



# Real-Time Sentiment Analysis of Reddit Movie Comments

High Performance Data Processing

**Group A**





# Introduction

Real-time data pipelines enable instant feedback and analysis.

Sentiment classification helps understand audience opinions effectively.

Machine learning models enhance the accuracy of sentiment detection.

Visualization tools reveal valuable insights from processed data.



# Project Objectives

Gathering data from various online movie discussion platforms.

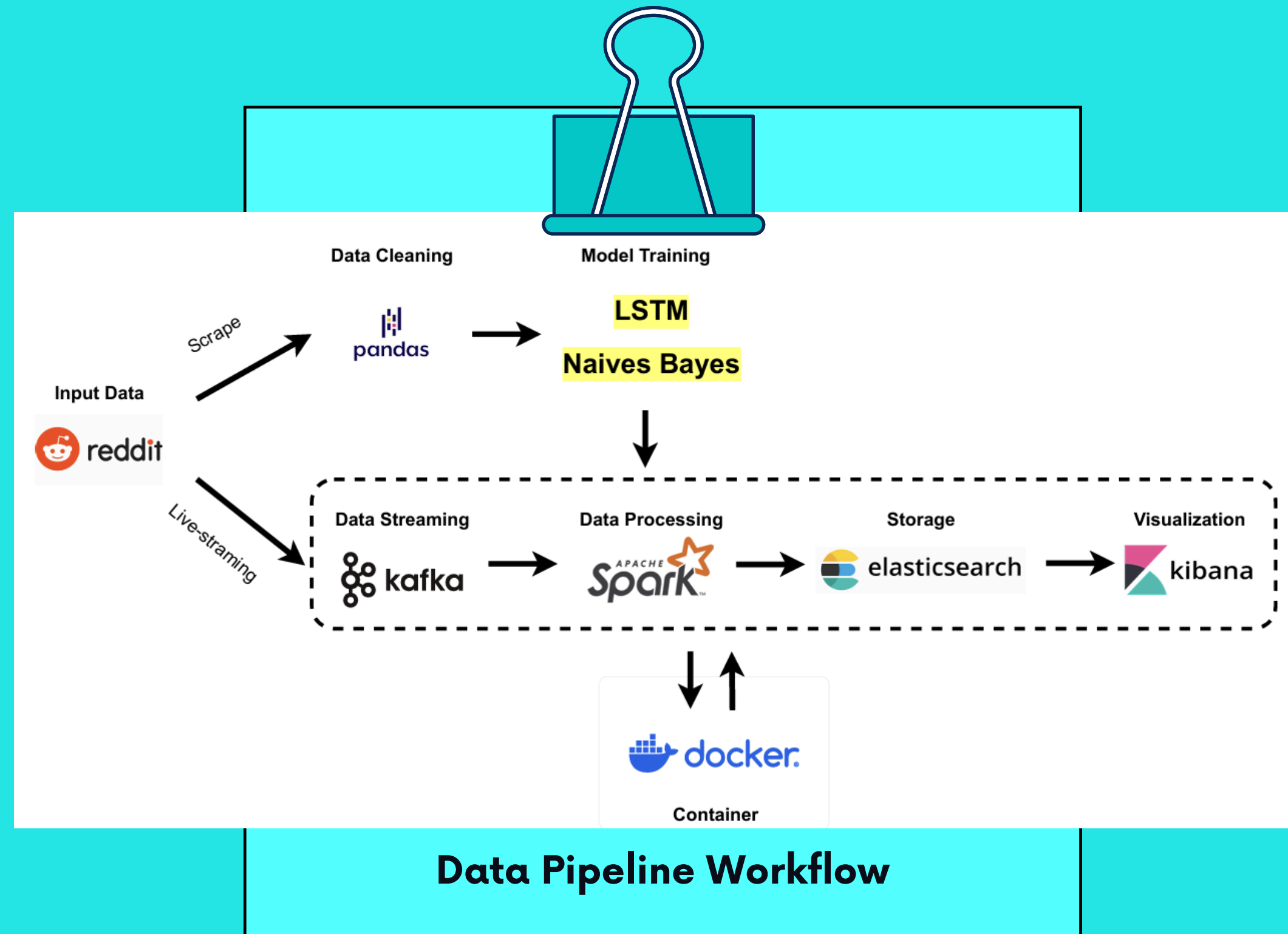
Developing and comparing traditional and deep learning sentiment models.

Implementing a continuous real-time processing pipeline for analysis.

Visualizing results to uncover patterns in audience sentiment.

# System Architecture & Workflow

- **Kafka:** Ingests real-time comments from the Reddit API.
- **Spark:** Consumes data from Kafka, cleans it, and applies the sentiment analysis model.
- **Elasticsearch & Kibana:** Stores the processed data and visualizes the results on an interactive dashboard.


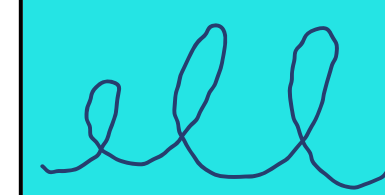




# Data Acquisition & Preprocessing

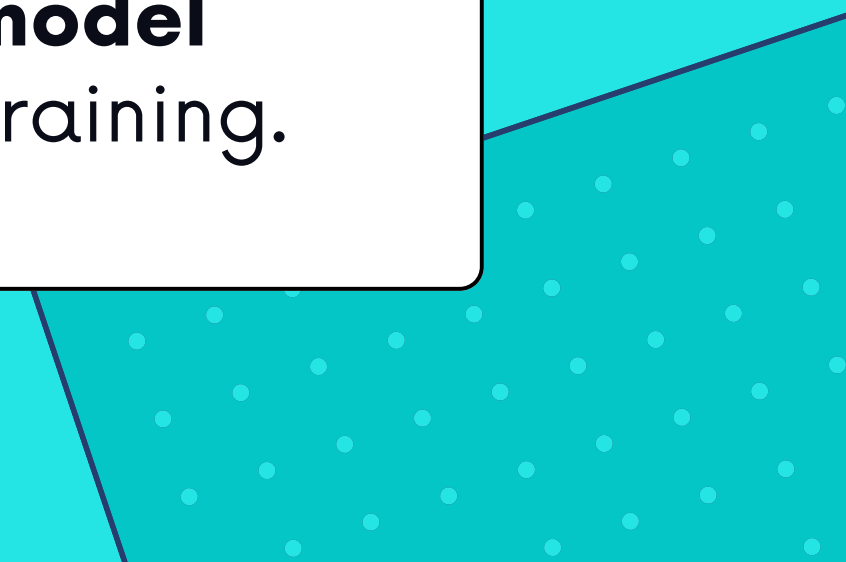
API-based scripts facilitate **efficient data collection** from online forums.

Cleaning workflows ensure **high-quality text** by removing irrelevant content.



Tokenization helps in breaking down text into **meaningful units** for analysis.

Normalization standardizes text format, enhancing **model performance** during training.





# Sentiment Model Development

Model Training Techniques: Selecting appropriate algorithms is crucial for accuracy.

Data Quality: Clean, well-prepared data significantly impacts model performance.

Evaluation Metrics: Using confusion matrices helps assess model effectiveness.

Continual Learning: Updating models with new data improves sentiment analysis over time.

# Model Evaluation & Comparison

Accuracy: A measure of correctly classified instances in the dataset.

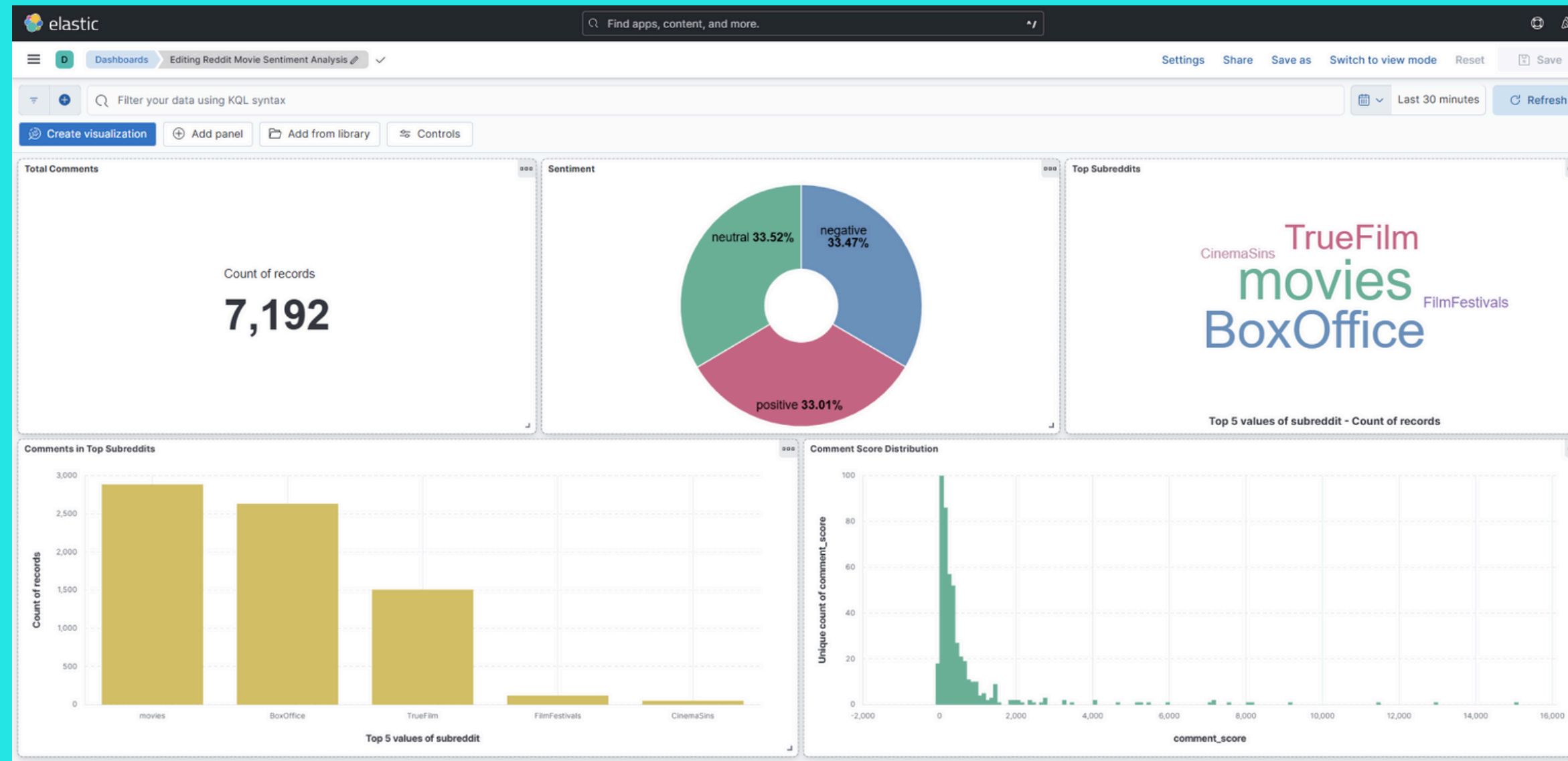
Precision: Evaluates the proportion of true positives among all predicted positives.

Recall: Focuses on the proportion of true positives detected out of actual positives.

F1 Score: A balanced measure combining precision and recall for model performance.



# Key Insights from Dashboard



**Total Comments Analyzed:**  
7,192

**Balanced Sentiment:**  
Neutral (33.52%),  
Negative (33.47%),  
Positive (33.01%).

**Top Subreddits:**  
'r/movies' and  
'r/BoxOffice' dominate  
discussions.



# Key Takeaways

The real-time data pipeline significantly enhances sentiment analysis efficiency.

Bi-LSTM model outperforms Naïve Bayes in accuracy and context understanding.

Visual dashboards provide valuable insights into audience sentiment trends.

Future improvements can further refine model performance and data scope.

# Future Work

Enhance model performance through advanced transformer techniques.

Broaden data sources to include diverse movie discussions.

Optimize processing pipeline for increased throughput and efficiency.

Explore additional analytics features for deeper sentiment insights.



# Thank You!

Group A

