Amazon Reviews Sentiment Analysis

Project Scoping

Problems

The project addresses several challenges businesses face (Amazon in this use case) when analyzing customer sentiment from user reviews:

- **Lack of Automation**: Sentiment analysis and insights generation from customer reviews often involve manual, time-consuming, and error-prone processes.
- **Data Scalability**: Handling, processing, and analyzing a growing volume of customer reviews at scale is difficult with traditional methods.
- **Sentiment Classification**: Current methods may not accurately categorize reviews into positive, negative, or neutral sentiments, impacting the quality of business decisions.
- **Monitoring and Model Retraining**: Machine learning models can degrade over time, and without automated monitoring and retraining, model performance can decline unnoticed.
- **Executive Insights**: Executives require detailed, real-time insights into customer satisfaction across various product categories, but existing dashboards lack the interactivity and detail needed for effective sentiment analysis.
- **Handling Edge Cases and Anomalies**: Reviews can be ambiguous, sarcastic, or irrelevant, making classification challenging, especially when dealing with complex or domain-specific language.

Current Solutions

Several existing solutions attempt to address these issues but have limitations:

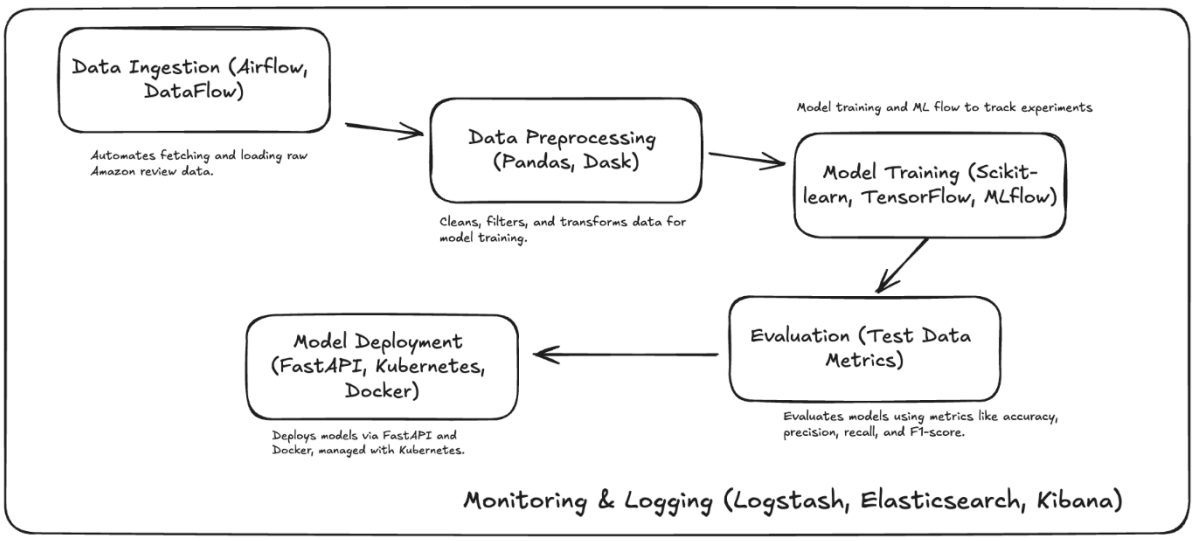
- **Manual Review Analysis**: Manually analyzing reviews doesn't scale for large datasets and can introduce biases.
- **Basic Sentiment Analysis Models**: Off-the-shelf models are available but often lack the nuance needed for specific domains like Amazon, leading to lower accuracy.
- **Traditional Dashboards**: Standard BI dashboards provide aggregated metrics but often miss specific insights related to customer sentiment by category.
- **Ad-hoc Monitoring**: Without automation, pipelines rely on manual monitoring, leading to delays in retraining models when performance drops.
- **Rigid Data Pipelines**: Many pipelines lack flexibility, making it difficult to adapt or improve without causing disruptions.

Proposed Solutions

The project will implement a robust MLOps pipeline that automates, scales, and customizes sentiment analysis with the following innovations:

- **Fully Automated MLOps Pipeline**: Using Airflow, Docker, Kubernetes, and GCP, the project will automate data ingestion, pre-processing, model training, evaluation, and deployment, ensuring scalability and reliability.
- **Scalable Cloud Infrastructure**: Google Cloud and Kubernetes will provide scalability to handle the growing volume of Amazon reviews, ensuring seamless performance.
- **Custom Sentiment Classification**: Custom sentiment classification models will be built using TensorFlow, designed to classify reviews as positive, negative, or neutral with greater accuracy by addressing domain-specific language (terms and phrases unique to Amazon's product reviews) and handling edge cases (ambiguous reviews, sarcasm, and spam content).
- **MLFlow for Experiment Tracking**: MLFlow will be used to track experiments, allowing for efficient comparison of different model configurations, hyperparameters, and performance metrics over time.
- **CI/CD with GitHub Actions**: Continuous integration and deployment will ensure that code changes, model updates, and configuration adjustments are automatically tested, validated, and deployed without manual intervention.
- **Comprehensive Monitoring with Kibana**: Real-time monitoring will be implemented using Kibana to track data quality, model performance, and detect anomalies, triggering retraining workflows as necessary.
- **Modular Pipeline Design**: Each component of the pipeline—data ingestion, preprocessing, modeling, deployment—will be designed to be modular and replaceable, enabling flexibility and easy integration of new components without breaking the pipeline.
- **Interactive Executive Dashboard**: The final product will include a dashboard that enables executives to drill down by product category, track sentiment over time, and interactively summarize reviews using a Retrieval-Augmented Generation (RAG) pipeline, providing actionable insights at both high-level and granular levels.

6. Current Approach Flowchart and Bottleneck Detection



Bottlenecks and Improvements

1. **Data Ingestion**:
    - **Bottleneck**: Slow ingestion due to large data volumes.

    - **Improvement**: Use batch processing and increase task concurrency in Airflow.
2. **Data Preprocessing**:
    - **Bottleneck**: Pandas struggles with large datasets.
    - **Improvement**: Use Dask for distributed data processing.
3. **Model Training**:
    - **Bottleneck**: Long training times on large datasets.
    - **Improvement**: Use distributed training on GCP (TPU/TF Distributed Strategy).
4. **Evaluation**:
    - **Bottleneck**: Limited test data affects generalization.
    - **Improvement**: Use cross-validation and regularly refresh test sets.
5. **Model Deployment**:
    - **Bottleneck**: High latency during deployment.
    - **Improvement**: Use Kubernetes rolling updates and autoscaling.
6. **Monitoring & Logging**:
    - **Bottleneck**: Delayed detection of model drift or pipeline issues.
    - **Improvement**: Set up real-time alerts for drift and performance drops.

Pipeline Bottlenecks

1. **Edge Cases**:
    - **Problem**: Model misclassifies ambiguous reviews.
    - **Improvement**: Log and address edge cases.
2. **Pipeline Breaks**:
    - **Problem**: Pipeline failure at any stage can stop processing.
    - **Improvement**: Use retry policies and circuit breakers to handle failures.
3. **Scaling**:
    - **Problem**: Struggles with large datasets or high traffic.
    - **Improvement**: Use distributed pipelines and GCP autoscaling.
4. **Model Drift**:
    - **Problem**: Model performance degrades over time.
    - **Improvement**: Monitor drift and automate retraining.

7. Metrics, Objectives, and Business Goals

Key Metrics

- **Accuracy**: Overall correctness in classifying reviews as positive, negative, or neutral.
- **Precision**: Correctly identifies positive/negative reviews without misclassifying neutral ones.

- **Recall**: Captures all relevant reviews in each sentiment category.
- **F1 Score**: Balances precision and recall, crucial for class imbalance.
- **True Negative Rate (Specificity)**: Accurately identifies negative reviews, critical for addressing customer dissatisfaction.

Project Objectives

- **Accurate Sentiment Classification**: Improve review classification for better organization and insights.
- **Product Insights**: Provide actionable insights for Amazon and sellers to enhance product offerings.
- **Customer Experience**: Help Amazon address customer issues through accurate sentiment tracking.
- **Data-Driven Decisions**: Use sentiment analysis to guide product improvements and business strategies.
- **Increase Sales and Reduce Returns**: Identifying and addressing negative feedback will drive better customer retention and lower return rates.

Operational Efficiency Objectives

- **Process Automation**: Automate the entire MLOps pipeline (data ingestion, preprocessing, training, deployment) to reduce manual intervention and ensure scalability.
- **Pipeline Health Monitoring**: Ensure real-time monitoring using metrics like:
    - **Latency**: The time taken for the model to process and classify reviews, aiming for low response times under load.
    - **Throughput**: The number of reviews processed per second, ensuring the system handles large datasets efficiently.
    - **Error Rates**: Track errors or failed model predictions, ensuring consistent performance.
    - **Resource Utilization**: Monitor CPU, GPU, and memory usage to optimize cost and performance.
    - **Retraining Triggers**: Set thresholds for automatic model retraining when performance drops below acceptable levels (e.g., F1 score or accuracy falls below a certain point).

Aligning Metrics with Business Goals

- **True Negative Rate**: Improves customer satisfaction by identifying and resolving negative reviews.
- **F1 Score**: Provides balanced sentiment classification for actionable insights.
- **Pipeline Efficiency**: High throughput and low latency ensure timely feedback processing, supporting real-time decision-making.
- **Dashboard Insights**: Offer real-time sentiment trends, enabling timely business interventions for improved performance. Additionally, it will feature interactive review summaries that allow users to extract key themes, identify customer pain points, and gain deeper insights into feedback, driving actionable business strategies.