como accuracy o F1 score. En cambio, para la variante del español (multiclase con siete etiquetas balanceadas) se utilizará el accuracy multiclase. Finalmente, para la predicción de los temas de interés (modelo no supervisado) se procederá a evaluar manualmente, de acuerdo con su coherencia y su interés, los temas que se han extraído del corpus, priorizando las clases que son de interés en marketing como deportes, música o belleza.

Se sugiere la implementación de la metodología SCRUM, que es una metodología ágil orientada a gestionar productos complicados, ya sean de software o hardware (Jiménez, 2021). Creada por Jeff Sutherland y Ken Schwaber en los años 90, SCRUM articula su provincia de funcionamiento a partir de tres elementos fundamentales: Roles, Eventos y Artefactos, de este modo se facilita el desarrollo ágil del proyecto, tal y como se muestra en la Figura 2 al respecto.

Figura 1

Ciclo de vida de SCRUM

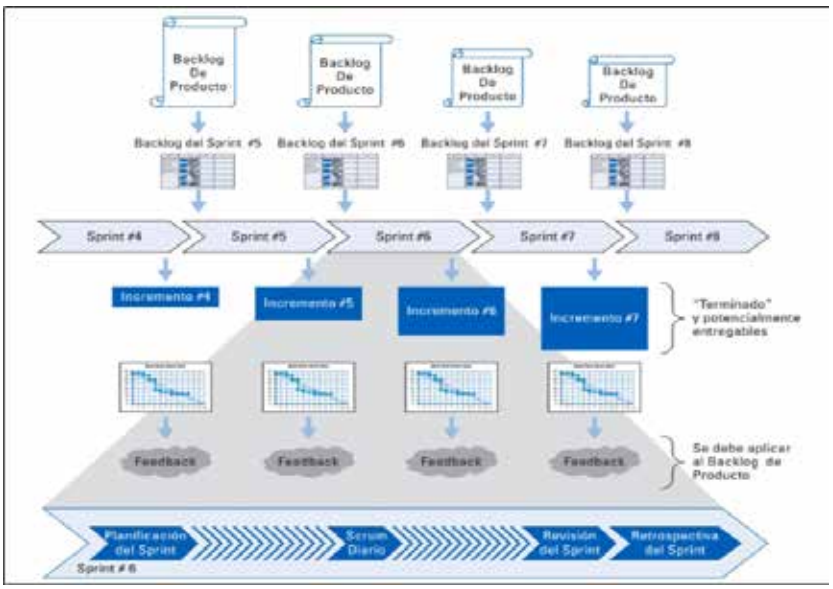


El ciclo SCRUM es la representación estructural del proceso SCRUM, y el ciclo de SCRUM, tal como vemos en la Figura 1, comienza a partir del product backlog, donde se priorizan las funcionalidades y se planifican las releases (versiones del producto). A continuación, encontramos el sprint backlog, donde se detallan las tareas que el equipo ha comprometido para cada sprint. Durante los sprint planning se define el objetivo y se estima el trabajo. El sprint dura entre dos y cuatro semanas e incluye reuniones clave: daily scrum, sprint review y retrospective. Al final de cada sprint, se genera un incremento del producto potencialmente entregable (Subra y Vannieuwenhuyze, 2018).

Seguidamente, en la Figura 2 se presenta el funcionamiento de los sprint en el proceso Scrum.

Figura 2

Funcionamiento de los sprint dentro de la metodología SCRUM.



La Figura 2 muestra la creación de un product backlog en el que se priorizan las funcionalidades del sistema y planifican los releases; y en el sprint backlog en el que se extienden las tareas que está dispuesto a entregar el equipo por sprint. En el sprint planning revisamos el product backlog y definimos el objetivo del sprint y la estimación de las tareas. De aquí, expondremos los elementos principales de la metodología SCRUM aplicada al proyecto de investigación.

Los roles dentro de la metodología SCRUM

Para poner en práctica la metodología SCRUM en el desarrollo del presente proyecto, es necesario que se asigne y se cumpla la relación y función de uno o varios roles importantes: Scrum Master, Product Owner, Data Scientists Data Analyst y Data Engineer. Cada uno de estos roles importantes tendrá una función en el equipo de desarrollo ágil del mismo para así asegurar el cumplimiento de los objetivos del proyecto, así como facilitar la colaboración efectiva. A continuación, se exponen con formato apropiado las tareas correspondientes a cada uno de los perfiles relacionados en las tablas correspondientes. A continuación tenemos la Tabla 1 que describe todo concerniente al rol Scrum Master:

Tabla 1

Rol de Scrum Master

Involucrado	Rol
Persona 1	Scrum Master
Descripción del rol	
Es el responsable de la aplicación de Scrum en el proyecto para el control del cumplimiento en tiempo y forma de cada sprint. Según (Chandan, 2019), el Scrum Master es el líder servidor del equipo Scrum. Es como un pastor para el equipo. El Scrum Master es un rol del equipo y él/ella es responsable de garantizar que el equipo respete sus valores y principios ágiles y siga los procesos y prácticas que el equipo acordó adoptar.	

Seguidamente, tenemos la Tabla 2 que describe todo concerniente al rol Product Owner:

Tabla 2

Rol de Product Owner

Involucrado	Rol
Persona 2	Product Owner
Descripción del rol	
El segundo individuo, en su papel de Product Owner, mantendrá una reunión con los usuarios finales para entender sus necesidades, así como para transmitir la visión de negocio al equipo técnico. Su rol es crucial en la comunicación, resolver dudas y detectar puntos críticos en la fase de desarrollo. Adicionalmente, será responsable de presentar en el negocio los resultados de cada sprint de tal modo que genere retroalimentación con respecto al producto. En palabras de Kelly (2019), el Product Owner es una figura solitaria, capaz de tomar decisiones, comprender los requerimientos del producto e integrarse inmediatamente como miembro del equipo.	

Después apreciamos la Tabla 3 que describe todo concerniente al rol Data Scientists:

Tabla 3

Rol de Data Scientists

Involucrado	Rol
Persona 3	Data Scientists
Descripción del rol	
Esta persona será la encargada de diseñar los modelos de procesamiento del lenguaje natural (NLP) para que nuestro proyecto pueda hacer el análisis de los *tuits*. Dentro de sus responsabilidades se encuentra el diseño del modelo, el entrenamiento y el ajuste adecuado de los hiperparámetros utilizando Cognitive Services y la máquina virtual de Azure. Según (Sarella, Srivastava, Jamberi y Syed, 2022), un científico de datos trabaja en el campo de la ciencia de datos. El nombre "científico de datos", nombre "científico de datos" se desarrolló en respuesta al hecho de que un científico de datos toma una gran cantidad de conocimientos de las disciplinas científicas y las aplica.	

Asimismo, tenemos la Tabla 4 que describe todo concerniente al rol Data Analyst:

Tabla 4

Rol de Data Analyst

Involucrado	Rol
Persona 4	Data Analyst
Descripción del rol	
Este rol, que será esencial para el desarrollo del proyecto y el diseño del modelo de Machine Learning, trabajará conjuntamente con el Product Owner para realizar el análisis de las necesidades de negocio, de la información de la que dispongamos en su caso y de los documentos de entrada del modelo. Sus funciones son clasificar comentarios de las redes sociales, detectar posibles sesgos (por ejemplo, procedencia geográfica) e identificar factores externos que pueden influir en las predicciones. Estará en constante comunicación con la persona Data Scientist para encauzar adecuadamente el modelo, como también asistirá a la persona Data Engineer en la preparación de los datos para el proceso del entrenamiento.	

Seguidamente, tenemos la Tabla 5 que describe todo concerniente al rol Data Engineer:

Tabla 5

Rol de Data Engineer

Involucrado	Rol
Persona 5	Data Engineer
Descripción del rol	
Esta persona se encargará junto con el Data Analyst del procesamiento de la información dentro de la infraestructura del proyecto. Sin embargo, a diferencia de la Data Analyst, esta persona tendrá un perfil más técnico y será el responsable de orquestar los diferentes pipelines en Data Factory, así como de diseñar las bases de datos en el Blob Storage y en el Data Lake Analytics. También será responsable de la limpieza de la información.	

Finalmente, tenemos la Tabla 6 que describe todo concerniente al rol Azure Specialist:

Tabla 6
Rol de Azure Specialist

Involucrado	Rol
Persona 6	Azure Specialist
Descripción del rol	
Esta persona se encargará junto con el Data Analyst del procesamiento de la información dentro de la infraestructura del proyecto. Sin embargo, a diferencia de la Data Analyst, esta persona tendrá un perfil más técnico y será el responsable de cualquier problema que se presente con el ambiente de Microsoft Azure. También será responsable de crear la documentación del proyecto.	

Según la información de las Tablas 1, 2, 3, 4, 5 y 6, observamos los roles del equipo Scrum con sus respectivas descripciones de cada uno de ellos, conformado por el Scrum master, Product Owner, Data Scientists, Data Analyst, Data Engineer y Azure Specialist.

El product backlog del proyecto

Para iniciar el desarrollo, el producto owner define el artefacto o documento que posee la lista completa de funcionalidades y requerimientos de los respectivos clientes o usuarios, llamado Product backlog que, según Subra, J. y Vannieuwenhuyze, A. (2018), contiene la expresión de las necesidades del Product Owner, traducidas en forma de User Stories. Asimismo, se ordenan según los criterios definidos por el Product Owner. El impacto de esto es que se tratan las Stories en el orden definido. Seguidamente, se va a especificar la lista de requerimientos que dan cumplimiento a la solicitud del cliente del presente proyecto que estamos proponiendo:

Definición de los requerimientos del cliente

En este paso, se realiza una reunión interna con el cliente para determinar los distintos requerimientos del usuario. A continuación, vamos a definir los requerimientos del usuario (Product Backlog Inicial):

Tabla 7
Product Backlog Inicial

Requerimientos
Preparar todo el proyecto para correr en un ambiente productivo y uno lanzamiento o pruebas.
Entrenamiento del modelo
Correcciones basadas en el feedback
Preprocesamiento de los datos (emojis, emoticones, slangs).
Uso del API que se conecta a X.
Obtener feedback de potenciales usuarios.
Limpieza de los datos
Análisis de los datos disponibles

Escribir las historias de usuarios

Seguidamente, se va a presentar el detalle de cada uno de los requisitos del producto backlog inicial, plasmados en historias del usuario. Aquí tenemos la Tabla 8 que describe todo concerniente a la historia de usuario 01:

Tabla 8

Historia de usuario 01: Analizar los datos disponibles.

Analizar los datos disponibles	
Como	Usuario
Quiero	Analizar las diferentes fuentes de información considerando las necesidades del negocio.
Para	Comprender los comentarios a grandes rasgos en la clasificación de estas mismas obtenidas en la red social X.
Criterios de aceptación	• Tener claro las necesidades del negocio. • Las fuentes de información tienen que ser de redes sociales.

Seguidamente, tenemos la Tabla 9 que describe todo concerniente a la historia de usuario 02.

Tabla 9

Historia de usuario 02: Uso del API que se conecta a Twitter.

Uso del API que se conecta a Twitter	
Como	Usuario
Quiero	Utilizar el API de X para el perfilado de los usuarios.
Para	Obtener los datos no estructurados de los comentarios de los usuarios.
Criterios de aceptación	Configuración del API de la red social X. Identificación del grupo de interés en la extracción de los datos en X.

A continuación, apreciamos la Tabla 10 que describe todo concerniente a la historia de usuario 03:

Tabla 10

Historia de usuario 03: Limpiar los datos.

Limpiar los datos	
Como	Usuario
Quiero	Realizar una limpieza de los datos extraídos a través del API de X.
Para	Obtener un dataset limpio y de buena calidad.
Criterios de aceptación	Identificar los parámetros correctos de la respectiva limpieza en el respectivo corpus o dataset.

Asimismo, tenemos la Tabla 11 que describe todo concerniente a la historia de usuario 04:

Tabla 11
Historia de usuario 04: Preprocesamiento de los datos

Pre-procesar los datos	
Como	Usuario
Quiero	Realizar el preprocesamiento de los datos.
Para	Tener una información de calidad.
Criterios de aceptación	Verificar que el dataset esté limpio. Identificar los **emojis, emoticones, slangs para el respectivo preprocesamiento**.

Seguidamente, tenemos la Tabla 12 que describe todo concerniente a la historia de usuario 05:

Tabla 12
Historia de usuario 05: Entrenamiento del modelo

Entrenar el modelo	
Como	Usuario
Quiero	Entrenar el modelo de Inteligencia artificial en el perfilado de autor.
Para	Lograr la predicción y clasificación acertada.
Criterios de aceptación	Verificar que los datos estén limpios y preprocesados. Seleccionar el modelo de I A

Continuando, tenemos la Tabla 13 que describe todo concerniente a la historia de usuario 06:

Tabla 13
Historia de usuario 06: Obtener del feedback de los resultados del modelo.

Obtener feedback de los resultados del modelo	
Como	Usuario
Quiero	Obtener feedback de los resultados del modelo a través de una bitácora de incidencias.
Para	Tener en cuenta todas las acciones de configuración en la implementación de los datos de entrada y salida.
Criterios de aceptación	Realizar un entrenamiento con datos preprocesados. Obtener los resultados al 100% si buenas.

Asimismo, tenemos la Tabla 14 que describe todo concerniente a la historia de usuario 07:

Tabla 14

Historia de usuario 07: Correcciones basadas en el feedback.

Correcciones basadas en el feedback	
Como	Usuario
Quiero	Aplicar la bitácora todo el feedback realizado.
Para	Poner a punto el modelo.
Criterios de aceptación	Registro de la bitácora observaciones al 100%. Verificar que los datos estén preprocesados.

Seguidamente, tenemos la Tabla 15 que describe todo concerniente a la historia de usuario 08.

Tabla 15

Historia de usuario 08. Preparar todo para correr en un ambiente productivo y lanzamiento.

Preparar todo para correr en ambiente productivo y lanzamiento	
Como	Usuario
Quiero	Ejecutar el modelo en un ambiente en producción para el lanzamiento formal.
Para	Obtener los resultados de las campañas de marketing digital.
Criterios de aceptación	Tener implementado el ambiente en producción. Tener desplegada la aplicación.

Planificación de los sprint del proyexto

Continuando con el proceso SCRUM, ahora se tendrá que desarrollar el sprint Planning. Es un evento que se lleva a cabo a través de un trabajo colaborativo de todo el equipo SCRUM. Asimismo, se da inicio al sprint con el propósito de definir qué se puede entregar al final de la interacción y cómo se conseguirá ese trabajo. Durante la reunión de planificación, el product owner describe las características con mayor prioridad al equipo. El equipo realiza las preguntas necesarias para poder convertir una historia de usuario en product backlog en tareas más específicas para el sprint backlog. Dentro de la planificación se debe de tener en cuenta dos eventos muy importantes:

a. El objetivo del Sprint
b. El sprint backlog

Seguidamente, se va a mostrar en la Tabla 16 la descripción del sprint backlog:

Tabla 16
Sprint backlog del Proyecto

Ítem	Historias de Usuarios	ID	Prioridad	Estimación	Sprint	Estado
1	Análisis de los datos disponibles	**HU01**	Muy Alto	60	1	En proceso
2	Uso del API que se conecta a X	**HU02**	Alto	50	2	En proceso
3	Limpieza de los datos	**HU03**	Alto	30	3	En proceso
4	Preprocesamiento de los datos (emojis, emoticones, slangs)	**HU04**	Alto	25	4	En proceso
5	Entrenamiento del modelo	**HU05**	Alto	30	5	En proceso
6	Obtener feedback de potenciales usuarios.	**HU06**	Medio	20	6	En proceso
7	Correcciones basadas en el feedback	**HU07**	Medio	15	7	En proceso
8	Preparar todo para correr en ambiente productivo y lanzamiento.	**HU08**	Medio	20	8	En proceso

Asimismo, en la Tabla 17 se describe las semanas y la estimación del sprint backlog del proyecto:

Tabla 17
Sprint backlog del Proyecto

Sprint	Historias de Usuarios	Semanas	Estimación
1	Análisis de los datos disponibles	2	60
2	Uso del API que se conecta a X	2	50
3	Limpieza de los datos	2	30
4	Preprocesamiento de los datos (emojis, emoticones, slangs)	2	25
5	Entrenamiento del modelo	2	30
6	Obtener feedback de potenciales usuarios	2	20
7	Correcciones basadas en el feedback	2	15
8	Preparar todo para correr en ambiente productivo y lanzamiento	2	20
Totales		**16**	**250**

Implementación de la propuesta

Implementación

La propuesta presentada para el desarrollo de este proyecto implica el uso de servicios en la nube, donde se priorizará una implementación escalable y eficiente. Se proponen herramientas de la oferta de Azure como Blob Storage para almacenamiento, Data Lake Analytics para ejecutar consultas masivas y Cognitive Services para la implementación de modelos de análisis

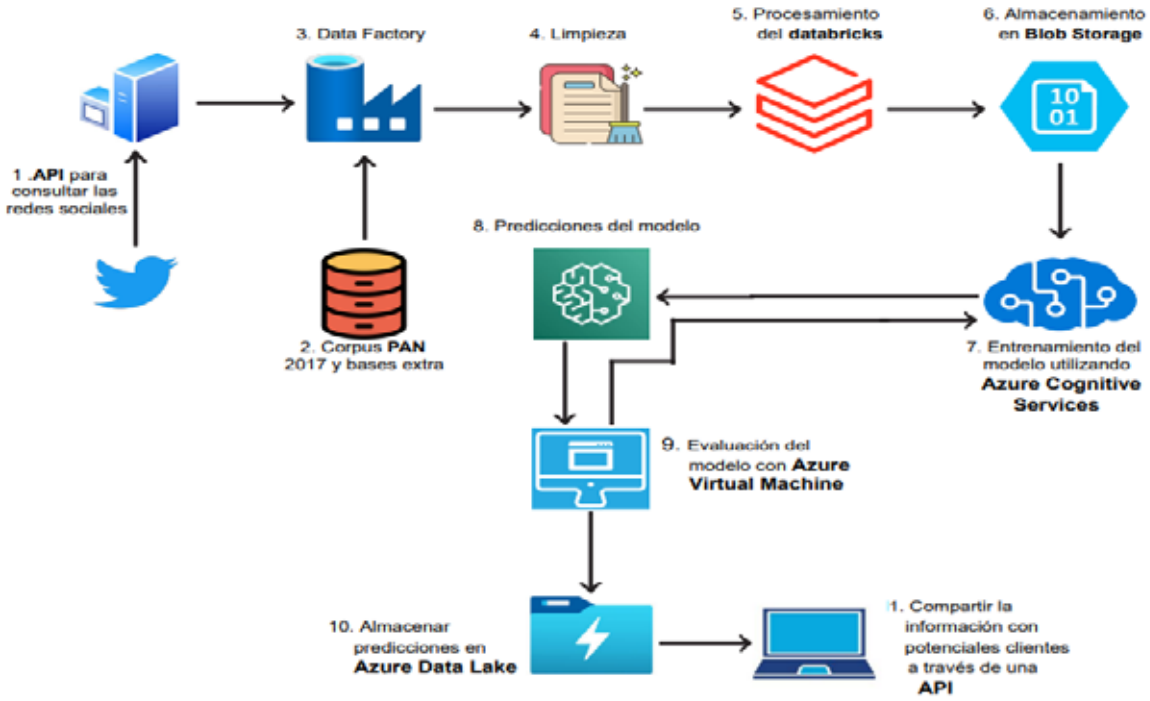
de sentimientos a partir de modelos de NLP ya entrenados.

Asimismo, se utilizarán Databricks para gestionar pipelines con Spark y Virtual Machines con GPU para entrenar modelos, además de Data Factory para orquestar todo el flujo del proyecto. El modelo se entrenará con el dataset creado para la competencia PAN 2017 (Rangel et al., 2016), donde se incluyen tuits en distintas variantes del español y la información correspondiente al género de la persona autora.

En fase productiva, los tuits estarán disponibles a través de X en una API de consulta. El conjunto de la arquitectura queda resumido en la Figura 3.

Figura 3

Arquitectura del proyecto.



Planificación y estimación

Esta sección incluye ocho sprints de dos semanas que suman 16 semanas de acuerdo con la estrategia Scrum, además de dos semanas para actividades de inicio y dos semanas para la clausura que no han sido tenidas en cuenta porque suponen actividades de "engagement" con los implicados en el proyecto. Por lo tanto, el tiempo estimado de proyecto es de 20 semanas en total, siendo la planificación representada en la Figura 7 en un diagrama de Gantt.

Tabla 18

Diagrama de Gantt para el Proyecto

Actividad	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8	Semana 9	Semana 10	Semana 11	Semana 12	Semana 13	Semana 14	Semana 15	Semana 16	Semana 17	Semana 18	Semana 19	Semana 20
Inicio del Proyecto	X	X																		
Sprint 1			X	X																
Sprint 2					X	X														
Sprint 3							X	X												
Sprint 4									X	X										
Sprint 5											X	X								
Sprint 6													X	X						
Sprint 7															X	X				
Sprint 8																	X	X		
Cierre del Proyecto																			X	X

Para esto, se puede consultar la Tabla 19 en donde se recopilan los precios actuales de cada componente que se mencionó para la arquitectura del proyecto.

Tabla 19

Precios de los componentes de Microsoft Azure

Servicio	Precio (en dólares)	Comentario
Azure Blob Storage	$ 0.15 por GB	Servicio pay-as-you-go
Azure Data Lake Analytics	$ 2 por hora	Servicio pay-as-you-go
Azure Cognitive Services for Language	$ 700 por 1 millón de textos	Servicio estándar
Azure Databricks	$ 0.40 por hora	Servicio estándar
Azure Virtual Machine	$ 383.98 por mes	Utilizando la versión NCas_T4_v3 Series de 28 GB de RAM, con GPU para correr modelos de AI
Azure Data Factory	$ 1 por hora	Utilizando la versión 2 e incluyendo "pipeline activity"

Se cree que el proyecto manipulará un total de menos de un millón de textos a lo largo de su desarrollo, es decir, incluyendo una prueba piloto previa a la salida definitiva; lo que indica que el peso estimado de los datos es de un máximo de 100 GB. Si se toma en consideración, el coste total que se estimó sería de $5,082.90 tal como se expone en la Tabla 20.

Tabla 20

Estimado de costo del proyecto en dólares y por componente utilizado de Microsoft Azure

Servicio	Precio (en dólares)	Comentario	Uso estimado	Costo final (en dólares)
Azure Blob Storage	$ 0.15 por GB	Servicio pay-as-you-go	100 GB	$ 15.00
Azure Data Lake Analytics	$ 2 por hora	Servicio pay-as-you-go	720 horas	$ 1,440.00
Azure Cognitive Services for Language	$ 700 por 1 millón de textos	Servicio estándar	1 millón de textos	$ 700.00
Azure Databricks	$ 0.40 por hora	Servicio estándar	720 horas	$ 288.00
Azure Virtual Machine	$ 383.98 por mes	Utilizando la versión NCas_T4_v3 Series de 28 GB de RAM, con GPU para correr modelos de AI	5 horas	$ 1,919.90
Azure Data Factory	$ 1 por hora	Utilizando la versión 2 e incluyendo "pipeline activity"	720 horas	$ 720.00
			Total	**$ 5,082.90**

Despliegue de los resultados

En esta sección se detallan los elementos técnicos clave para el despliegue de la herramienta. El proyecto utiliza un modelo de Deep Learning basado en BERT multilingüe (Devlin et al., 2019) para predecir género, nacionalidad y temas de interés en usuarios de Twitter. Este modelo será ajustado (*fine-tuning*) con el corpus PAN 2017 (Rangel et al., 2016), que incluye tuits en distintas variantes del español y etiquetas de género. Así, se logra un modelo adaptado a las particularidades del idioma. Los resultados serán almacenados en Azure Data Lake y accesibles mediante una API, permitiendo a los usuarios visualizarlos fácilmente en herramientas como Excel o Power BI.

- **Plan de contingencias en el despliegue**

Varias posibles amenazas al objetivo del proyecto son recopiladas y son presentadas con medidas atenuadoras y, en adicción, también una matriz de riesgos que nos ayudará a evaluar el impacto y la probabilidad de que los riesgos se efectúen. Leemos así que una de las amenazas mayores es el bajo desempeño del modelo, que podría tener que ejecutarse en modo retrain (reentrenar el modelo) o modelo retune (ajustar el modelo); también el data drift, consecuencia de los cambios posibles en el lenguaje propio de Twitter por parte de la comunidad de usuarios de esa red.

El riesgo de que las predicciones no sean útiles para los usuarios de marketing obliga a rediseñar la propuesta. Amenazas menores serían las caídas de Azure, el hecho de que el usuario final no comprenda la herramienta (que podría solucionarse con un programa de capacitaciones) y la aparición de modelos nuevos que superen a BERT, lo que obligaría en este caso a mantener el modelo al día con los nuevos modelos de mayor desempeño.

Finalmente, en la Tabla 21, se muestra el resumen de todos los riesgos presentados con su respectiva valoración sobre el impacto y probabilidad de que ocurran.

Tabla 21
Matriz de riesgos del Proyecto

Riesgo	Descripción	Probabilidad	Impacto
1	Mal desempeño del modelo	**Media**	**Media**
2	Data drift	**Baja**	**Baja**
3	Temas predichos poco relevantes	**Baja**	**Alto**
4	Problemas con Microsoft Azure	**Baja**	**Baja**
5	Poca comprensión del usuario final	**Media**	**Baja**
6	Cambio en las tecnologías disponibles	**Muy Alta**	**Medio**

Mantenimiento

El mantenimiento referido al modelo que proponemos radicará solamente en el mantenimiento de su componente de inteligencia artificial, dado que la explotación del servicio cloud de Microsoft Azure hace que deleguemos el mantenimiento tanto del hardware como del software básico en la entidad proveedora. Referido al modelo, habrá que realizar cambios de vez en cuando para paliar el data drift, es decir, el cambio en el lenguaje en Twitter que podría afectar el desempeño del modelo.

También hay que tener en cuenta la API de Twitter, pues si se producen cambios en su estructura hay que hacer cambios en el código. Por último, habrá que considerar la evolución tecnológica, ya que probablemente se produzcan modelos más avanzados que BERT, lo que haría necesario hacer cambios en la arquitectura para seguir manteniendo la pertinencia del sistema. En la medida en la que consideramos lo anterior, se estima que el ciclo de vida del proyecto sea de unos cinco años hasta que haya que renovar por completo la propuesta presentada.

Validación y diseño experimental

Este capítulo presenta la definición del Mínimum Viable Product (MVP) del trabajo entendido como la mejor muestra de éxito del proyecto. El MVP estará compuesto por tres modelos. El primero de ellos, para predecir el género; el segundo, para identificar la nacionalidad del autor, de acuerdo con la variante de español, y el tercero, para la identificación de sus temas de interés. Para los dos primeros modelos se estipula un mínimo del 85 % de accuracy como el umbral mínimo a alcanzar. Para el modelo no supervisado, que predice los intereses, se validará la predicción mediante el feedback continuo de los usuarios, en el marco de los sprints, siguiendo metodologías Scrum y LEAN Startup.

De esta manera, se garantiza la mejora del producto, a través de una mejora iterativa y de su aceptación. Finalmente, se considerará que el MVP es exitoso si al menos el 85 % de los usuarios

que lo prueban a lo largo del proyecto lo aprueban.

Conclusiones y trabajo futuro

De esta investigación, podemos extraer dos conclusiones muy importantes. En primer lugar, el uso de la inteligencia artificial en el marketing está más que justificado, pudiendo decir que el perfilado de los consumidores y la personalización del tipo de campañas basadas en bases de datos es una oportunidad estratégica a la que las empresas que lo utilicen se tendrán que adaptar y cuya responsabilidad moral tienen que asumir, debería tenerse en cuenta que esto no significa moralizar la investigación señalada, sino que se tiene que tener una fuerte responsabilidad con lo que decidimos hacer con esos datos.

En segundo lugar, el modelo estadístico que se modeliza en esta investigación podría y debería tener un margen importante de mejora, sobre todo en base a las nuevas tecnologías como podría ser la integración con GPT-4 para obtener mejores predicciones junto con el uso de variables que no se han utilizado para la obtención de predicciones que no son satisfactorias (metadatos, tipos sociales y socioeconómicos, etc.) o incluso como el uso de la inteligencia artificial para hacer la obtención de predicciones cada vez más contextualizadas, pero teniendo siempre en consideración al uso ético y responsable de los datos en modelos de inteligencia artificial.

Referencias

Chandan , L. (2019). *The Scrum Master Guidebook A Reference for Obtaining Mastery. Notion Press . https://books.google.com.pe/books?id=nsnBDwAAQBAJ&source=gbs_navlinks_s*

Chiu, F., Sandroni, R. y Paraboni, I. (2018). *Author profiling from facebook corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation . https://aclanthology.org/L18-1407.pdf*

Davenport, T., Guha A. y Grewal, D. (2021). *How to Design an AI Marketing Strategy*. Harvard Business Review. https://hbr.org/2021/07/how-to-design-an-ai-marketingstrategy

Delmondes, J. P. y Paraboni, I. (2022). *Multi-source BERT stack ensemble for cross-domain author profiling*. *Expert Systems*, volumen(*39)*. https://doi.org/10.1111/exsy.12869

Devlin , J., Chan , M., Lee , K. y Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT* , 4171-4186 .https://stewarthu.com/papers/LLM/bert.pdf

Jiménez-Villar, V. (2020). *Aumento de datos para tareas relacionadas al perfilado de autor.* https://inaoe.repositorioinstitucional.mx/jspui/handle/1009/2164

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M. y Stein, B. (2016). *Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations*. https://ceur-ws.org/Vol-1609/16090750.pdf

Mamgain, S., Balabantaray, R. y Das, A. K. (2019). *Author profiling: Prediction of gender and language variety from document. In 2019 International Conference on Information Technology* , 473-477. https://ieeexplore.ieee.org/abstract/document/9031930

Mendivelso, H. y Lobos, F. (2019). *La evolución del marketing: una aproximación integral - Revista Chilena de Economía y Sociedad.* https://rches.utem.cl/articulos/la-evolucion-del-marketing-una-aproximacion-integral/

Sarella,V.,Srivastava, S.,Jamberi , K. y Syed Khasim (2022).*Datascience.* GCS .https://www.google.com.pe/books/edition/DATA_SCIENCE/hFduEAAAQBAJ?hl=es-419&gbpv=0&kptab=overview

Subra , J y Vannieuwenhuyze , A.(2018). *Scrum un método ágil para sus proyectos*. ENi. Barcelona. España. https://books.google.com.pe/books?id=TyQuFpGhZ8sC&source=gbs_navlinks_s

Veenhoven, R., Snijders, S., van der Hall, D. y van Noord, R. (2018). *Using translated data to improve deep learning author profiling models. In Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, volumen(2125). https://pure.rug.nl/ws/portalfiles/portal/78994327/paper_178.pdf