Bollywood Movie Analysis Project

Project Overview

The objective of this project is to curate and analyze a
in order to study how themes such as Hindu–Muslim relations, gender relations, and nationalism are depicted over time, using reproducible and transparent data science best practices.

Pipeline Breakdown

1. Dataset Download & Sampling

- : The Indian Movie Database from Kaggle (https://www.kaggle.com/datasets/pncnmnp/the-indian-movie-database).
- : Selected a random sample of 100 movies released after 2010, preserving all available attributes for each.
  - : Sampling is handled by a script located in the `scripts/` folder, which allows others to reproduce or modify the sampling process using the same random seed or parameters.
  - : The sampled movies are saved as a CSV within the `data/` directory.

2. Movie Materials Collection

For each film in the sample, the following are collected and organized:

- (`data/subtitles/`): Obtained from OpenSubtitles via API.
- (`data/descriptions/`): Fetched from OMDb using API queries to ensure rich summaries.
- (`data/posters/`): Downloaded via TMDb API.

All files maintain a consistent directory structure for ease of navigation and reproducibility.

3. Descriptive Metadata & Thematic Coding

- : Both human and LLM (Large Language Model)-based analyses were conducted on subtitle and description text to identify and categorize:
  - *Hindu–Muslim relations*
  - *Gender relations*
  - *Nationalism*

- : LLMs (e.g., Google Gemini API) are used to not only detect the presence of these themes, but also their sentiment and nuance:
  - Assigned classifications: *Exclusionary vs. Secular (Inclusive)*, *Positive or Negative*, *Progressive or Conservative*.
-

4. Analytical and Visualization Outputs

- : Thematic codes and sentiment labels are visualized in R using ggplot2.
  - Main plot: Annual trends (from 2010 onward) in the frequency and sentiment of each coded theme.
  - Additional plots illustrate further nuances and comparisons, all available in the `data/` directory.
- : All R scripts required to recreate these analyses from the CSVs are provided.

Repository Structure

```
bollywood-movie-analysis/
│
├── data/
│   ├── descriptions
│   │   └── descriptions_all.csv
│   ├── metadata/
│   │   └── metadata_extended.csv
│   ├── posters/
│   │   └── posters_all.csv
│   ├── sampled/
│   │   └── movies_sampled.csv
│   ├── subtitles/
│   │   └── subtitles_all.csv
│   ├── bollywood_2010-2019.csv
│   ├── bollywood_meta_2010-2019.csv
│   ├── bollywood_ratings_2010-2019.csv
│   ├── bollywood_text_2010-2019.csv
│   ├── bollywood_coding_clean.csv
│   ├── bollywood_coding.csv
│   ├── theme_frequency.png
│   ├── theme_sentiment_breakdown.png
│   ├── theme_sentiment_overtime.png
│   └── violence_measure.csv
│
```

```
├── scripts/
│   ├── description.py
│   ├── metadata.py
│   ├── poster.py
│   ├── proposed_sentiment.py
│   ├── sampled_movies.py
│   ├── subtitles.py
│   ├── thematic_clean.py
│   ├── thematic_coding.py
│   └── visualisation.R
│
├── supplementary_ideas.md
├── requirements.txt # Python dependencies
├── requiremnet_r.txt # R dependencies
├── .gitignore
├── .gitattributes
├── .Rprofile
├── LICENSE
├── renv.lock
├── renv/ # R virtual environment
├── venv/ # Python virtual environment
├── README.md
```

🚀 How to Run This Project Locally

1. Clone the Repository

```
git clone https://github.com/Econbee007/bollywood-movie-analysis.git
cd bollywood-movie-analysis
```

2. Setup Python Environment

```
python -m venv venv
source venv/Scripts/activate  # On Windows
# Or on Mac/Linux: source venv/bin/activate
pip install -r requirements.txt
```

3. Download Full Dataset from Kaggle

Visit: Indian Movie Database on Kaggle

Place the files inside the `data/` folder.

4. Sample 100 Movies

```
python scripts/sampled_movies.py
```

Output: `data/sampled/movies_sampled.csv`

5. Collect Subtitles, Descriptions, and Posters

```
python src/subtitles.py
python src/description.py
python src/poster.py
```

Output:

- data/subtitles/subtitles_all.csv
- data/descriptions/descriptions_all.csv
- data/posters/posters_all.csv

These scripts use the OpenSubtitles API, OMDb API, and TMDb API respectivel

Before running them, ensure that your API keys are saved in a `.env` file in

6. Perform Thematic Coding using LLM

```
python src/thematic_coding.py
```

Output: `data/thematic_coding.csv`

Now to clean the data:

```
python src/thematic_coding_clean.py
```

Output: `data/thematic_coding_clean.csv`

This is the final output for thematic coding.

Themes:

- Hindu–Muslim relations
- Gender relations
- Nationalism

Sentiment Attributes:

- Exclusionary ↔ Inclusive
- Positive ↔ Negative
- Progressive ↔ Conservative

Additional Thematic Measure: Violence Representation

We introduce a new variable: violence_representation , which classifies mov

This provides a scalable proxy for thematic violence, useful for analyzing

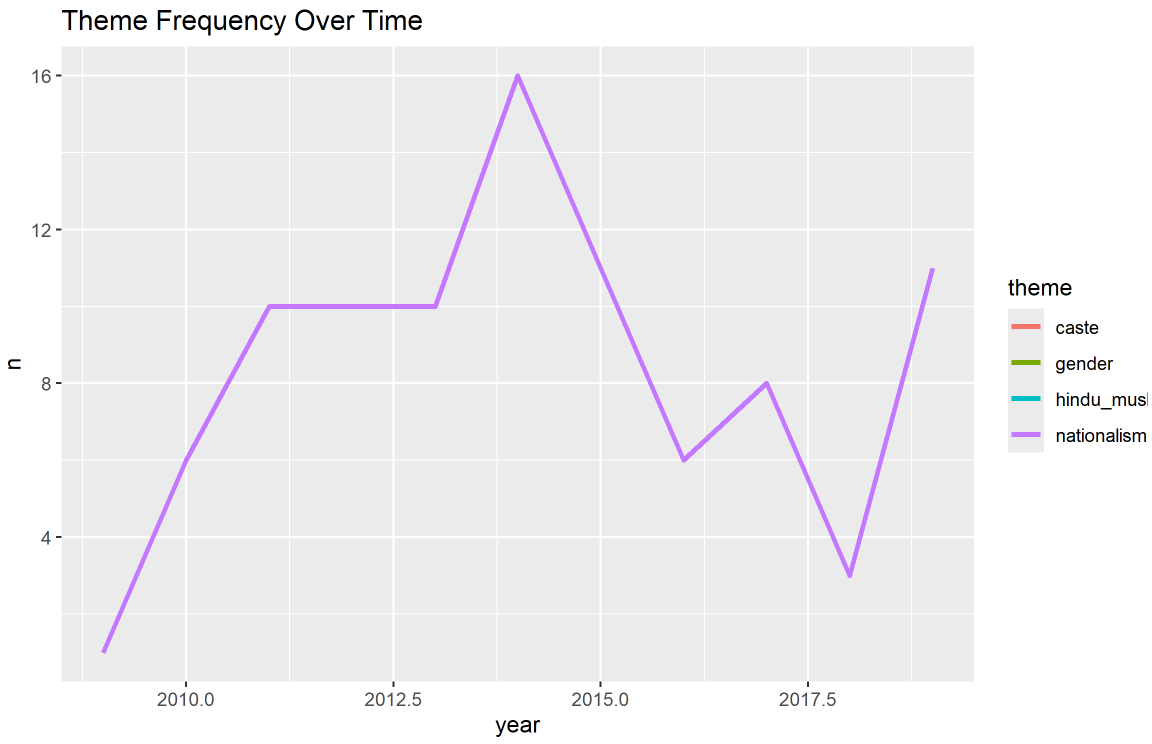
```
python src/proposed_sentiment.py
```

Output: data/violence_measure

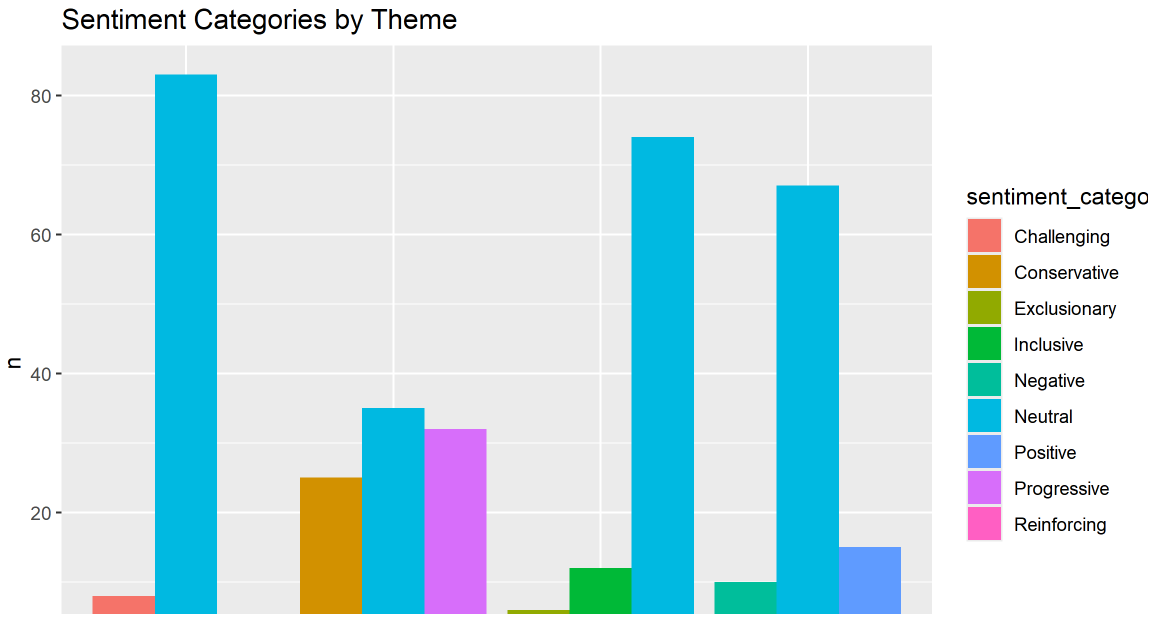
7. Visualize Trends using R

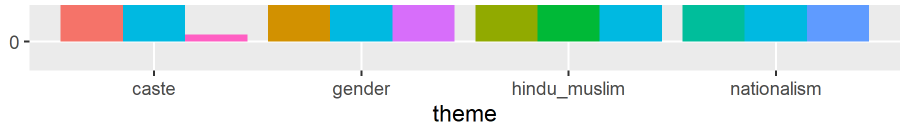
Run visualisation.R after installing all R dependencies from requiremnet_r

Theme Frequency Over Time

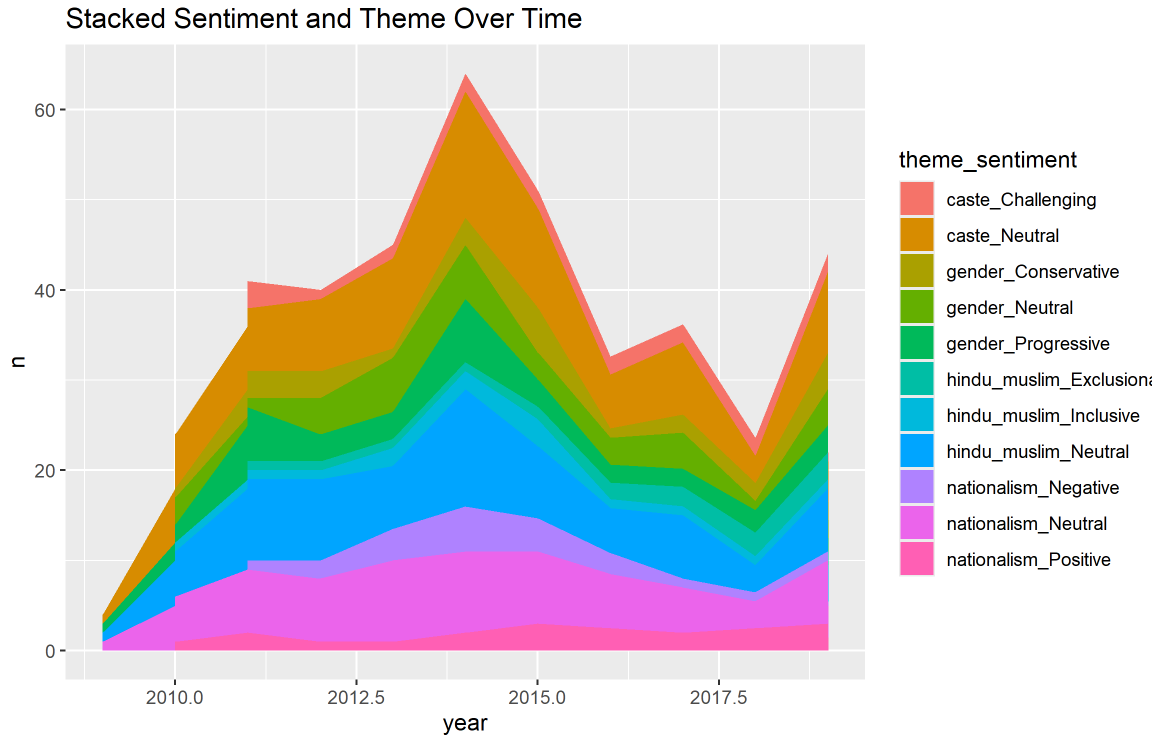


Theme Sentiment Breakdown





Theme Sentiment Over Time



🔍 Key Design Principles

- **Transparency & Reproducibility**: Every step, from data acquisition to them
- **AI Integration**: LLMs are used for scalable and auditable theme classifica
- **Extensibility**: Modular file and folder structure allows addition of new t

---

Supplementary Ideas

To make this dataset more insightful and meaningful, here are a few additio

---

1. User Reviews and Audience Sentiment

Sometimes, what a movie tries to say and how people feel about it can be ve

- **Where to get it**: IMDb reviews (via scraping or using IMDbPY), or even pla
- **How to use it**: We could extract text reviews, run them through basic sent

---

2. Cast Diversity (Gender, Religion, Nationality)

Representation matters. Looking at who stars in these movies – across gende

- **Where to get it**: IMDb (using IMDbPY) for cast lists, and Wikipedia/Wikida
- **How to use it**: For each film, we could check the top-billed cast and recc

---

3. Awards and Critical Recognition

Awards and nominations signal what kinds of films the industry values. Incl

- **Where to get it**: Wikipedia pages for each film often list awards, or we c
- **How to use it**: Add columns like `won_award`, `num_awards`, or even break it

---

Each of these additions would make the dataset richer and allow for more nu

---

License

This project is released under the **MIT License**.

This repository is intended as a transparent and extensible resource for re