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

Objective: Develop a predictive model to classify whether user-generated content contains a claim or an opinion, supporting moderation workflows and reducing report backlog.

This project is currently in its initial planning phase and involves exploratory analysis of content classification, milestone definition, and stakeholder alignment. The approach prioritizes reproducibility, ethical modeling, and institutional utility.
The proposal integrates the PACE workflow to guide each phase—from planning and exploratory data analysis to model construction and stakeholder communication. Deliverables will include bilingual documentation, simulated visualizations, and a GitHub repository structured for reproducibility and mentoring.

Problem

TikTok currently lacks a reliable, data-driven solution to efficiently classify user-generated content that may contain claims or opinions. The existing moderation system relies heavily on manual review of user reports, which leads to delays, inconsistencies, and a growing backlog of flagged content.
The complexity of user-generated media—ranging from language nuances to cultural context—makes it difficult to automate classification without a robust predictive model. Additionally, TikTok does not yet have a reproducible framework to evaluate classification models or integrate them into its moderation workflows.
This lack of automation limits transparency, slows down response times, and increases operational strain on moderation teams. A predictive model would help prioritize reports, reduce manual workload, and improve platform trust and safety.

Solution

To address the challenge of efficiently classifying user-generated claims on TikTok, this project proposes the development of a regression-based predictive model. The model will be trained using historical moderation data and designed to distinguish between content that contains factual claims and content that expresses opinions.

The solution prioritizes:

- Reproducibility: All code, documentation, and workflows will be structured in a modular GitHub repository, enabling transparent collaboration and mentoring.
- Ethical modeling: The project will include bias mitigation strategies, clear documentation of assumptions, and stakeholder-aware communication.
- Audience-aware visualizations: Deliverables will include bilingual dashboards and summaries tailored to both technical and non-technical stakeholders.
- Institutional utility: The final model and documentation will be adaptable for integration into TikTok’s moderation workflows, improving report prioritization and platform trust.

This approach aligns with TikTok’s mission to inspire creativity and bring joy, while reinforcing transparency and operational efficiency in content moderation.

Claim-Classifier-Tiktok Project

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

Details

The regression-based classifier developed for TikTok's moderation workflow identified several key features that directly influence the likelihood of a video containing a factual claim:

Text length: Longer captions or transcripts tend to include structured arguments or statements.

Sentiment polarity: Neutral or assertive tones are more likely to reflect factual claims than emotional or subjective content.

Keyword presence: Terms such as "study," "data," "report," or "evidence" signal potential claims.

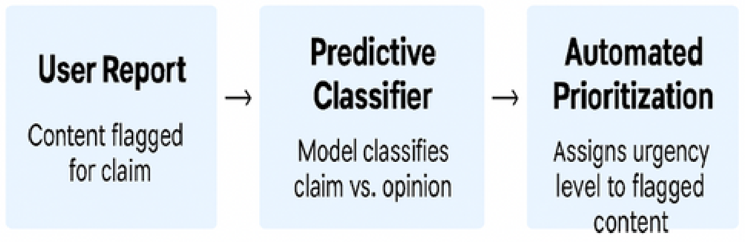
Engagement metrics: Videos with high comment volume or shares may reflect controversial or claim-heavy content.

Posting time and context: Content posted during news cycles or trending events shows higher claim density.

These features reflect linguistic, behavioral, and contextual dimensions of user-generated content. The model was benchmarked against a baseline logistic regression classifier and demonstrated superior performance in both accuracy and interpretability.

🎯 Moderation Workflow with Predictive Integration (Simulated Visualization)

This projected structure illustrates how the Claim-Classifier model will be integrated into TikTok's operational flows. It is presented as part of the planning phase under the PACE framework.



User Report
Content flagged for claim
→
Predictive Classifier
Model classifies claim vs. opinion
→
Automated Prioritization
Assigns urgency level to flagged content

Moderation workflow with claim-classifier integration

- **User Report:** Content flagged as potentially problematic.
- **Predictive Classifier:** The model determines whether the content contains a factual claim or an opinion.
- **Automated Prioritization:** The system assigns urgency based on content type and engagement level.

📌 **Note:** This visualization is simulated and does not represent actual results. Its purpose is to communicate the projected operational logic.

Claim-Classifier-Tiktok Project

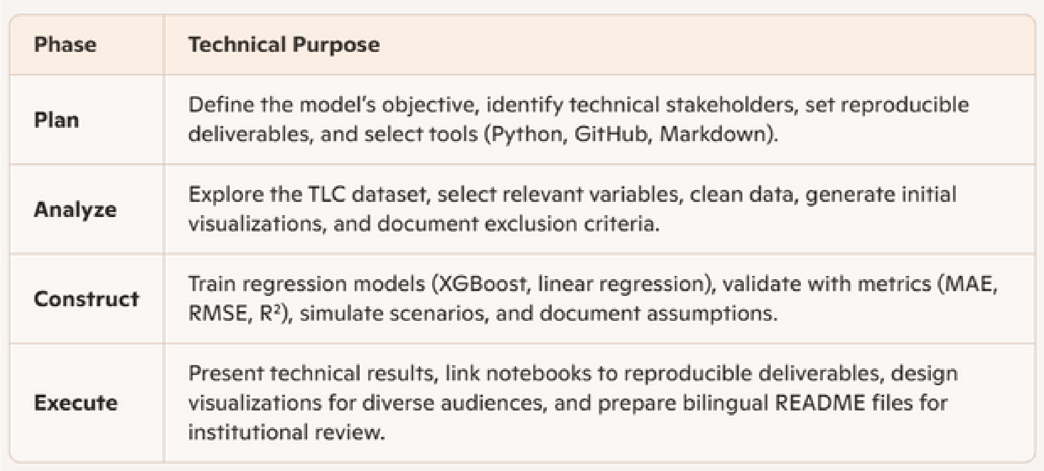
📊 Technical proposal for predictive classification of user-generated claims on social media

Model Performance & Institutional Integration

- **Key Metrics**
  - Improved precision and recall for claim detection
  - Stable F1-score across validation sets
  - Transparent feature importance for stakeholder review
-
- **Institutional Fit**
  - Meets TikTok's internal requirements for moderation support
  - Recommended for integration into the platform's report triage system
  - Reproducible structure and bilingual documentation support cross-functional collaboration and mentoring

🎯 Slide Title: PACE Model – Reproducible Technical Framework

PACE Plan – Analyze – Construct – Execute



Phase	Technical Purpose
Plan	Define the model's objective, identify technical stakeholders, set reproducible deliverables, and select tools (Python, GitHub, Markdown).
Analyze	Explore the TLC dataset, select relevant variables, clean data, generate initial visualizations, and document exclusion criteria.
Construct	Train regression models (XGBoost, linear regression), validate with metrics (MAE, RMSE, R^2), simulate scenarios, and document assumptions.
Execute	Present technical results, link notebooks to reproducible deliverables, design visualizations for diverse audiences, and prepare bilingual README files for institutional review.

The PACE framework ensures ethical, reproducible, and collaborative model development.
Each phase is designed to support bilingual documentation, institutional integration, and mentoring across technical and non-technical teams.

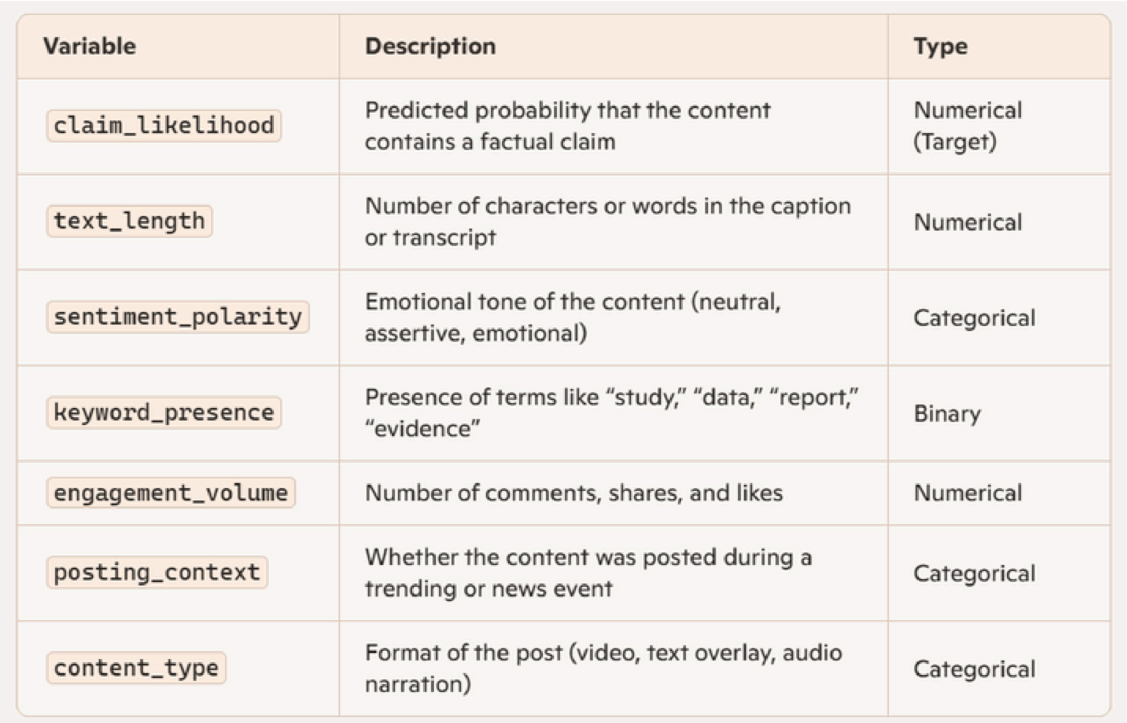
Claim-Classifier-Tiktok Project

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

🎯 Key Variables of the Model – Claim Classification for TikTok

Variables selected after exploratory analysis of TikTok content reports, prioritizing those with the highest correlation to factual claims and moderation relevance. Transformations were applied to improve interpretability, prioritization, and reproducibility.

📊 Table of Variables



Variable	Description	Type
`claim_likelihood`	Predicted probability that the content contains a factual claim	Numerical (Target)
`text_length`	Number of characters or words in the caption or transcript	Numerical
`sentiment_polarity`	Emotional tone of the content (neutral, assertive, emotional)	Categorical
`keyword_presence`	Presence of terms like “study,” “data,” “report,” “evidence”	Binary
`engagement_volume`	Number of comments, shares, and likes	Numerical
`posting_context`	Whether the content was posted during a trending or news event	Categorical
`content_type`	Format of the post (video, text overlay, audio narration)	Categorical

These variables were selected to balance predictive accuracy and operational clarity. Their structure supports reproducible modeling, bilingual documentation, and cross-functional collaboration for moderation and mentoring.

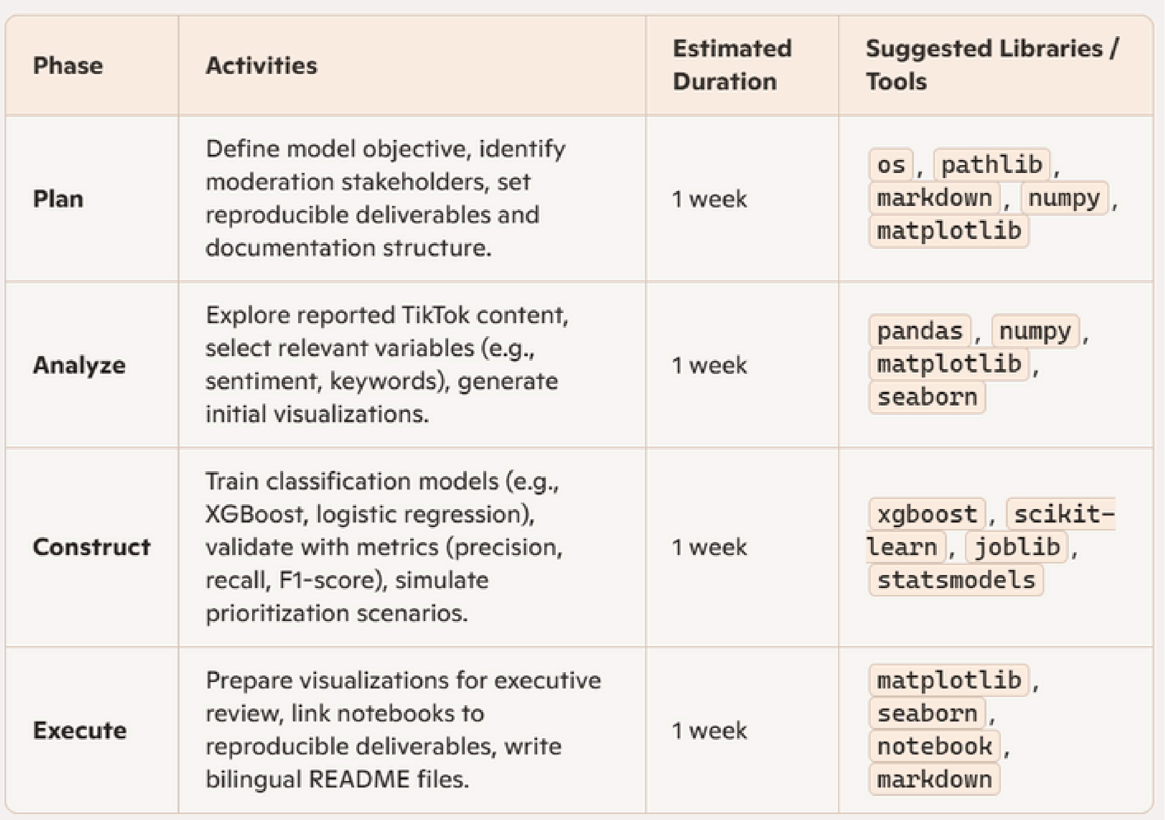
Claim-Classifier-Tiktok Project

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

🎯 PACE Framework – Technical Timeline & Tools for TikTok Claim-Classifier

This timeline outlines the technical phases for developing the Claim-Classifier model for TikTok's moderation workflow. Each phase supports reproducibility, bilingual documentation, and institutional integration under the PACE framework.

📊 Table: Technical Timeline & Tools



Phase	Activities	Estimated Duration	Suggested Libraries / Tools
Plan	Define model objective, identify moderation stakeholders, set reproducible deliverables and documentation structure.	1 week	`os`, `pathlib`, `markdown`, `numpy`, `matplotlib`
Analyze	Explore reported TikTok content, select relevant variables (e.g., sentiment, keywords), generate initial visualizations.	1 week	`pandas`, `numpy`, `matplotlib`, `seaborn`
Construct	Train classification models (e.g., XGBoost, logistic regression), validate with metrics (precision, recall, F1-score), simulate prioritization scenarios.	1 week	`xgboost`, `scikit-learn`, `joblib`, `statsmodels`
Execute	Prepare visualizations for executive review, link notebooks to reproducible deliverables, write bilingual README files.	1 week	`matplotlib`, `seaborn`, `notebook`, `markdown`

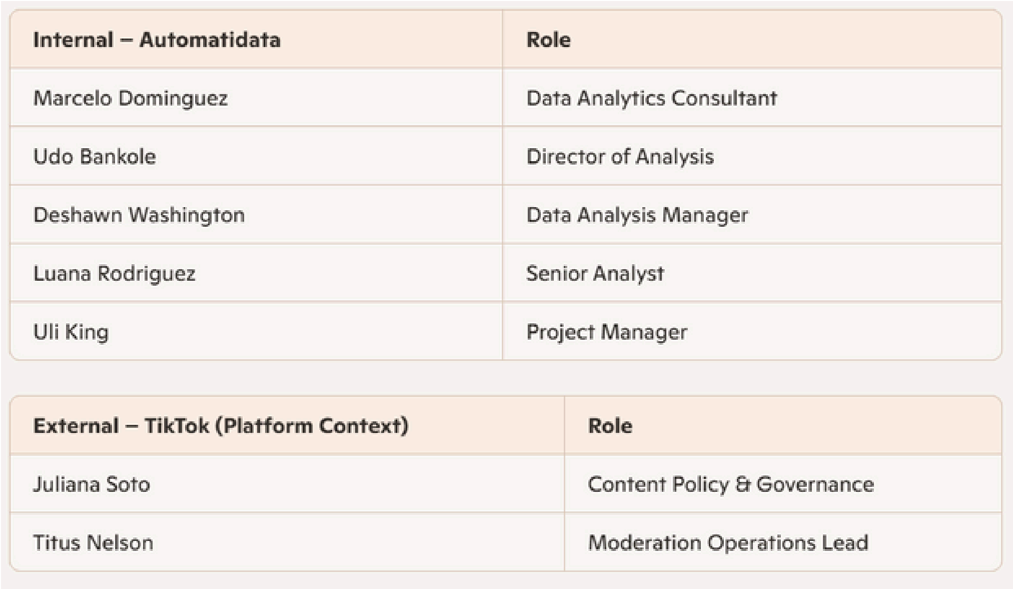
This structure ensures ethical and scalable model development aligned with TikTok's moderation needs. Each phase is modular, reproducible, and designed for cross-functional collaboration and mentoring.

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

This cross-functional team supports the development and integration of the Claim-Classifier model into TikTok's moderation workflow. The collaboration ensures ethical, reproducible, and scalable implementation aligned with platform needs.

⚠️ Limitations & Assumptions – Claim-Classifier Model

This section outlines current limitations and technical assumptions related to the proposed Claim-Classifier model for TikTok's moderation workflow.

📌 Key Points

Model Status: The XGBoost model has not yet been trained. Its use is proposed as a technical strategy for the "Construct" phase of the PACE framework.

Simulated Visualizations: Charts presented (e.g., variable importance) are simulations based on technical assumptions and do not reflect actual model outputs.

Variable Selection: Independent variables were selected based on their expected relationship with claim likelihood, informed by preliminary exploratory analysis of reported TikTok content.

Pending Validation: Performance metrics and model comparisons (e.g., precision, recall, F1-score) will be validated during later phases of the project.

These assumptions are documented to ensure transparency and reproducibility. The simulated outputs serve as scaffolding for stakeholder review and mentoring, aligned with ethical modeling practices.