Syracuse University

School of
Information Studies

IST 718

Big Data Analytics

Commercial Detection in News Broadcasting

Team Members:

Karan Shah

Vik Dasari

Maxwell McFadden

CONTENTS

Project Overview

Problem Statement

The media landscape is crowded with content ranging from news broadcasts to commercial advertisements, and differentiating between these types can enhance user experience, compliance monitoring, and content management. This project focuses on the automatic classification of clips from TV news broadcasts into two distinct categories: 'advertisement' and 'non-advertisement'. This classification helps in managing and organizing content more effectively, ensuring compliance with broadcast regulations, and potentially enhancing viewer experience by identifying and segmenting content based on viewer preferences.

Dataset Description

The dataset comprises several thousand clips extracted from various TV news broadcasts. Each clip has been labeled as either 'advertisement' or 'non-advertisement' through a manual review process conducted by media experts. The features include:

- **Auditory Features**: Acoustic signals processed to capture the dynamics and characteristics typical of commercials (e.g., higher volume, distinctive jingles).
- **Visual Features**: Image data analysis results indicating rapid scene changes or the presence of known commercial brands and logos, which are prevalent in advertisement clips.
- **Textual/Bag of Words Features**: Frequencies of specific words or phrases that are commonly used in advertising content but rarely used in regular news segments.

Stakeholders

The primary stakeholders in this project include:

- **Broadcast Networks**: Interested in optimizing ad placements and ensuring compliance with regulations.
- **Advertisers**: Seeking verification that their advertisements are aired as intended and analyzing the impact of their ad placements.
- **Regulatory Bodies**: Monitoring compliance with broadcasting standards and advertisement regulations.
- **Viewers**: Benefiting from enhanced content delivery, where advertisements can be identified, potentially skipped or muted, especially in recorded or on-demand content.
- **Digital Platforms:** Such as streaming services, which could use this technology to enhance user control over content viewing, offering options to skip or highlight advertisements.

By automating the classification of TV clips, this project not only aims to streamline broadcast content management but also enhances the viewing experience by providing insights that could lead to more viewer-centric content delivery. This report will delve into the methodologies employed, the performance of the machine learning models used, and discuss the implications of our findings on the stakeholders involved.

Prediction, inference, and other goals

The aim of this project is to classify TV broadcast clips into 'advertising' and 'non-advertisement' categories using machine learning techniques. Given the variety in the dataset, particularly with a number of semi-normally distributed (and quite robust) audio and visual features, and the high number of categorical bag of words features, we anticipated that certain models would excel:

- **Random Forest** for its robust handling of mixed data types and complex patterns.
- **Logistic Regression** as a baseline for performance due to its efficacy with binary outcomes.
- **Support Vector Machines (SVM)**, appropriate for high-dimensional data, which could leverage the diverse feature set effectively.

We expected the visual and auditory features to hold the highest importance due to their strong indicators in distinguishing commercials, such as scene transitions and background jingles. Given the robustness of these features, a high accuracy in classifying the clips was anticipated. This setup was expected to provide clear insights into the effectiveness of applying machine learning in practical media applications.

Data exploration

Exploring the dataset is a crucial initial step in understanding its characteristics and preparing it for further analysis. In our project, we investigated two distinct datasets from BBC and CNN, each containing a different number of rows and columns.

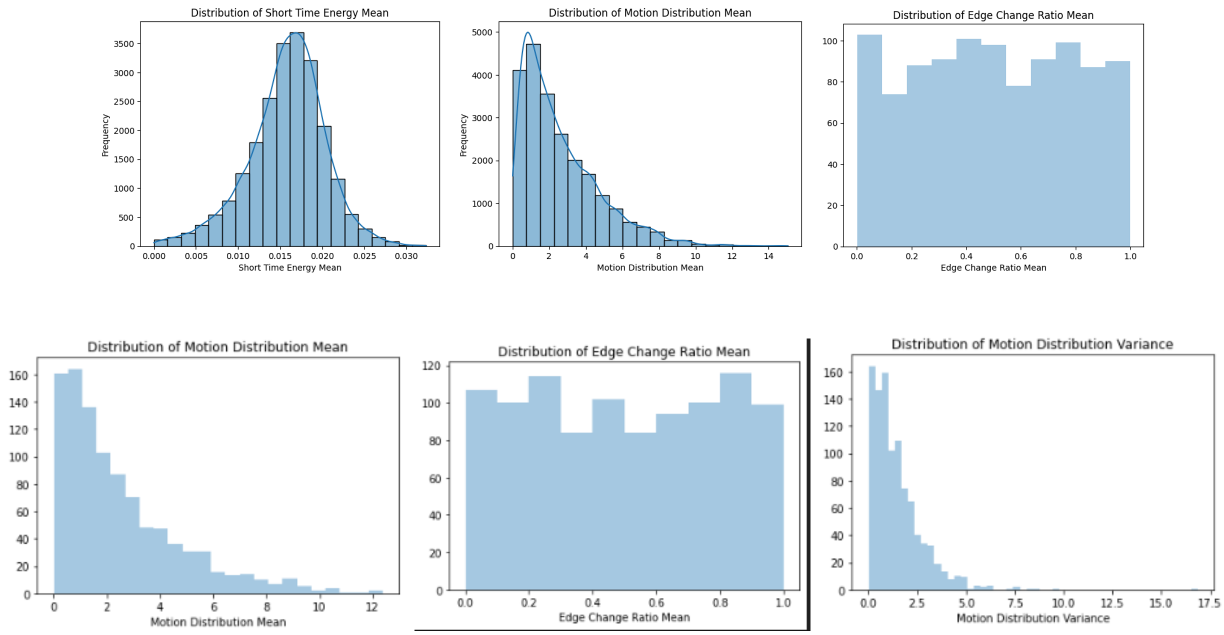
1. BBC Dataset:
    - Number of Rows: 17,702
    - Number of Columns: 232
    - The BBC dataset comprises 17,702 instances, each representing a sample from the BBC news broadcast. With 232 columns, this dataset encompasses a comprehensive set of features, including both audio and visual attributes extracted from the video shots.
2. CNN Dataset:
    - Number of Rows: 22,545

- Number of Columns: 231
- The CNN dataset consists of 22,545 instances, each corresponding to a sample extracted from CNN news broadcasts. Similar to the BBC dataset, this dataset contains 231 columns, representing various audio and visual features extracted from the video shots.

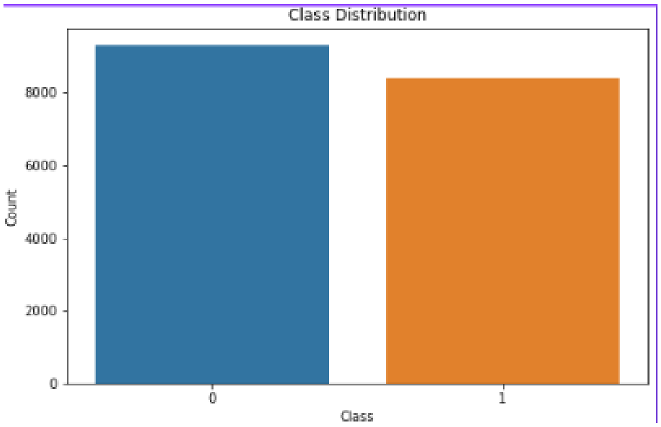
3. **Audio Features**
    - **ZCR**: Measures audio signal sign changes, indicating noise or percussive elements.
    - **Short-Time Energy:** Computes energy in signal frames, useful for intensity changes detection.
    - **Fundamental Frequency:** Represents perceived pitch, crucial for pitch detection and voice recognition.
    - **MFCC Bag of Words**: Represents power spectrum, used in clustering or classification.
    - **Spectral Centroid:** Indicates power spectrum center, higher values denote dominant frequencies.
    - **Spectral Flux**: Measures power spectrum change rate over time, useful for onset detection.

4. Audio Features
    - **Edge Change Ratio:** Estimates the net motion at object boundaries by analyzing changes in edge pixels. It's useful for detecting moving objects or changes in the scene.
    - **Shot Length:** Represents the duration of a shot in frames. It's a fundamental metric in video analysis and editing, often used for scene segmentation and understanding pacing in storytelling.
    - **Text Distribution:** Describes the spatial or temporal distribution of text within the video frames. It's relevant for tasks like subtitle extraction, text-based video indexing, or understanding the prominence of textual elements in videos.

To get a better idea of the shape of the data we were dealing with, we plotted the distributions of a number of the auditory and visual (numeric) features. Some of the examples are shown below:



Figures A-F: Audio Feature Distributions



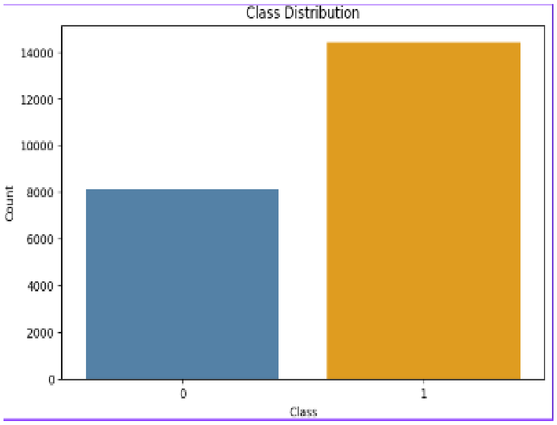


Figure G: Class Distribution for BBC ***Figure H: Class Distribution for CNN***

As displayed in the graphs above, these features are distributed either semi-normally, or semi-uniformly. This trend held with nearly all of the auditory and visual features in the data. This piece of exploration was encouraging, as it implies that the measurements taken were consistent enough that normal distributions were recorded over that large number of instances, which is what one would expect from something like audio measurements in a TV broadcast.

Additionally, this dataset was already cleaned, so we did not have to worry about handling NA values or anything of that nature.

Summary of methods used to solve the problem

Models used:

1. SVM
2. Naive Bayes
3. Logistic regression
4. Random Forest

In this project, we utilized a range of machine learning models within a PySpark environment to classify TV broadcast clips as either 'advertisement' or 'non-advertisement'. The models implemented included Support Vector Machine (SVM), Logistic Regression, Random Forest, and Naive Bayes. Each model was integrated into PySpark pipelines, facilitating systematic transformations and streamlined evaluations.

To assess the performance of each model, we employed two key metrics: the area under the Receiver Operating Characteristic (ROC) curve and the area under the Precision-Recall (PR) curve. These metrics provided comprehensive insights into model accuracy and the trade-offs between sensitivity and specificity, particularly useful in the context of our imbalanced dataset where distinguishing between classes accurately is crucial.

For the SVM model, we applied grid search cross-validation to fine-tune the hyperparameters. This approach allowed us to explore a wide range of parameter settings and identify the optimal combination that maximized prediction accuracy. By using cross-validation, we ensured that the model's performance was robust and generalizable across different subsets of the data, thus mitigating the risk of overfitting.

Results Summary

Algorithm	Dataset	Accuracy	Precision	Recall	F1 Score	AUC ROC	AUC PR
Support Vector Machine	BBC	0.85	0.82	0.91	0.84	N/A	N/A
Support Vector Machine	CNN	0.91	0.89	0.86	0.91	N/A	N/A
Naive Bayes	BBC	0.76	0.76	0.76	0.76	0.54	0.48
Random Forest	BBC	0.86	0.85	0.85	0.84	0.93	0.93
Logistic Regression	BBC	0.86	0.87	0.87	0.86	0.94	0.94

Our top-performing model, which underwent fine-tuning using grid search, achieved an accuracy of **0.91**.

Problems Encountered

We faced a number of obstacles during the project that required thoughtful deliberation and calculated problem-solving. These difficulties covered a wide range of project components, from model performance to data comprehension.

- **Technical Jargon/Interpreting Features:** A major challenge we encountered was the technical jargon that was used in nearly every column. None of us had significant prior knowledge in auditory or visual features, so it proved difficult at times to interpret the meaning of the numeric features in our data.

- **Class Imbalance:** The CNN dataset had a significant class imbalance, which meant that the performance of our classifiers may have been slightly inflated, since the classifiers may have benefited from overly favoring the more commonly occurring classification.
- **Model Generalization over Other Networks:** Ensuring our model's generalization over many networks presented another difficulty for us. While our initial training did prove fairly effective on the datasets from CNN and the BBC, it is important to carefully consider any potential biases and differences in varied broadcasting networks, as the performance of the models would likely be impacted by different broadcasting styles and conventions.

Performance Summary

Our project aimed to classify TV broadcast clips as 'advertisement' or 'non-advertisement' achieved notable success, particularly with the Support Vector Machine (SVM) and Logistic Regression models. Both models were implemented within a PySpark environment and optimized to handle the complexity and volume of our dataset effectively.

The SVM and Logistic Regression models demonstrated exceptional performance, each achieving an Area Under the Receiver Operating Characteristic (ROC) Curve of 0.96 and an Area Under the Precision-Recall (PR) Curve of 0.97. These metrics not only indicate high accuracy and precision but also reflect the models' ability to manage the balance between sensitivity and specificity effectively—critical in our context of binary classification.

These results were facilitated by the strategic use of PySpark pipelines, which streamlined the data processing and model evaluation processes, allowing for robust handling of the data and efficient computation. The use of grid search cross-validation with the SVM model played a pivotal role in fine-tuning parameters to reach optimal performance levels.

In conclusion, the outstanding performance of the SVM and Logistic Regression models underscores the effectiveness of our analytical approach and methodologies in achieving our inference and prediction goals. These results provide a strong foundation for future enhancements and potential applications in broadcast content management and regulatory compliance monitoring.

Citations

(PDF) Commercial Block Detection in Broadcast News Videos, www.researchgate.net/publication/300663545_Commercial_Block_Detection_in_Broadcast_News_Videos. Accessed 30 Apr. 2024.

Apoorv Vyas Dept. of EEE IIT, et al. “Commercial Block Detection in Broadcast News Videos: Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing.” *ACM Other Conferences*, 1 Dec. 2014, dl.acm.org/doi/10.1145/2683483.2683546.

Shah, K., Shinde, A., Vaghasia, S. and Arora, B., 2022. AI IN ENTERTAINMENT - MOVIE RECOMMENDATION. [online] Irjet.net. Available at: [Accessed 28 April 2022]

K. Shah, B. Arora, A. Shinde and S. Vaghasia, "AI in Entertainment – Movie Recommendation using cosine similarity," *2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA*, Pune, India, 2022, pp. 1-4, doi: 10.1109/ICCUBEA54992.2022.10010973.