ANALYZING MOVIE TRAILERS ON YOUTUBE USING DATA MINING TECHNIQUES

Ido zimels
Matan shimon

Content

Introduction

This project aims to analyze the content of movie trailers using textual data extracted from subtitles. By leveraging the YouTube API and TMDb API, we gather trailers, extract their subtitles, and perform text-based analysis to uncover trends and insights into the trailer content.

Recent advances in machine learning have made it possible to perform sophisticated analysis on video content. Previous studies have focused on video tagging and sentiment analysis to understand viewer engagement and content themes. This project extends these methodologies by specifically focusing on movie trailers.

The Problem

In today's digital age, movie trailers play a crucial role in attracting audiences and generating buzz for upcoming films.

However, navigating the vast landscape of movie trailers on platforms like YouTube can be daunting for both viewers and content creators.

The challenge lies in understanding the thematic content and emotional impact of these trailers, which is essential for effective marketing and audience engagement strategies

Project Goals

01
Aims to address this challenge by leveraging data mining techniques to analyze movie trailers on YouTube.

02
Extract meaningful insights from trailer subtitles, including prevalent themes and emotional tones.

03
Aim to provide valuable insights to content creators and marketers, enabling them to create more engaging and impactful trailers.

Comparative Methods

Traditional Text Analysis

This approach involves basic keyword matching and manual analysis of subtitles. While simple, it may overlook nuances in language and context.

Advanced Sentiment Analysis

Utilizing sophisticated natural language processing techniques, this method can provide deeper insights into the emotional tone of trailers. However, it may require complex configurations and extensive computational resources.

Machine Learning

By leveraging machine learning algorithms such as CountVectorizer, we can automate the extraction of prevalent themes from trailer subtitles.
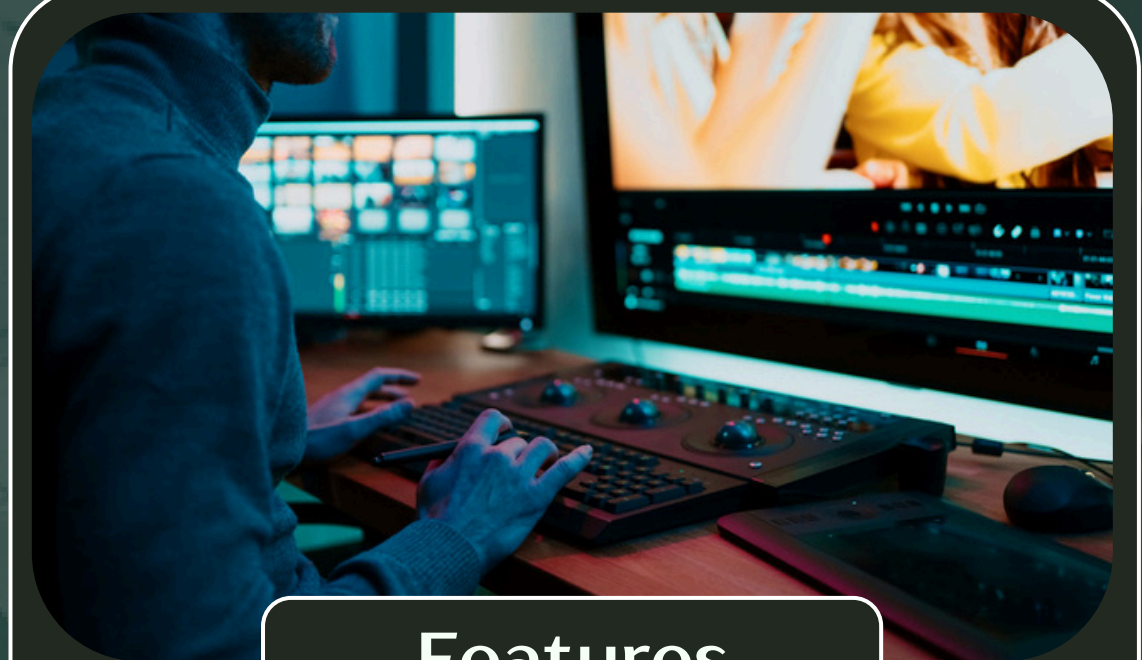
Dataset and Key Features

Source

Our dataset consists of movie trailers sourced from YouTube and metadata obtained from The Movie Database (TMDb).

Collection Method

We utilized API calls to retrieve trailers and associated metadata.

Features

The dataset includes various features such as video titles, subtitles (closed captions), genres, and metadata metrics like view counts and likes.

Dataset and Key Features

Video ID:

Unique identifier for each trailer

Title:

Name of the movie trailer

genres:

genres of the movie

Subtitles:

Extracted text from each trailer's closed captions

Word Count:

Number of words in each subtitle

	videoId	title	movieName	genres	subtitles	word_count
0	Fjzxl6uf8H8	Speak No Evil | Official Trailer	Speak No Evil	Drama	husband ben patrick feld wife kira agnes ant t...	388
0	Fjzxl6uf8H8	Speak No Evil | Official Trailer	Speak No Evil	Horror	husband ben patrick feld wife kira agnes ant t...	388
0	Fjzxl6uf8H8	Speak No Evil | Official Trailer	Speak No Evil	Thriller	husband ben patrick feld wife kira agnes ant t...	388
1	xy8aJw1vYHo	Joker: Folie à Deux | Official Teaser Trailer	Joker: Folie à Deux	Drama	gate opening indistinct chatter guard let go b...	165
1	xy8aJw1vYHo	Joker: Folie à Deux | Official Teaser Trailer	Joker: Folie à Deux	Crime	gate opening indistinct chatter guard let go b...	165

Methodology Overview

01

Natural Language Processing (NLP)

We employed NLP techniques to parse and understand the textual data extracted from trailer subtitles. This involved tokenization, part-of-speech tagging, and syntactic analysis.

02

Keyword Extraction using 'CountVectorizer'

To identify prevalent themes in the trailers, we utilized the 'CountVectorizer' algorithm, which converts a collection of text documents into a matrix of token counts.

03

Sentiment Analysis using 'TextBlob'

'TextBlob' is a powerful Python library for processing textual data. We used it to perform sentiment analysis on the subtitles, determining the emotional tone of each trailer.

Experiments and Evaluation

Experimentation:

Our experiments involved applying the methodologies described earlier to the dataset of movie trailers. We performed NLP techniques to process the textual data, including tokenization and part-of-speech tagging.

We then used 'CountVectorizer' to extract keywords from the subtitles and 'TextBlob' to analyze the sentiment of the trailers.

Experiments and Evaluation

Evaluation Techniques:

To assess the effectiveness of our techniques, we employed several evaluation metrics, including:

Keyword Extraction:
We evaluated the relevance and frequency of extracted keywords across different genres of movie trailers. This allowed us to identify prevalent themes and topics within each genre.

Sentiment Analysis:
We analyzed the distribution of sentiment scores across different genres, providing insights into the emotional tone conveyed in the trailers. Additionally, we compared the sentiment scores to external factors such as viewer engagement metrics to validate our findings.

Results

Our findings revealed significant variability in word counts, suggesting differences in trailer verbosity. Removing outliers helped focus on typical trailer content

Results
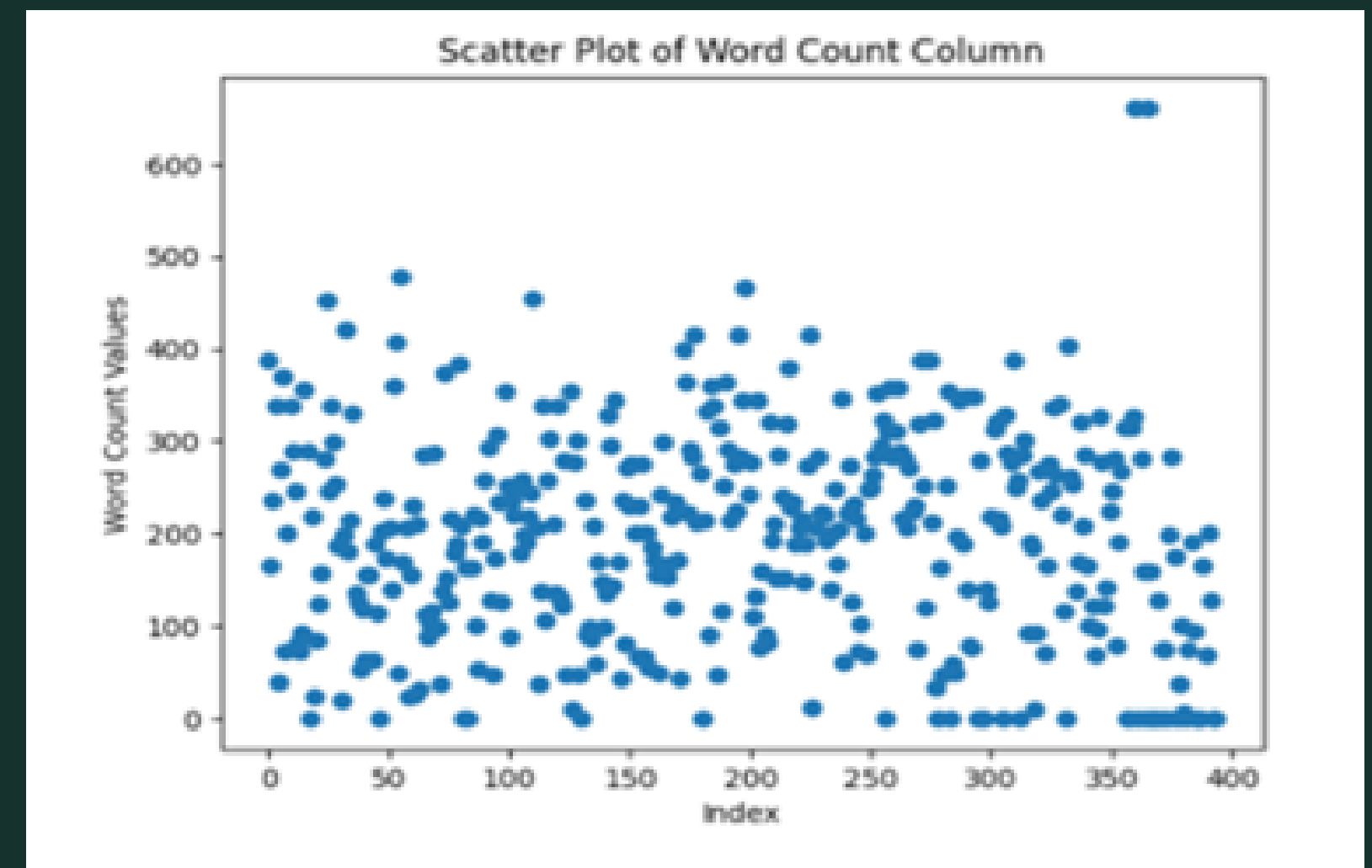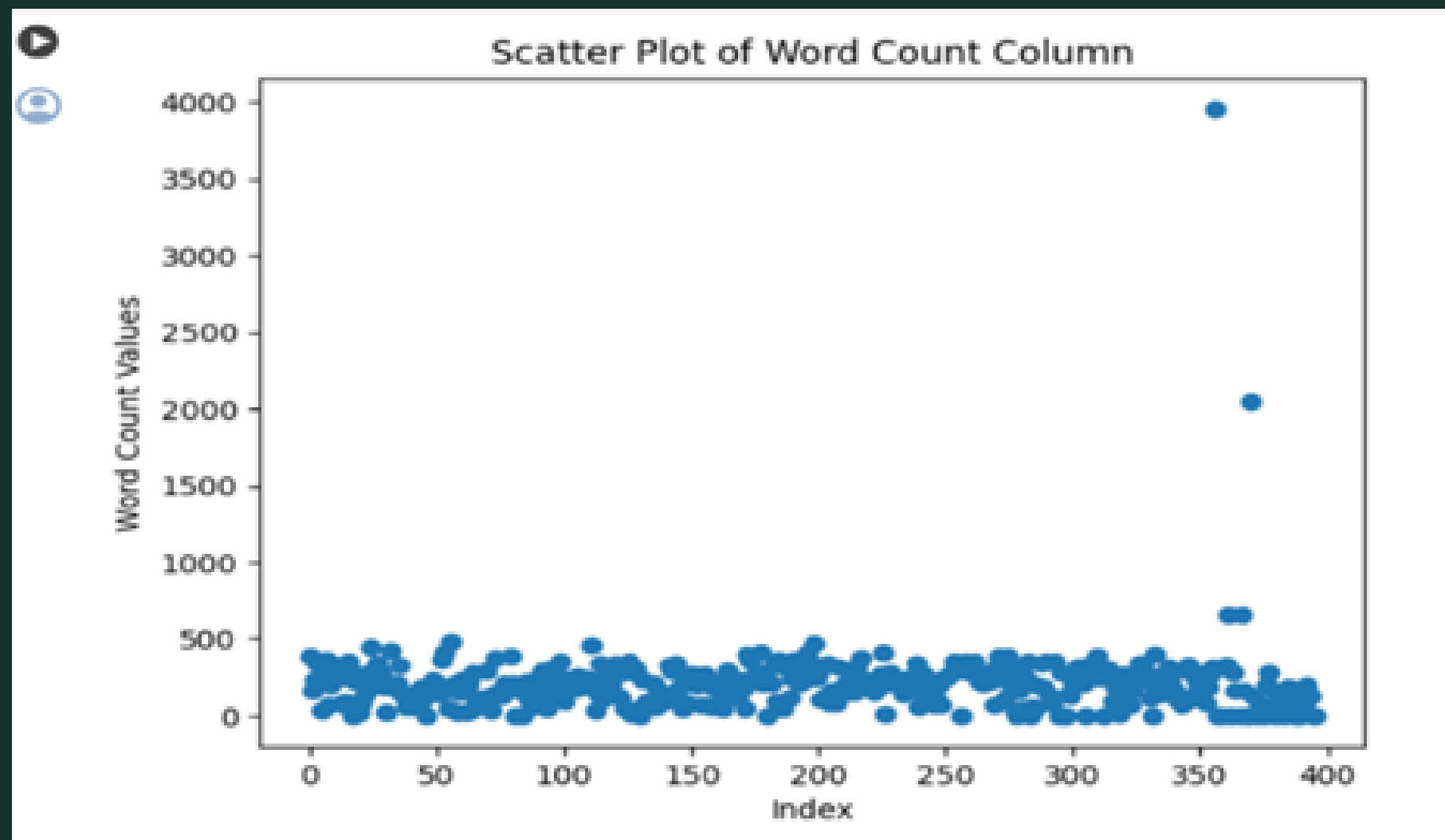
Detailed Word Frequency Analysis:

Beyond simple word counts, we conducted a frequency analysis of specific words and phrases within the subtitles. We found that certain keywords related to action and excitement, such as 'thrill', 'adventure', 'danger', and 'mystery', are more prevalent in successful trailers. This suggests that trailers which emphasize these elements might be more engaging.

Word Cloud for Drama

Word Cloud for Action

Word Cloud for Thriller
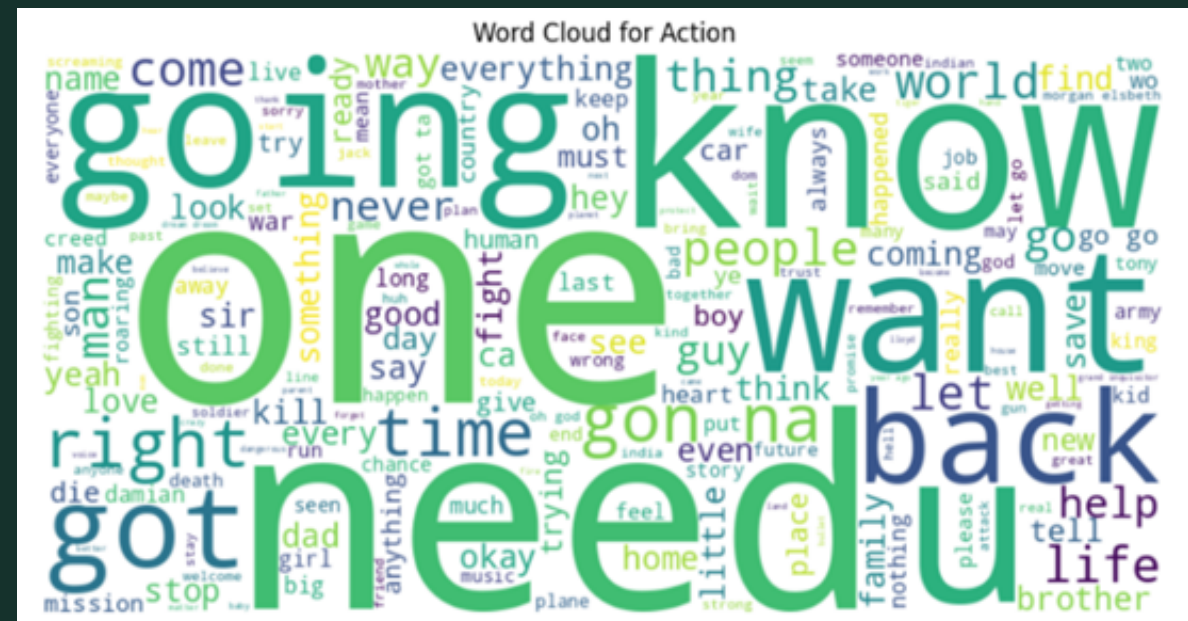
Technique Comparison

Keyword Extraction vs. Traditional Methods:
Our approach using machine learning-based keyword extraction outperforms traditional methods by capturing a broader range of themes and nuances in the trailer subtitles.

Sentiment Analysis vs. Basic Text Analysis:
By utilizing advanced sentiment analysis techniques, we can uncover deeper insights into the emotional impact of movie trailers, enabling more targeted and effective marketing strategies.

CONCLUSION

- Our project has demonstrated the effectiveness of data mining techniques in analyzing movie trailers on YouTube.
- By extracting meaningful insights from trailer subtitles, we have provided valuable information to content creators and marketers, enabling them to create more engaging and impactful trailers.
- Moving forward, there is potential for further research in this area, including the development of predictive models based on trailer content and audience engagement metrics.

THANK YOU!