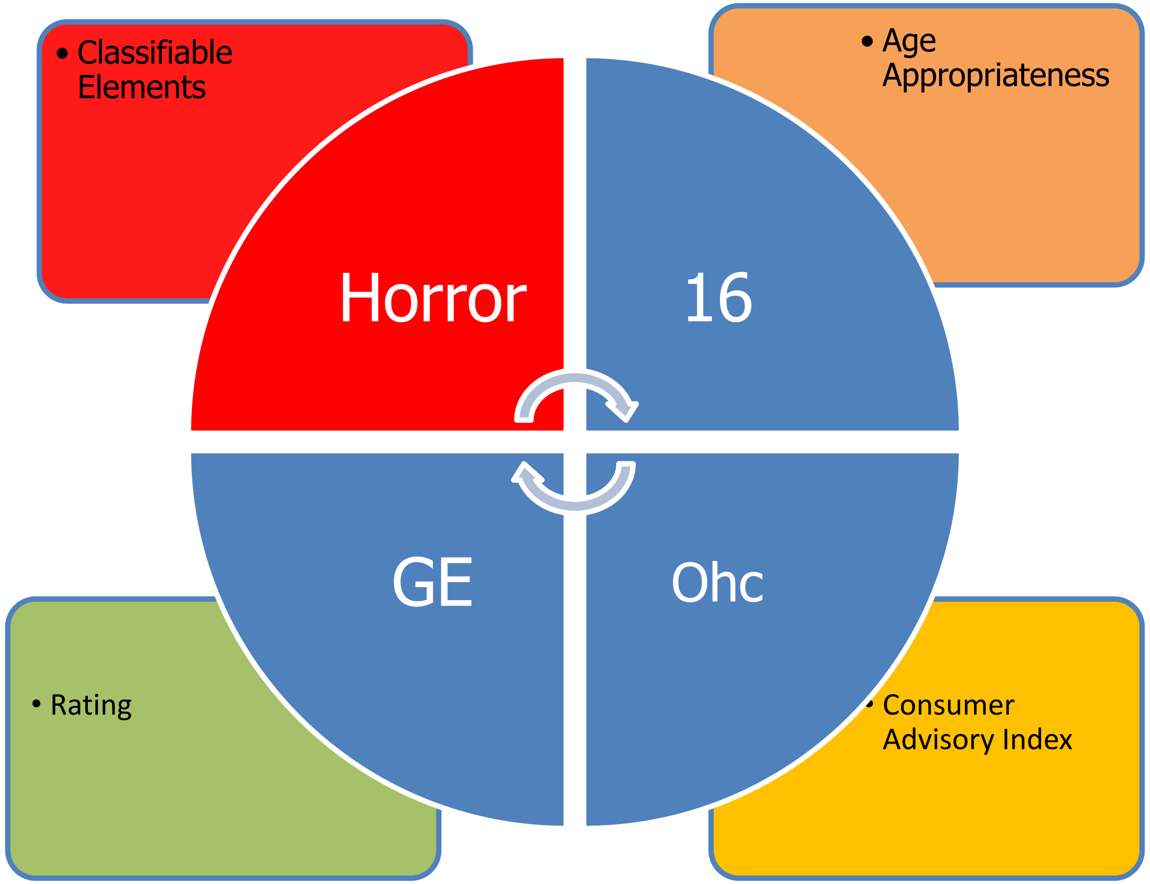
Automating Classification of Audio-Visual Content and Rating for Regulation and Personal Use



DSC-PT-10

Authors:

- Gibson Ngetich
- Cindy Minyade
- Ayaya Vincent
- Maryan Daud
- Edwin Korir
- Margaret Njoroge

July 25th, 2025

PROJECT OVERVIEW

The project builds a machine learning and natural language processing pipeline to automatically classify audiovisual content and predict its rating based on the official rating (GE, PG, 16, 18, Restricted) in Kenya. It aims to support regulatory bodies, parents, educators, and streaming platforms by enhancing the speed, accuracy, and scalability of content classification.

PROBLEM STATEMENT

As digital content explodes across platforms like YouTube, TikTok, and local streaming services, the manual task of classifying each piece of content for age-appropriateness becomes impractical.

This poses a challenge for the regulator to serve clients by setting standards.

OBJECTIVES

- Automate film classification using machine learning to predict appropriate age ratings
- Improve regulatory efficiency by providing AI-based rating suggestions
- Support parental controls through age-based content filtering tools.
- Promote and recommend local content through enhanced categorization and discoverability.

DATA USED

Film classification data

From the country's regulator

From Which Period?

FY 2022/2023 to FY 2024/2025

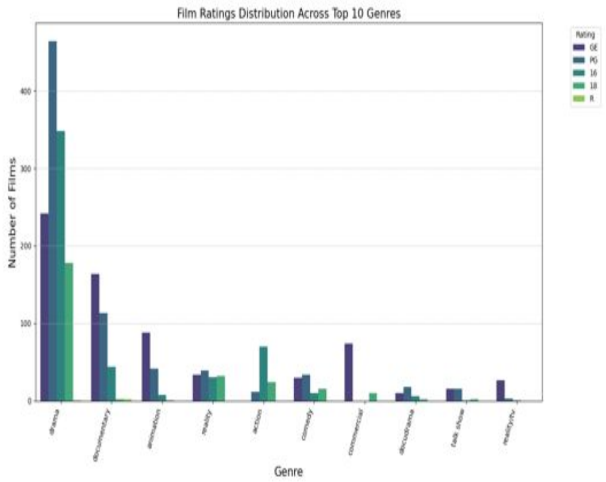
Three(3) years data

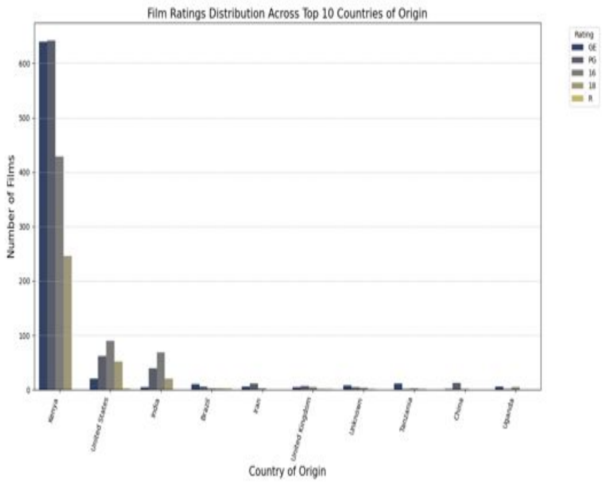
Composition of the Data

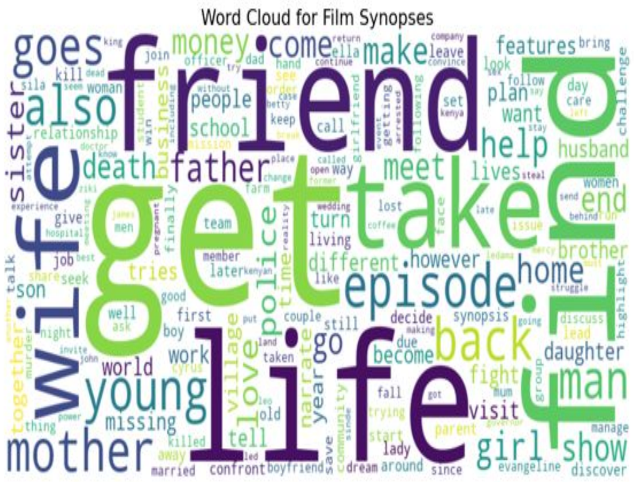
- Film title
- Genre
- Classification
- Synopsis
- Rating, and
- Country of origin

2,574 Film Records → 15 key Attributes → Data was cleaned and standardized

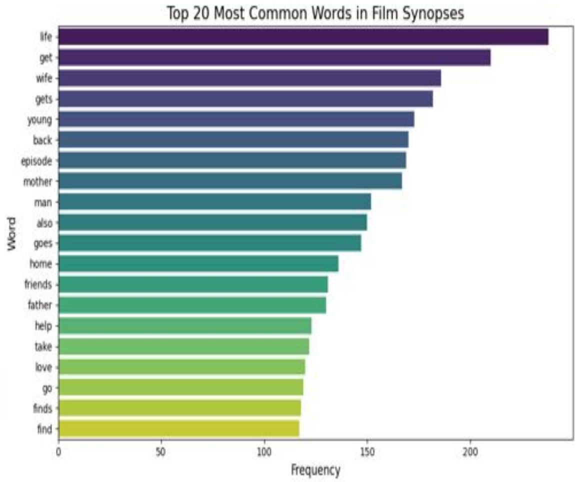
EDA Summary

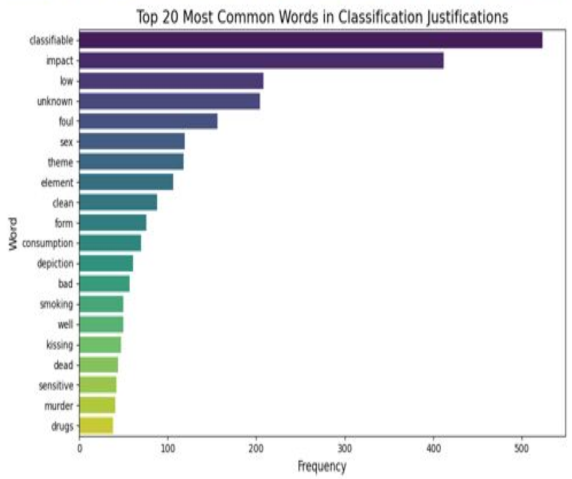

Film Ratings Distribution Across Top 10 Genres
Number of Films
Genre
Rating
GE
PG
16
18
R
drama
documentary
animation
reality
action
comedy
commercial
docudrama
talk show
reality/tv


Film Ratings Distribution Across Top 10 Countries of Origin
Number of Films
Country of Origin
Rating
GE
PG
16
18
R
Kenya
United States
India
Brazil
Iran
United Kingdom
Unknown
Tanzania
China
Uganda


Word Cloud for Film Synopses
friend
get
life
take
find
wife
young
episode
mother
back
man
goes
also
father
death
help
home
girl
show


Word Cloud for Classification Justifications
classifiable
impact
unknown
low
foul
element
theme
sex
clean
consumption
depiction
form
contain
community
sensitive
kissing
smoking
drugs
murder
violent


Top 20 Most Common Words in Film Synopses
Word
Frequency
life
get
wife
gets
young
back
episode
mother
man
also
goes
home
friends
father
help
take
love
go
finds
find


Top 20 Most Common Words in Classification Justifications
Word
Frequency
classifiable
impact
low
unknown
foul
sex
theme
element
clean
form
consumption
depiction
bad
smoking
well
kissing
dead
sensitive
murder
drugs

MODELLING

Logistic Regression Model(Baseline)

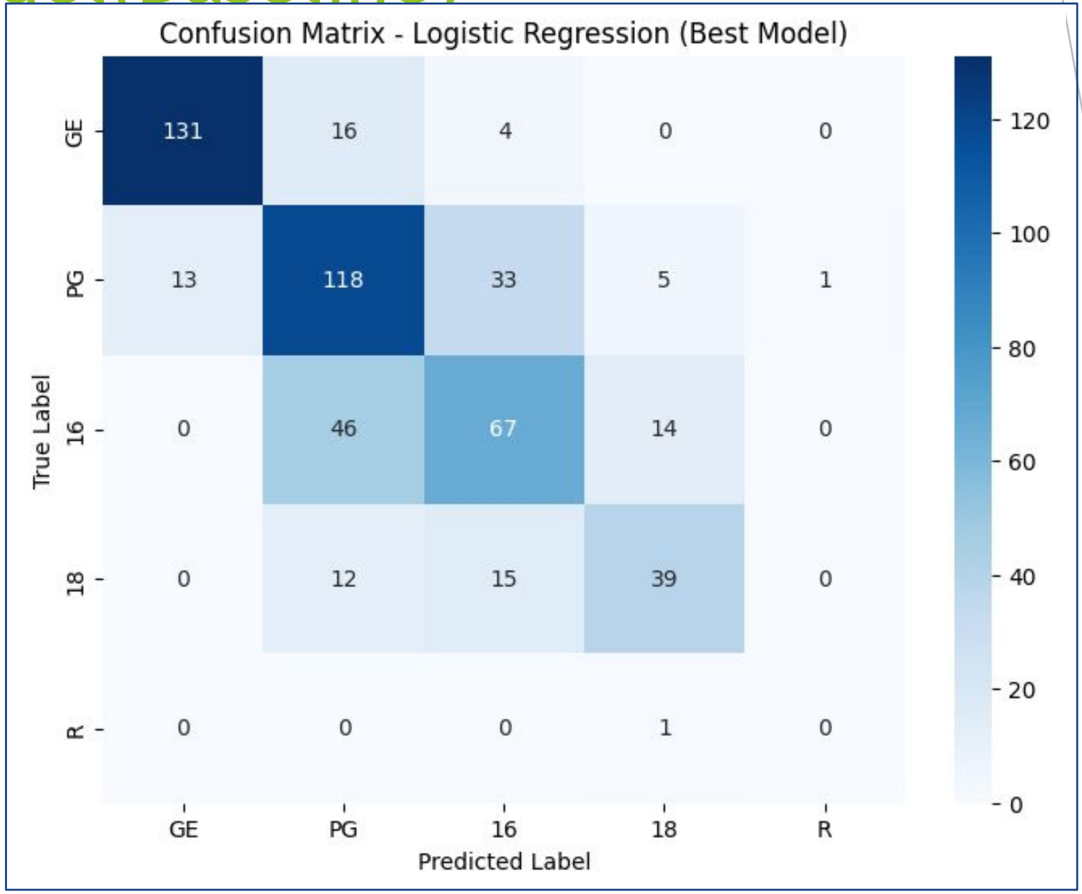
1. Logistic Regression

- **Best Parameters**: `C=1`, `solver='lbfgs’`

- **Accuracy**: 0.69

- **Best F1-Weighted Score (CV)**: 0.69

- **Notes**: Serves as a solid baseline with decent precision for 'GE' and 'PG' classes. Underperforms for rare classes like 'R'.



- The model performs well in predicting GE ratings, with minimal misclassifications. However, PG, 16, and 18 show significant overlap many PG films are misclassified as 16, and vice versa suggesting feature similarity in borderline content. The model struggles to correctly classify Restricted films due to the very limited training data in that class. Improving class balance and incorporating more discriminative features could enhance overall rating accuracy.

MODELLING

Best Model : XGBoost Classifier

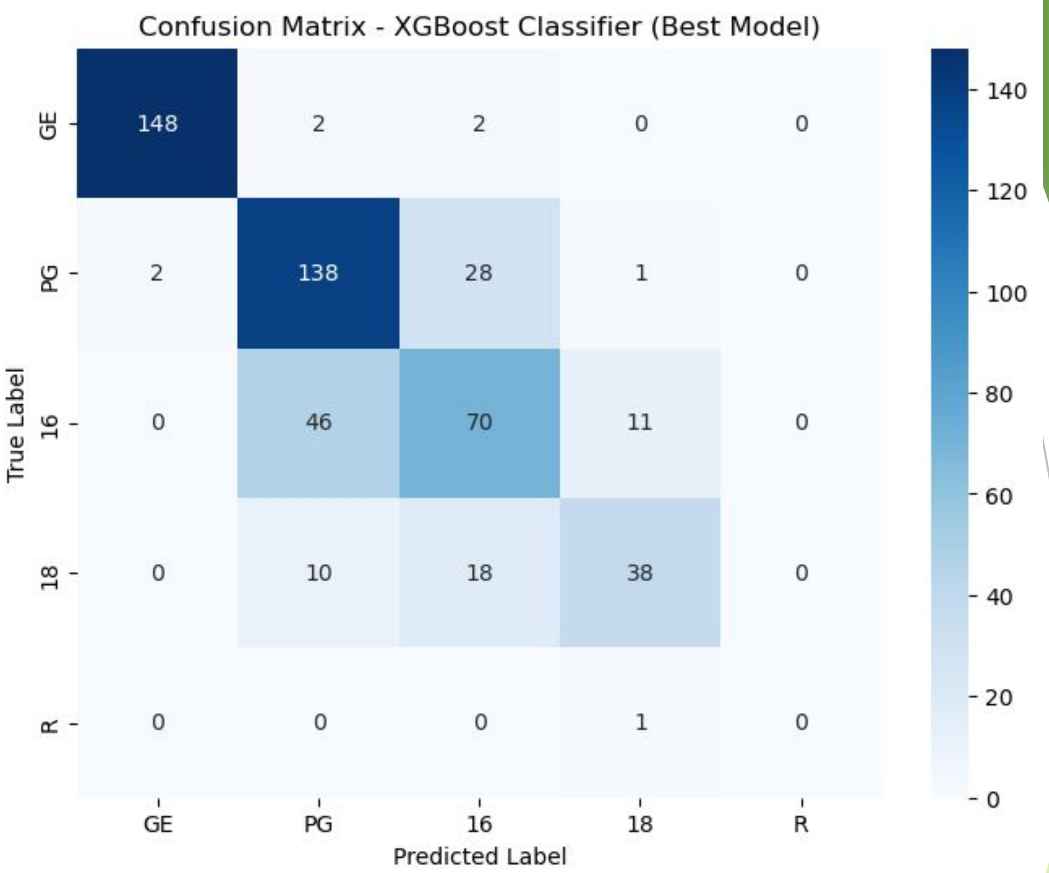
XGBoost Classifier

- Accuracy: **0.77**

- F1-Weighted Score (Cross-Validation): **0.76**

- Excellent balance between performance and interpretability

- Robust against overfitting and handles both categorical and numerical features effectively



High True Positives across major classes like "PG" and "GE", indicating strong classification performance on the most frequent categories.
- **Misclassifications** are more common between similar or adjacent rating classes (e.g., "16" misclassified as "18"), suggesting overlap in content characteristics.
- The **"R" and "18" classes**, which are less represented, show slightly lower recall , typical in imbalanced datasets.

MODEL EVALUATION SUMMARY

Model	Best Parameters	Accuracy	F1-Weighted	Notes
Logistic Regression	C=1, solver=lbfgs	0.69	0.69	Baseline; weak on rare class "R"
Decision Tree	max_depth=None	0.70	0.71	Captures "18"; risk of overfitting
Random Forest	n_estimators=200	0.76	0.76	Balanced; strong overall
XGBoost	lr=0.1, n_estimators=200	0.77	0.76	Top performer; efficient
LightGBM	lr=0.1, n_estimators=200	0.75	0.75	Fast; comparable to XGBoost
Naive Bayes	alpha=1.0	0.75	0.75	Great with text; good for "16"

CONCLUSION

1.

- Built a machine learning model to classify films based on age-appropriateness using the regulators' guidelines

2.

- XGBoost classifier performed best (Accuracy: **0.77**, F1: **0.76**).

3.

- Text features like synopses and justifications were key in improving prediction.

CONCLUSION

4.

- EDA revealed rating patterns across genres, platforms, and countries

5.

- The solution supports regulators, parents, and content platforms in faster, scalable, and objective classification.

6.

- The solution can assist the regulator on time taken to classify content.

Project Challenges

Missing Data: Key columns like VENUE and CONTACT had many null values.

Data Cleaning: Inconsistent formats in fields like DURATION(MINS) required extensive preprocessing

Class Imbalance: Rare ratings like 'R' had very few samples, hurting model recall.

Similar Class Overlap: Models confused PG, 16, and 18 due to feature similarity.

Text Feature Complexity: High-dimensional TF-IDF features from SYNOPSIS increased model complexity.

Evaluation Limitation: Low support for rare classes affected confusion matrix reliability

RECOMMENDATIONS

Use the ML model as a **pre-screening tool** for faster content review

Integrate API with the regulator or streaming platforms for real-time classification.

Switch to **transformer models** (e.g., BERT) for better text analysis.

RECOMMENDATIONS

Apply **SMOTE or class weighting** to handle rating imbalance.

Add **human-in-the-loop** feedback to improve accuracy over time.

Build a **parental control app** to help filter content by rating.

RECOMMENDATIONS

Include **image/audio features** for richer content classification.

Perform **regular audits** to detect and correct model bias.

NEXT STEPS

Advanced NLP: Use BERT/RoBERTa for better text understanding
Multimodal: Add image/audio features
API & Dashboard: Deploy for real-time use
Bias Audit: Fix class imbalance, monitor fairness
Human Feedback: Improve model via user input
Scaling: Localize for other countries/languages

🙏 **Thank You**

Thank you for your time and attention.

We appreciate your support and interest in our project.

— *The Team*