AML 2304 – Natural Language Processing

Movie Sentiment Analysis

(Mid Submission)

November 14, 2023

Submitted to :
Prof. Bhavik Gandhi

Submitted by Group 3 :
Abhishek Natani
Auradee Castro
Bhumika Rajendra Babu
Miraj Sinya
Olivia Deguit
Rochan Mehta
Roger Mais
Varun Sharma

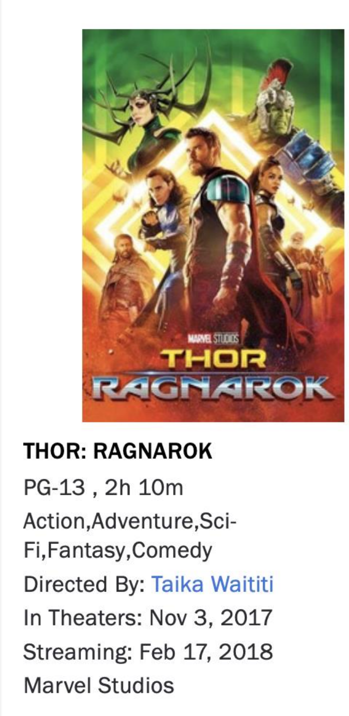
Lambton College

Case Background

- Movie Reviews have always been a reference point for the audience to decide weather or not to watch movies but from a production standpoint it has not be utilized to its full extent.
- This application will allow the producers in the mass media and entertainment industry to not only understand the sentiment of the audience but also the reasons behind those sentiments.
    - Behind the scenes, the application uses Natural Language Processing models along with clustering techniques to make sense of the sentiment and get the intent behind those sentiments giving deeper insights into audience opinions, preferences, and pain points.
    - Help potential clients to potentially get insights on trends in the audience's sentiment towards specific genre over the years, actors, directors which will aid them in making informed decisions.
- Dataset containing movie reviews, details of the movie (genre, budget, cast and crew, ratings) is taken from Kaggle along with the data accessible to the public on IMDb, which is used for cross-validating the data collected from Kaggle.
    - Dataset containing 500K reviews for movies

Case Study (1/4) : Marvel Studios

Thor: Love and Thunder released in 2022 rated significantly lower than its prequal Thor: Ragnarok released in 2017



THOR: RAGNAROK

PG-13 , 2h 10m

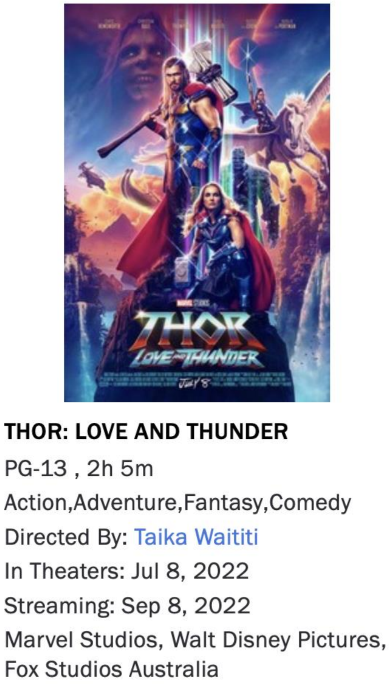
Action,Adventure,Sci-Fi,Fantasy,Comedy

Directed By: Taika Waititi

In Theaters: Nov 3, 2017

Streaming: Feb 17, 2018

Marvel Studios



THOR: LOVE AND THUNDER

PG-13 , 2h 5m

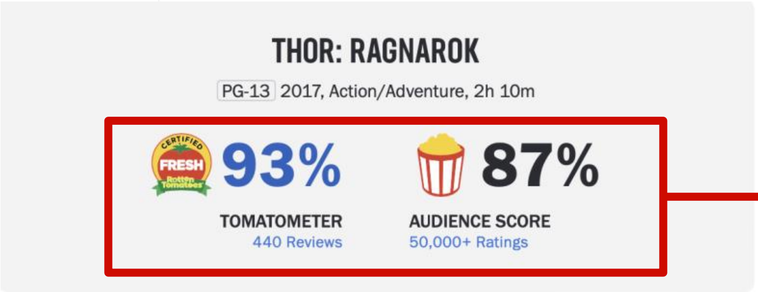
Action,Adventure,Fantasy,Comedy

Directed By: Taika Waititi

In Theaters: Jul 8, 2022

Streaming: Sep 8, 2022

Marvel Studios, Walt Disney Pictures, Fox Studios Australia



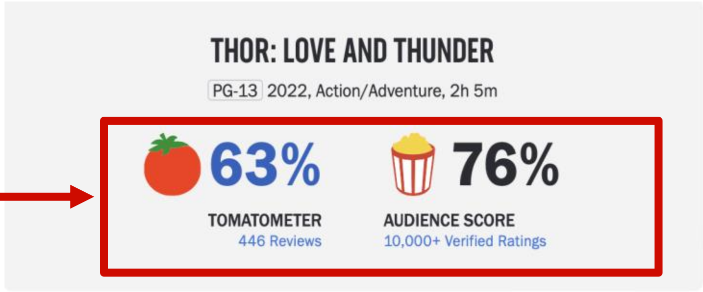
THOR: RAGNAROK

PG-13 2017, Action/Adventure, 2h 10m

93% TOMATOMETER 440 Reviews

87% AUDIENCE SCORE 50,000+ Ratings

→



THOR: LOVE AND THUNDER

PG-13 2022, Action/Adventure, 2h 5m

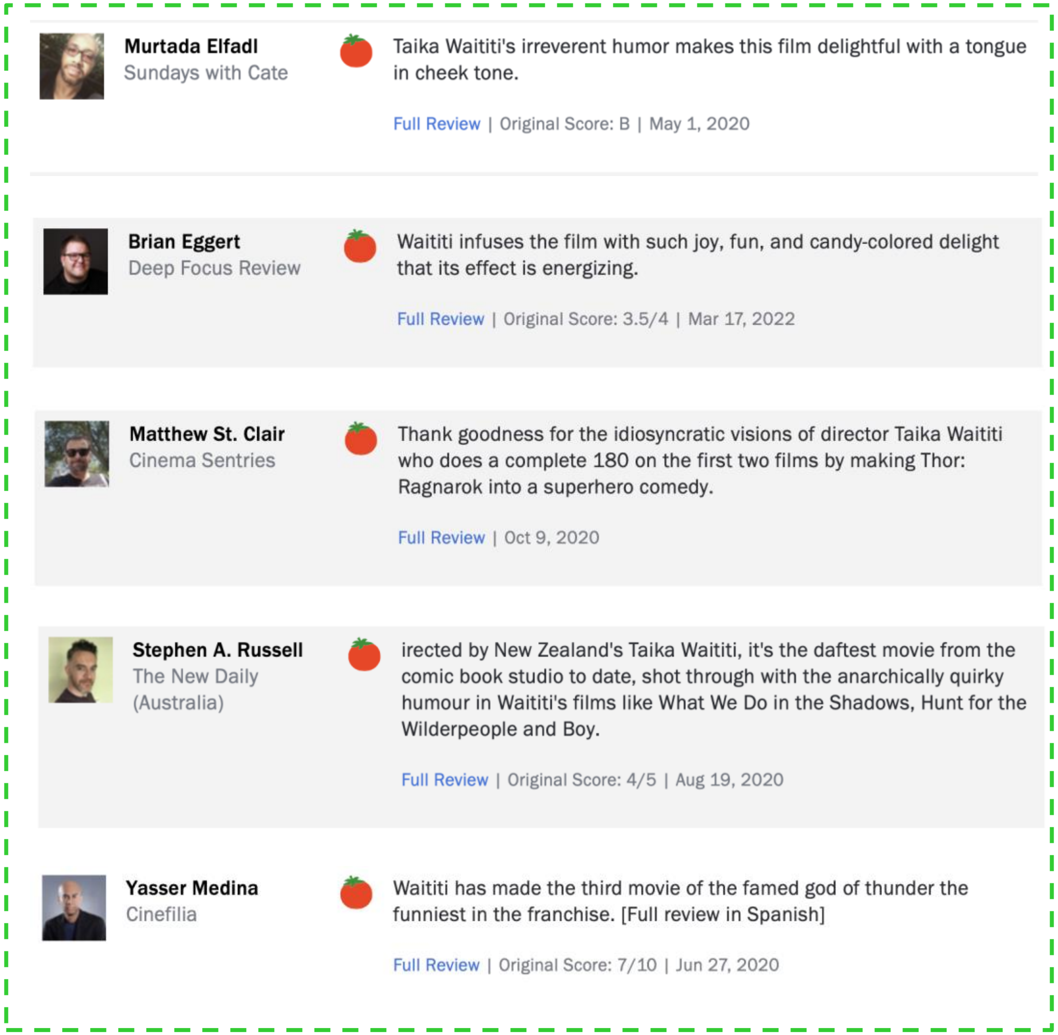
63% TOMATOMETER 446 Reviews

76% AUDIENCE SCORE 10,000+ Verified Ratings

Case Study (2/4)

Factors contributing to the movie being well received by the audience

Thor: Ragnarok Reviews



Murtada Elfadl
Sundays with Cate

Taika Waititi's irreverent humor makes this film delightful with a tongue in cheek tone.

Full Review | Original Score: B | May 1, 2020

Brian Eggert
Deep Focus Review

Waititi infuses the film with such joy, fun, and candy-colored delight that its effect is energizing.

Full Review | Original Score: 3.5/4 | Mar 17, 2022

Matthew St. Clair
Cinema Sentries

Thank goodness for the idiosyncratic visions of director Taika Waititi who does a complete 180 on the first two films by making Thor: Ragnarok into a superhero comedy.

Full Review | Oct 9, 2020

Stephen A. Russell
The New Daily (Australia)

irected by New Zealand's Taika Waititi, it's the daftest movie from the comic book studio to date, shot through with the anarchically quirky humour in Waititi's films like What We Do in the Shadows, Hunt for the Wilderpeople and Boy.

Full Review | Original Score: 4/5 | Aug 19, 2020

Yasser Medina
Cinefilia

Waititi has made the third movie of the famed god of thunder the funniest in the franchise. [Full review in Spanish]

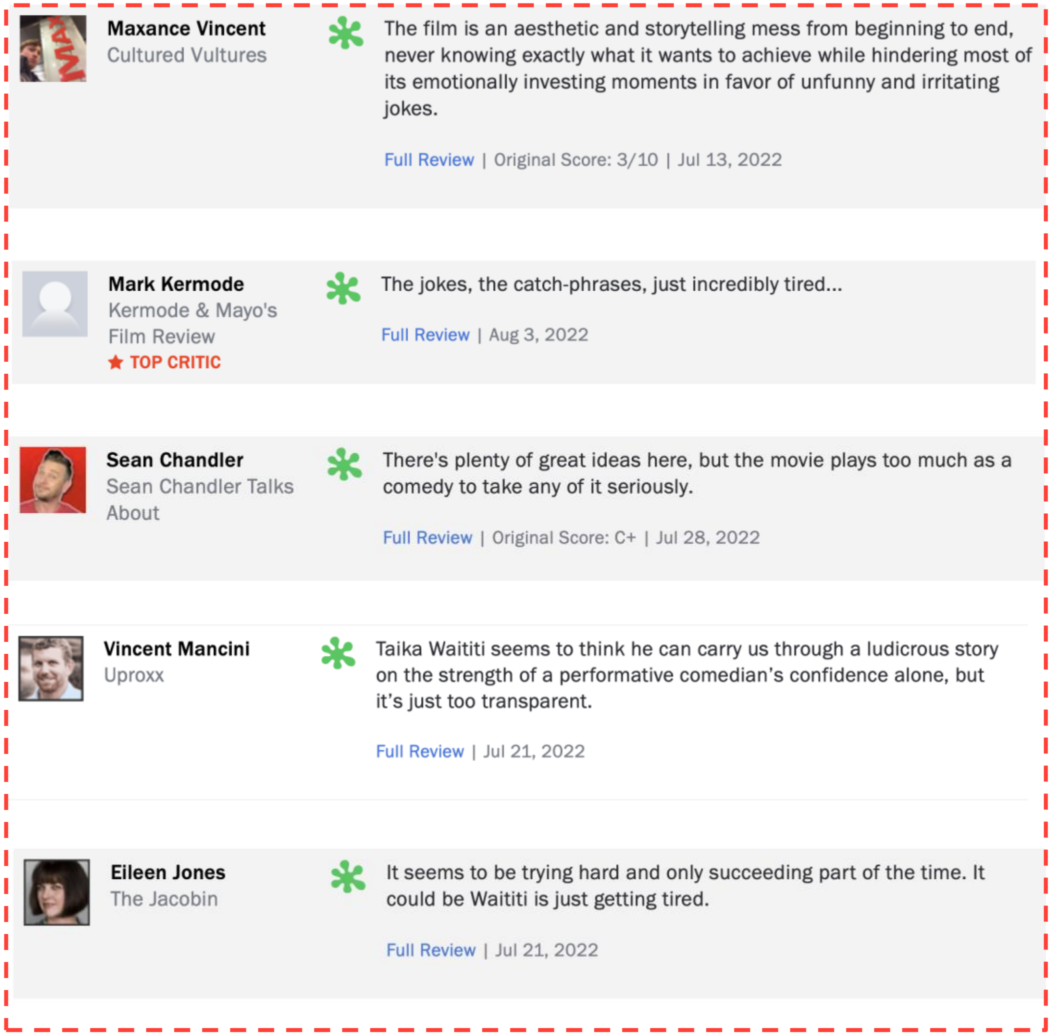
Full Review | Original Score: 7/10 | Jun 27, 2020

Significant mentions of director Taika and directing style infusing comedy into the franchise in the positive reviews of the first movie

Case Study (3/4)

Factors contributing to the movie being criticized by the audience

Thor : Love and Thunder Reviews



Maxance Vincent
Cultured Vultures

The film is an aesthetic and storytelling mess from beginning to end, never knowing exactly what it wants to achieve while hindering most of its emotionally investing moments in favor of unfunny and irritating jokes.

Full Review | Original Score: 3/10 | Jul 13, 2022

Mark Kermode
Kermode & Mayo's Film Review
★ TOP CRITIC

The jokes, the catch-phrases, just incredibly tired...

Full Review | Aug 3, 2022

Sean Chandler
Sean Chandler Talks About

There's plenty of great ideas here, but the movie plays too much as a comedy to take any of it seriously.

Full Review | Original Score: C+ | Jul 28, 2022

Vincent Mancini
Uproxx

Taika Waititi seems to think he can carry us through a ludicrous story on the strength of a performative comedian's confidence alone, but it's just too transparent.

Full Review | Jul 21, 2022

Eileen Jones
The Jacobin

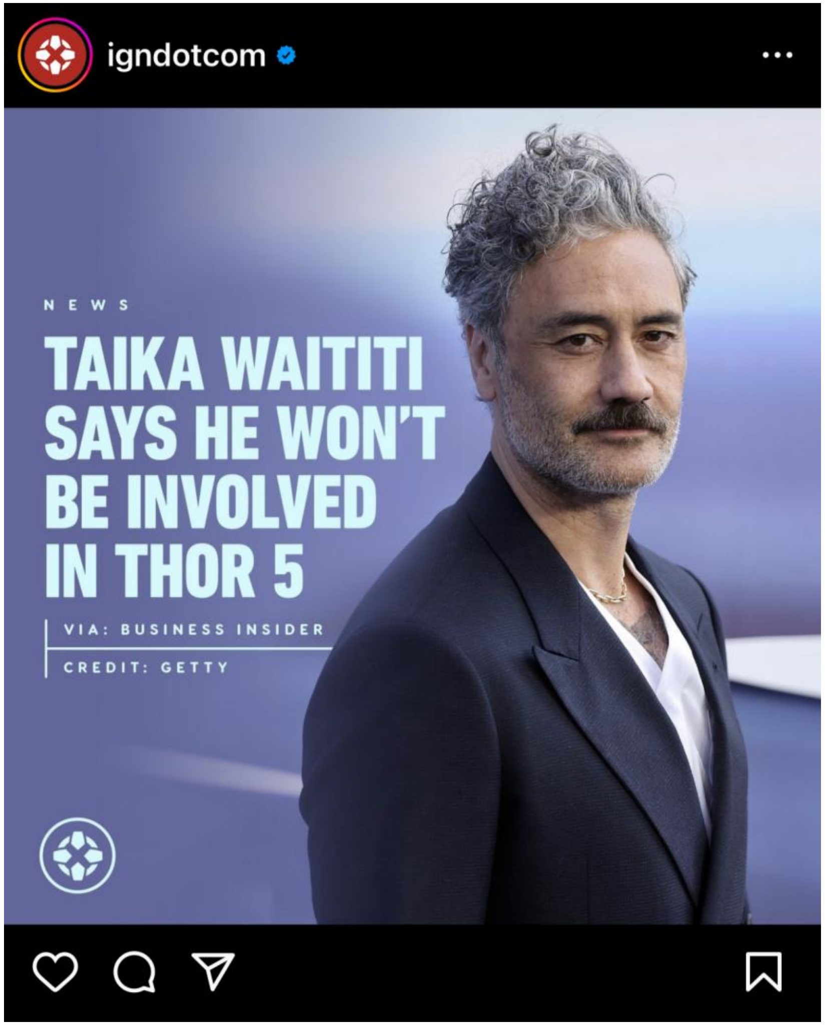
It seems to be trying hard and only succeeding part of the time. It could be Waititi is just getting tired.

Full Review | Jul 21, 2022

Significant mentions of director Taika, movie humor and story in the negative reviews of the first movie

Case Study (4/4)

Failure of sequel along with negative reviews of movies potentially led to Director Taika Waititi's subsequent departure from the Thor franchise



TAIKA WAITITI SAYS HE WON'T BE INVOLVED IN THOR 5

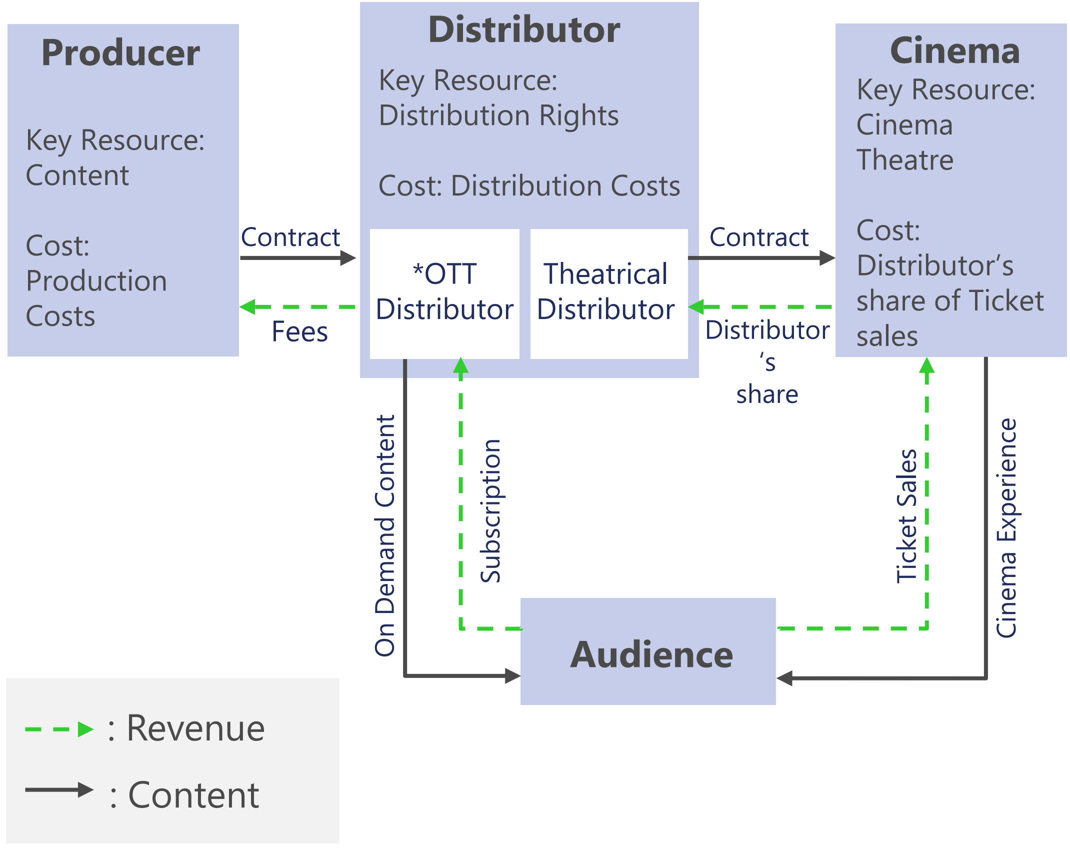
VIA: BUSINESS INSIDER

CREDIT: GETTY

Mass Media & Entertainment Domain (1/2)

Producers, a potential customer for the movie review sentiment analyzer

Business Model



Details

Description

- Movie Production Studios establish contractual relation with Distributors to distribute their original content
- Distributors own the distribution rights to movie content which is shown across cinema or OTT platforms
  - ○ Theatrical distributors distribute film content to movie theatre for a portion of revenue generated from ticket sales
  - ○ OTT distributors make film content available to audience on demand on their platform in exchange for subscription fees

Top Players

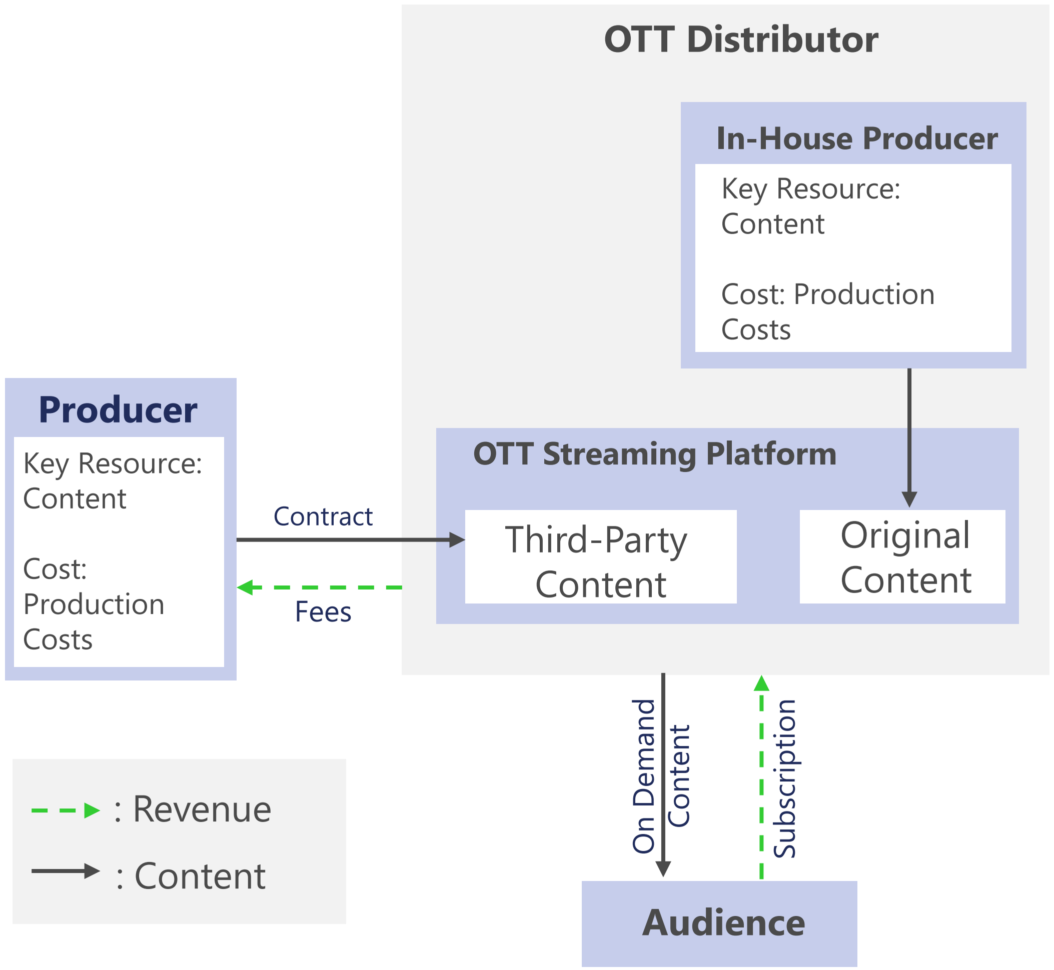
- Movie Production Studio:
  - The Walt Disney Studios (Global Box Office Sales: $79B)
  - Warner Bros. Entertainment Inc. (Global box Office Sales: $48B)
  - Universal City Studios LLC (Global Box Office Sales: $47.9B)

* OTT : Over-The-Top (Film content over the internet)
Global Box Office Sales figures based on 2022 data

Mass Media & Entertainment Domain (2/2)

Movie Review Sentiment Analyzer a useful tool for OTT Distributors Original Content creation

Business Model



Details

Description

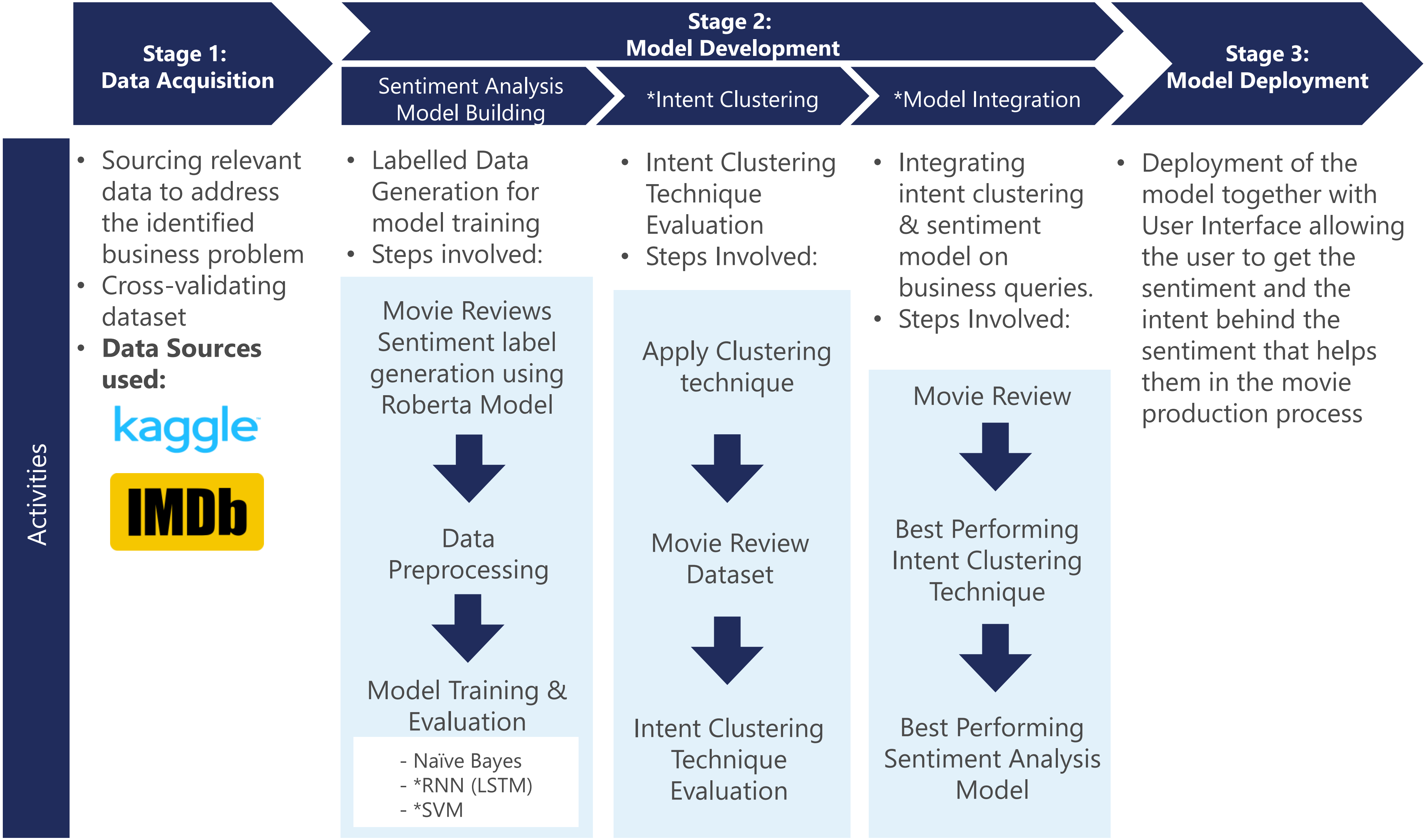
- Some OTT distributors like Netflix, Amazon prime Video, Disney+ have also entered the content production space
  - Inhouse production unit to develop and stream original content
- Subscription models adopted by popular OTT distributors:
  - Subscription Video On Demand (SVOD): Monthly plan for unlimited access to content anytime
  - Transactional Video On Demand (TVOD): Only pay for content audience want to watch
  - Hybrid Model: Combination of SVOD & TVOD

Top Players

- OTT Distributor:
  - Netflix (Global Subscribers: 231M)
  - Amazon Prime Video (Global Subscribers: 200M)
  - Disney+ (Global Subscribers: 138M)

* OTT : Over-The-Top (Film content over the internet)
Global Subscriber figures based on 2023 data

Data Pipeline (1/3) : Overview



Stage 1:
Data Acquisition

Activities

- Sourcing relevant data to address the identified business problem
- Cross-validating dataset
- **Data Sources used:**

kaggle

IMDb

Stage 2:
Model Development

Sentiment Analysis Model Building

- Labelled Data Generation for model training
- Steps involved:

Movie Reviews Sentiment label generation using Roberta Model

Data Preprocessing

Model Training & Evaluation

- Naïve Bayes
- *RNN (LSTM)
- *SVM

*Intent Clustering

- Intent Clustering Technique Evaluation
- Steps Involved:

Apply Clustering technique

Movie Review Dataset

Intent Clustering Technique Evaluation

*Model Integration

- Integrating intent clustering & sentiment model on business queries.
- Steps Involved:

Movie Review

Best Performing Intent Clustering Technique

Best Performing Sentiment Analysis Model

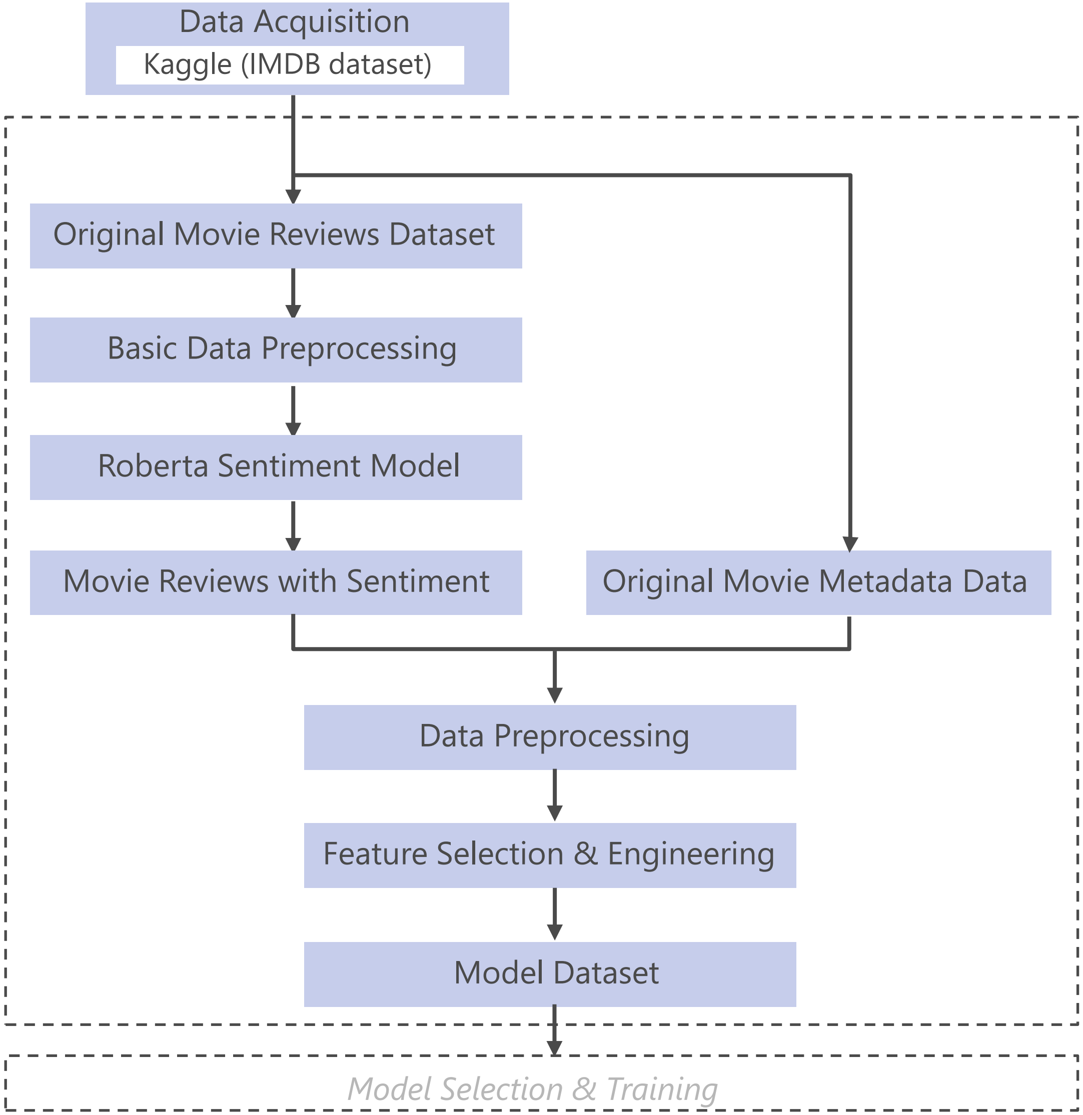
Stage 3:
Model Deployment

- Deployment of the model together with User Interface allowing the user to get the sentiment and the intent behind the sentiment that helps them in the movie production process

* To be performed after the mid-point delivery

Data Pipeline (2/3) : Data Preprocessing & Creation

Data Flow



Data Acquisition

Kaggle (IMDB dataset)

Original Movie Reviews Dataset

Basic Data Preprocessing

Roberta Sentiment Model

Movie Reviews with Sentiment

Original Movie Metadata Data

Data Preprocessing

Feature Selection & Engineering

Model Dataset

Model Selection & Training

Details

Description

- RoBERTa Sentiment Model for generating sentiment labels on movie reviews since it has higher accuracy than Vader

Data Pre-Processing

- Null values and duplicate records
- ** Non-grammatical text (emails and URLs)
- Non-ascii and diacritics characters
- Emojis, Slangs and * Abbreviations
- Word contractions
- Name Entity Recognition
- *Spellchecker and * POS Tagging
- Lowercasing and ** whitespaces
- Non-alphanumeric characters
- ** Stopwords and lemmatization (Spacy)

Feature Selection & Engineering

- Count Vectorizer (better than TF-IDF)
- Label encoding:

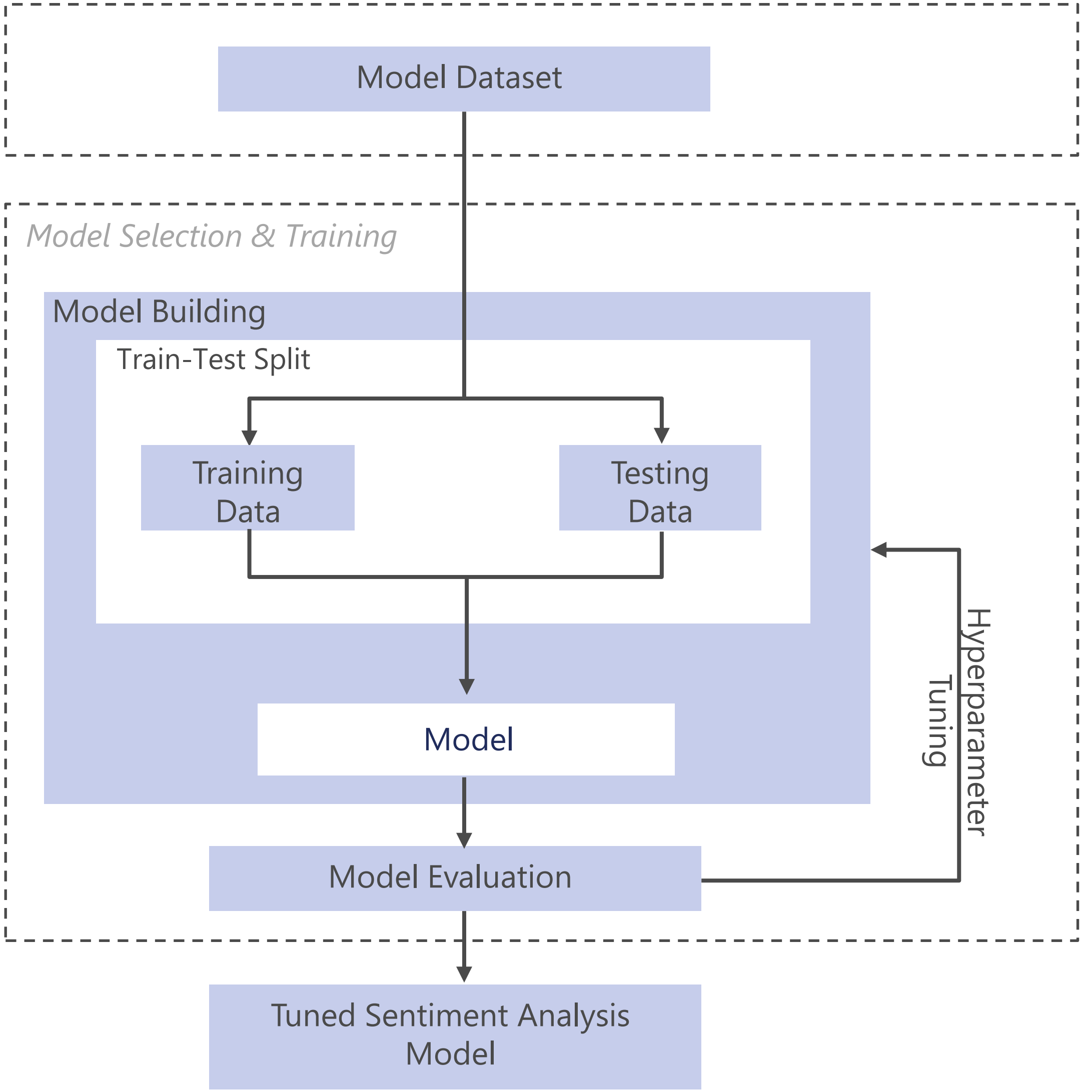
  Sentiment Labels → Numbers
  - 0: Negative
  - 1: Neutral
  - 2: Positive

* Excluded from the most recent sentiment analysis model

** Basic data pre-processing used on sentiment generator for movie reviews

Data Pipeline (3/3) : Model Building and Evaluation

Data Flow



Details

Model Creation

- Employed **Multinomial Naïve Bayes** for the initial model as it is robust for overfitting
- Percentage of training data : 80%

Model Evaluation

- Classification Report
  - Accuracy: 0.71
  - **Precision**: 0.64 (Negative), 0.63 (Neutral), 0.78 (Positive)
  - Recall
  - F1-score
- *ROC curve and AUC:
  - 0.88 (Negative, Positive), 0.76 (Neutral)

Hyperparameter Tuning

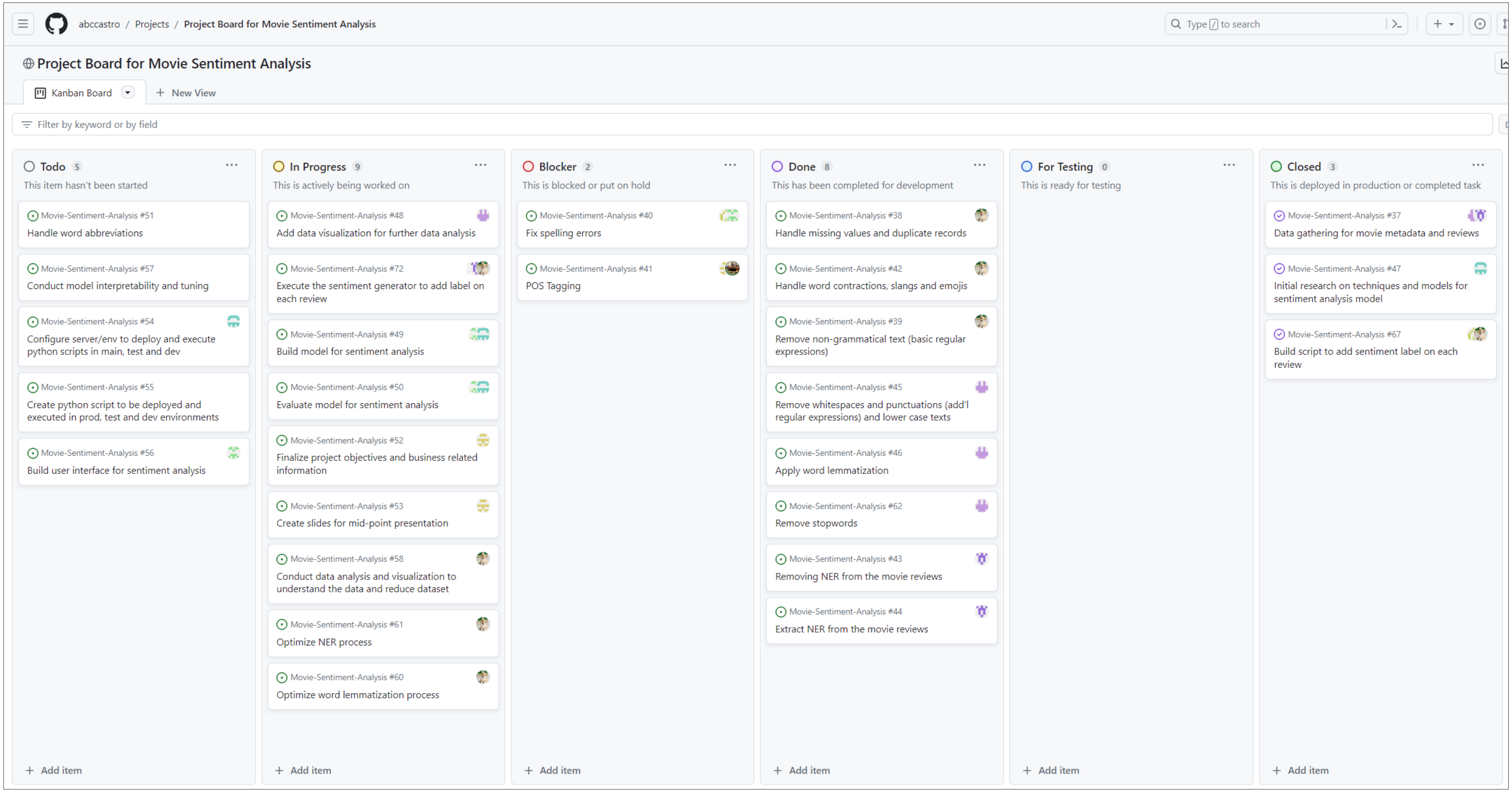
- GridSearchCV for hyperparameter tuning
- Parameter: Alpha

Note: Minor change when adjusting alpha; higher alpha values improve accuracy, while lower alpha values improve precision. Alpha 1 has an optimal balance across accuracy, precision, recall, and F1-score

* ROC (Receiver Operating Characteristic) and AUC (Area Under the ROC Curve)

Project Board Walkthrough

Using Github's Project to create Kanban Board for model development



abccastro / Projects / Project Board for Movie Sentiment Analysis

Type / to search

Project Board for Movie Sentiment Analysis

Kanban Board | + New View

Filter by keyword or by field

Todo 5

This item hasn't been started

- Movie-Sentiment-Analysis #51 — Handle word abbreviations
- Movie-Sentiment-Analysis #57 — Conduct model interpretability and tuning
- Movie-Sentiment-Analysis #54 — Configure server/env to deploy and execute python scripts in main, test and dev
- Movie-Sentiment-Analysis #55 — Create python script to be deployed and executed in prod, test and dev environments
- Movie-Sentiment-Analysis #56 — Build user interface for sentiment analysis

+ Add item

In Progress 9

This is actively being worked on

- Movie-Sentiment-Analysis #48 — Add data visualization for further data analysis
- Movie-Sentiment-Analysis #72 — Execute the sentiment generator to add label on each review
- Movie-Sentiment-Analysis #49 — Build model for sentiment analysis
- Movie-Sentiment-Analysis #50 — Evaluate model for sentiment analysis
- Movie-Sentiment-Analysis #52 — Finalize project objectives and business related information
- Movie-Sentiment-Analysis #53 — Create slides for mid-point presentation
- Movie-Sentiment-Analysis #58 — Conduct data analysis and visualization to understand the data and reduce dataset
- Movie-Sentiment-Analysis #61 — Optimize NER process
- Movie-Sentiment-Analysis #60 — Optimize word lemmatization process

+ Add item

Blocker 2

This is blocked or put on hold

- Movie-Sentiment-Analysis #40 — Fix spelling errors
- Movie-Sentiment-Analysis #41 — POS Tagging

+ Add item

Done 8

This has been completed for development

- Movie-Sentiment-Analysis #38 — Handle missing values and duplicate records
- Movie-Sentiment-Analysis #42 — Handle word contractions, slangs and emojis
- Movie-Sentiment-Analysis #39 — Remove non-grammatical text (basic regular expressions)
- Movie-Sentiment-Analysis #45 — Remove whitespaces and punctuations (add'l regular expressions) and lower case texts
- Movie-Sentiment-Analysis #46 — Apply word lemmatization
- Movie-Sentiment-Analysis #62 — Remove stopwords
- Movie-Sentiment-Analysis #43 — Removing NER from the movie reviews
- Movie-Sentiment-Analysis #44 — Extract NER from the movie reviews

+ Add item

For Testing 0

This is ready for testing

+ Add item

Closed 3

This is deployed in production or completed task

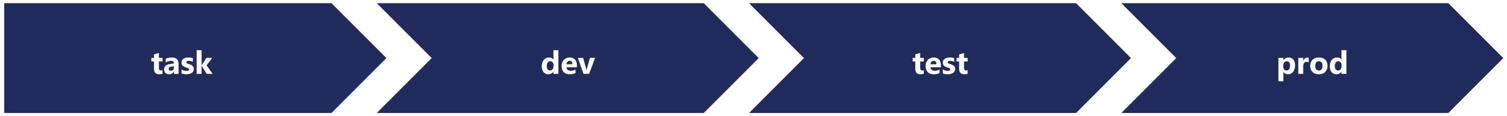
- Movie-Sentiment-Analysis #37 — Data gathering for movie metadata and reviews
- Movie-Sentiment-Analysis #47 — Initial research on techniques and models for sentiment analysis model
- Movie-Sentiment-Analysis #67 — Build script to add sentiment label on each review

+ Add item

* Project Board for Movie Sentiment Analysis: https://github.com/users/abccastro/projects/1

Team’s Best Practices

- **Feature Breakdown:** Dividing features into smaller tasks for better prioritization and more manageable components
- **Distributed Responsibilities:** Assigning tasks to individual team members to ensure clear responsibility and accountability
- **Project Progress Checkpoint:** Conducting regular team meetings, via MS Teams or in person, for updates, blockers and planning next task
- **Model Development Workflow:** Leveraging * GitHub as a model development repository, adhering to industry standards with distinct branches for production, testing, development, and tasks, and conducting code reviews before merging



* Movie Sentiment Analysis repository: https://github.com/abccastro/Movie-Sentiment-Analysis

Lambton
College

References

- Cheema, Ramish. (2022, November 14). 5 Biggest Movie Production Companies in the World. *Insider Monkey*. https://www.insidermonkey.com/blog/5-biggest-movie-production-companies-in-the-world-1085847/?singlepage=1
- Violini, Marcello. (2023, October 15) Top 10 OTT Video Streaming Services 2023. Teyuto. https://teyuto.com/blog/top-10-ott-video-streaming-services-2023#:~:text=Which%20ott%20platform%20has%20the,200%20million%20subscribers%20in%202023
- (2022 December 07). The Business Model in Film and Television Industries https://www.storybiz.tech/entertainment/business-model-film-television/
- Banik, R. (n.d.). The Movies Dataset. Kaggle. https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
- Zhang, X. (n.d.). IMDb Vision and NLP Dataset. Kaggle. https://www.kaggle.com/datasets/raynardj/imdb-vision-and-nlp/data