Claim-Classifier-Tiktok Project

📊 Technical proposal for predictive classification of user-generated claims on social media

Overview

Project Title: Claim-Classifier-Tiktok
Client: TikTok Data Team

🎯 **Objetivo**: Desarrollar un modelo predictivo que clasifique si el contenido generado por usuarios contiene una afirmación factual o una opinión, con el fin de apoyar los flujos de moderación y reducir el volumen de reportes acumulados.

Este proyecto se encuentra en fase inicial de planificación e incluye análisis exploratorio de clasificación de contenido, construcción de hitos y simulación de flujos operativos. El informe contempla antecedentes, metodología y resultados simulados.

Se utilizará el marco PACE para guiar cada fase: desde la planificación y el análisis exploratorio hasta la construcción del modelo y la comunicación con stakeholders. Los entregables incluyen documentación del modelo, visualizaciones simuladas y un repositorio GitHub estructurado para reproducibilidad y mentoring.

Problem

TikTok actualmente carece de una solución confiable y basada en datos para clasificar eficientemente contenido generado por usuarios que pueda contener afirmaciones o juicios de opinión. El sistema de moderación existente depende en gran medida de la revisión manual de reportes, lo que genera demoras, inconsistencias y una acumulación creciente de contenido marcado.

La complejidad del contenido multimedia —desde matices lingüísticos hasta formatos diversos— dificulta la automatización sin un modelo predictivo robusto.
El marco propuesto busca evaluar modelos de clasificación e integrarlos en los flujos de moderación. Esta falta de automatización limita la transparencia, ralentiza los tiempos de respuesta, aumenta la carga operativa interna y reduce la eficacia del sistema de moderación.

Un modelo predictivo sólido permitirá priorizar reportes, reducir operaciones manuales y mejorar la seguridad y equidad de la plataforma.

Solution

Para abordar el desafío de clasificar eficientemente afirmaciones generadas por usuarios en TikTok, se propone desarrollar un modelo predictivo basado en regresión, utilizando datos históricos de moderación. Este modelo permitirá distinguir entre afirmaciones factuales y contenido de opinión, optimizando los flujos de revisión y priorización.

🔑 Prioridades del Enfoque

- Reproducibilidad Todo el código, documentación y flujos de trabajo estarán estructurados en un repositorio modular en GitHub, facilitando la colaboración transparente y el mentoring técnico.
- Modelado Ético El proyecto incluirá estrategias para mitigar sesgos, documentación clara de supuestos y comunicación adaptada a los stakeholders.
- Visualizaciones orientadas al público Los entregables incluirán dashboards bilingües y resúmenes adaptados tanto para públicos técnicos como no técnicos.
- Utilidad institucional El modelo final y su documentación serán adaptables para integrarse en los flujos de moderación de TikTok, mejorando la priorización de reportes y fortaleciendo la confianza en la plataforma.

🧭 Nota final
Este enfoque se alinea con la misión de TikTok de inspirar creatividad y generar alegría, al mismo tiempo que refuerza la transparencia y el compromiso con procesos de moderación responsables.

Claim-Classifier-Tiktok Project

📊 Technical proposal for predictive classification of user-generated claims on social media

Details

El clasificador predictivo basado en regresión, desarrollado para apoyar los flujos de moderación de TikTok, identificó varias características que influyen directamente en la probabilidad de que un video contenga una afirmación factual:

- **Longitud del texto:** Las descripciones o transcripciones más extensas tienden a incluir argumentos estructurados o declaraciones explícitas.
- **Polaridad del sentimiento:** Los tonos neutros o asertivos reflejan con mayor frecuencia afirmaciones factuales, en contraste con contenido emocional o subjetivo.
- **Presencia de palabras clave:** Términos como “estudio”, “datos”, “informe” o “evidencia” indican posibles afirmaciones.
- **Métricas de interacción:** Videos con alto volumen de comentarios o compartidos pueden reflejar contenido polémico o cargado de afirmaciones.
- **Momento y contexto de publicación**: El contenido publicado durante ciclos de noticias o eventos en tendencia presenta mayor densidad de afirmaciones.

Estas variables reflejan dimensiones lingüísticas, conductuales y contextuales del contenido generado por usuarios. El modelo fue comparado con un clasificador base de regresión logística, demostrando un rendimiento superior tanto en precisión como en interpretabilidad.

🔄 Flujo de Moderación con Integración Predictiva (Visualización Simulada)

Esta estructura proyectada ilustra cómo el modelo Claim-Classifier será integrado en los flujos operativos de TikTok. Se presenta como parte de la fase de planificación dentro del marco metodológico PACE.

Moderation workflow with claim-classifier integration

🔹 **Reporte de Usuario**
Contenido marcado como posible desinformación.

🔹 **Clasificador Predictivo**
El modelo determina si el contenido contiene una afirmación factual o una opinión.

🔹 **Priorización Automatizada**
El sistema asigna un nivel de urgencia según el tipo de contenido y su nivel de interacción.

📝 **Nota**
Esta visualización es simulada y no representa resultados reales. Su propósito es comunicar la lógica operativa proyectada.

Claim-Classifier-Tiktok Project

📊 Technical proposal for predictive classification of user-generated claims on social media

📊 Desempeño del Modelo e Integración Institucional

🔹 **Métricas Clave**

- Mejora en la precisión y el recall para la detección de afirmaciones
- F1-score estable en los conjuntos de validación
- Importancia de variables transparente para revisión por parte de stakeholders

🔹 **Alineación Institucional**

- Cumple con los requisitos internos de TikTok para apoyar los flujos de moderación
- Recomendado para su integración en el sistema de priorización de reportes de la plataforma
- Estructura reproducible y documentación bilingüe que facilitan la colaboración transversal y el mentoring

🎯 Modelo PACE – Marco Técnico Reproducible

PACE Plan – Analyze – Construct – Execute

Esta estructura proyectada ilustra cómo el modelo Claim-Classifier se integrará en los flujos operativos de TikTok. Forma parte de la fase de planificación bajo el marco metodológico PACE, y comunica la lógica operativa prevista para la clasificación predictiva de contenido generado por usuarios.

Fase	Propósito Técnico
Planificar	Definir el objetivo del modelo, identificar stakeholders técnicos, establecer entregables reproducibles y seleccionar herramientas (Python, GitHub, Markdown).
Analizar	Explorar el conjunto de datos TLC, seleccionar variables relevantes, limpiar y generar métricas, y realizar análisis exploratorio de datos (EDA).
Construir	Entrenar modelos de regresión (XGBoost, regresión lineal), validar con métricas (MAE, RMSE, R^2), simular escenarios y documentar supuestos.
Ejecutar	Presentar resultados técnicos, vincular notebooks con entregables reproducibles (por ejemplo, GitHub, README), redactar resúmenes para públicos diversos y preparar archivos README bilin para revisión institucional.

El marco PACE garantiza un desarrollo de modelos ético, reproducible y colaborativo. Cada fase está diseñada para apoyar la documentación bilingüe, la integración institucional y el mentoring entre equipos técnicos y no técnicos.

Claim-Classifier-Tiktok Project

📊 Technical proposal for predictive classification of user-generated claims on social media

🎯 Variables Clave del Modelo – Clasificación de Afirmaciones en TikTok

📌 Las variables fueron seleccionadas tras un análisis exploratorio de reportes de contenido en TikTok, priorizando aquellas con mayor correlación con afirmaciones factuales y relevancia para los flujos de moderación. Se aplicaron transformaciones para mejorar la interpretabilidad, la priorización y la reproducibilidad del modelo.

📊 Tabla de variables

Variable	Descripción	Tipo
Claim_Likelihood	Probabilidad estimada de que el contenido contenga una afirmación factual	Numérica
text_length	Número de caracteres o palabras en el contenido	Numérica
sentiment_polarity	Tono emocional del contenido (neutral, positivo o negativo)	Numérica
keyword_presence	Presencia de palabras clave (ej. “evidencia”, “por primera vez”, “informe”)	Categórica
engagement_volume	Número de comentarios, compartidos y “me gusta”	Numérica
posting_context	Si el contenido fue publicado durante una crisis o evento en tendencia	Categórica
content_type	Tipo de medio utilizado (texto, video, audio, etc.)	Categórica

📌 Estas variables fueron seleccionadas tras un análisis exploratorio de reportes de contenido en TikTok, priorizando su correlación con afirmaciones factuales y su relevancia para la moderación. El objetivo fue mejorar la interpretabilidad, priorización y rendimiento del modelo.

Claim-Classifier-Tiktok Project

📊 Technical proposal for predictive classification of user-generated claims on social media

📊 Cronograma Técnico y Herramientas – Modelo Claim-Classifier

📌 Este cronograma describe las fases técnicas para el desarrollo del modelo Claim-Classifier dentro del flujo de moderación de TikTok. Cada fase respalda la reproducibilidad, la documentación bilingüe y la integración institucional bajo el marco metodológico PACE.

📊 Table: Technical Timeline & Tools

Fase	Actividades Técnicas	Duración Estimada	Librerías / Herramientas Sugeridas
Planificar	Definir el objetivo del modelo, identificar stakeholders técnicos, establecer entregables reproducibles y estructura de documentación.	1 semana	`os`, `pathlib`, `markdown`, `numpy`, `matplotlib`
Analizar	Explorar contenido reportado en TikTok, seleccionar variables relevantes, generar visualizaciones iniciales y documentar criterios de exclusión.	1 semana	`pandas`, `numpy`, `matplotlib`, `seaborn`
Construir	Entrenar modelos de clasificación (XGBoost, regresión logística), validar con métricas (precisión, recall, F1-score), simular escenarios y documentar supuestos.	1 semana	`xgboost`, `scikit-learn`, `joblib`, `statsmodels`
Ejecutar	Presentar resultados técnicos, vincular notebooks con entregables reproducibles, diseñar visualizaciones para públicos diversos y preparar README bilingües.	1 semana	`matplotlib`, `seaborn`, `notebook`, `markdown`

📌 Esta estructura garantiza un desarrollo de modelos ético y escalable, alineado con las necesidades de moderación de TikTok. Cada fase es modular, reproducible y está diseñada para facilitar la colaboración transversal y los procesos de mentoring.

Taxi Fare Estimation – Automatidata Project

Technical proposal for fare modeling using regression

👥 Stakeholders – Automatidata & TLC

👥 Stakeholder Table

Internal – Automatidata	Role
Marcelo Dominguez	Data Analytics Consultant
Udo Bankole	Director of Analysis
Deshawn Washington	Data Analysis Manager
Luana Rodriguez	Senior Analyst
Uli King	Project Manager

External – TikTok (Platform Context)	Role
Juliana Soto	Content Policy & Governance
Titus Nelson	Moderation Operations Lead

Este equipo transversal apoya el desarrollo e integración del modelo Claim-Classifier en el flujo de moderación de TikTok. La colaboración garantiza una implementación ética, reproducible y escalable, alineada con las necesidades de la plataforma.

⚠️ Limitaciones y Supuestos – Modelo Claim-Classifier

Esta sección describe las limitaciones actuales y los supuestos técnicos relacionados con el modelo propuesto para el flujo de moderación de TikTok.

📌 **Puntos Clave**

- **Estado del Modelo**: El modelo XGBoost aún no ha sido entrenado. Su uso se propone como estrategia técnica para la fase “Construir” del marco PACE.
- **Visualizaciones Simuladas:** Los gráficos presentados (por ejemplo, importancia de variables) son simulaciones basadas en supuestos técnicos y no reflejan salidas reales del modelo.
- **Selección de Variables**: Las variables independientes fueron seleccionadas en función de su relación esperada con la probabilidad de afirmación, informada por análisis exploratorio preliminar de contenido reportado en TikTok.
- **Validación Pendiente:** Las métricas de desempeño y las comparaciones entre modelos (precisión, recall, F1-score) serán validadas en fases posteriores del proyecto.

📌 Estos supuestos se documentan para garantizar transparencia y reproducibilidad. Las salidas simuladas funcionan como andamiaje para revisión institucional y procesos de mentoring, en línea con prácticas éticas de modelado.