Data & Novelty

IMDB

Movie

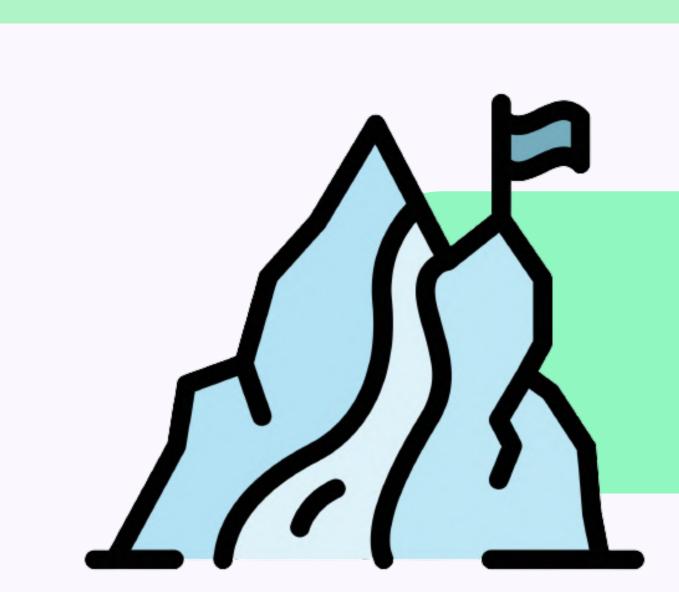
Negative

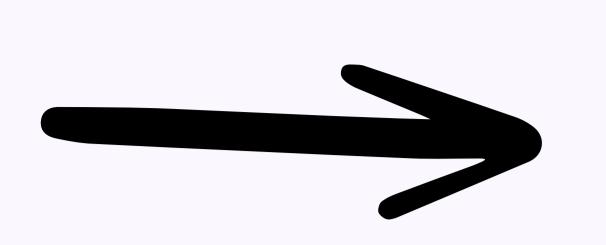
Rating

Positive

Rating...

² Pre-processing & Pipelining

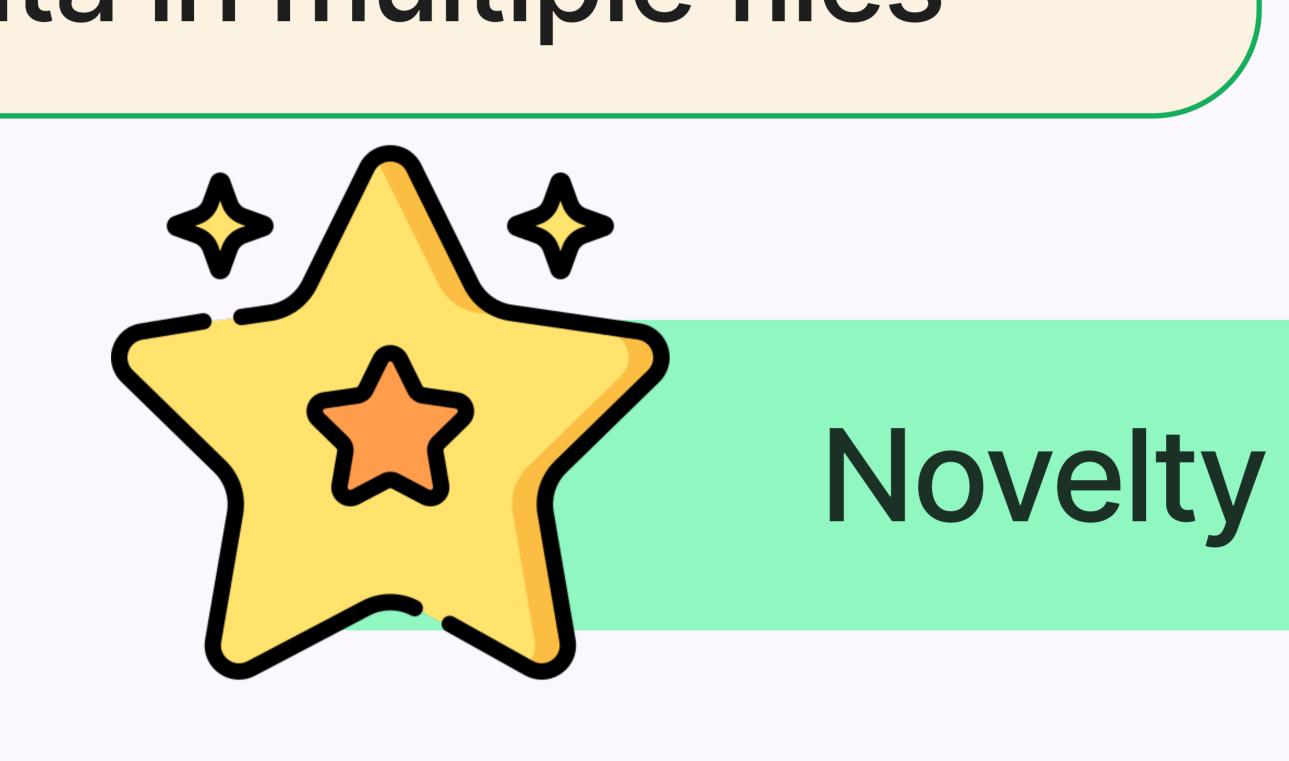




Predict a positive or negative IMDB movie rating...

Data exploration

- 10.000 IMDB movies with a Negative or Positive Rating (8k training, 1k test, 1k validation)
- Dirty data: loose csv files, typo's, multiple genre's etc.
- Training data in multiple files



- Generated novel contextual features by generating 10K descriptions of the Movie Plots (max 128 tokens) through OpenAl's GPT3 API, prompting movie name, year and duration.
- Generated the variable Awards won, through OpenAl chatGPT, prompting movie name, year and duration.
- Import new feature: Genre from Wikipedia
- Feeding the movie plot generations to a BERT-language model to predict the probabilities of good/bad rating.
- Ensembling: Averaging both the BERT probabilities and the predictions of an LGBM model, trained on other features.

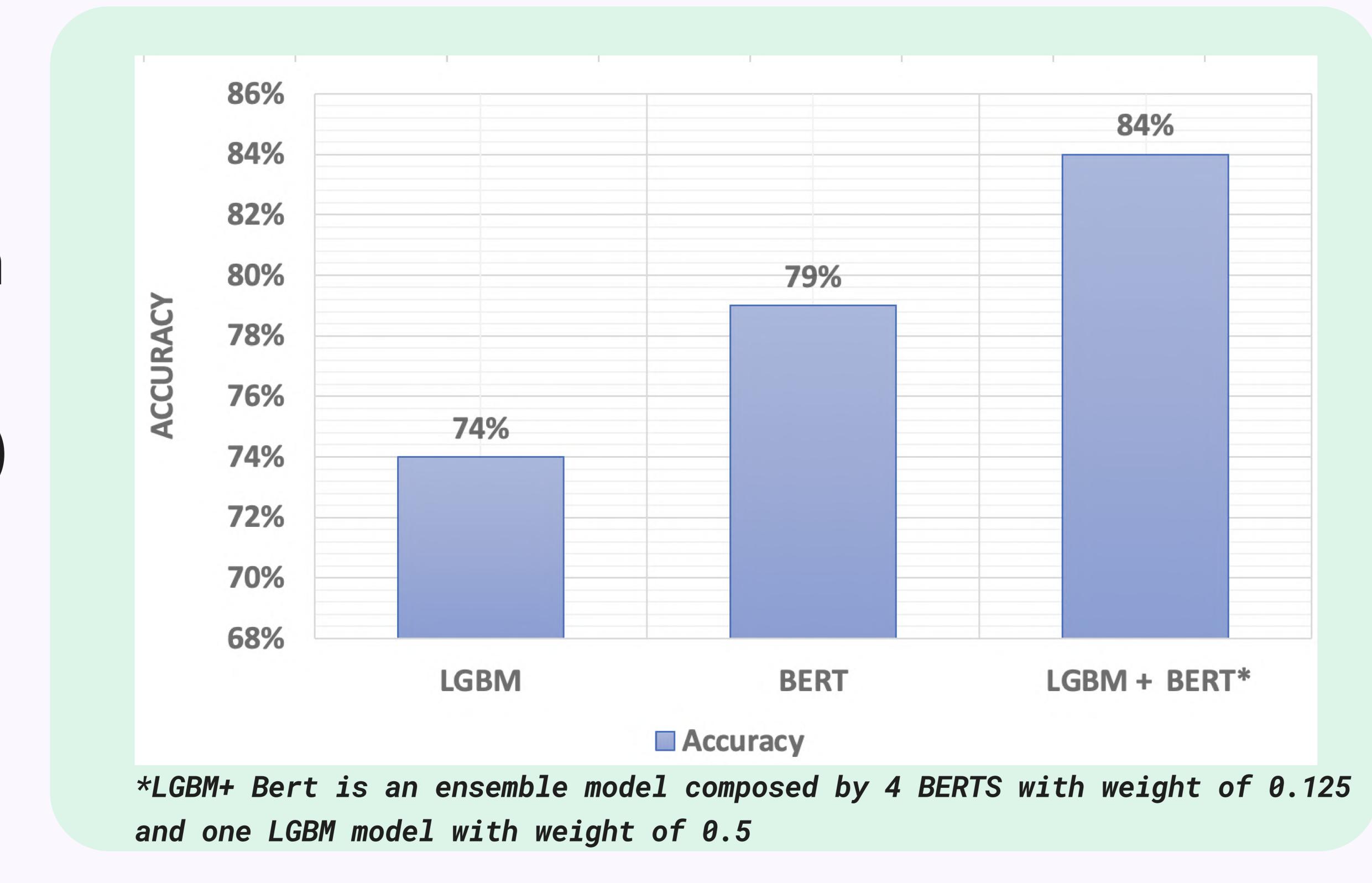
Learning performance

BERT Training Params:

- 7959 inputs each 128 tokens
- Bert for Sequence Classification
- Data tokenized and encoded
- Scheduler (with warm-up steps)
- Adam optimizer.

BERT + LGBM Training Params:

- 7959 rows, 22 features
- Default parameter settings



Limitations & Future Work

- Limited size of 'Movie Plot' Feature up to 128 tokens.
- Improve pipeline re-usability by taking more advantage of Spark's Pipeline Classes also in ML part.
- Improve stability by controlling repeatability of GPT's prompt output by lowering randomness (or temperature setting) in API.
- Increase accuracy by using advanced language model like GPT4
- Increase model accuracy by collecting descriptive information on Movie Ratings such as country, marketing budget, and network effects on various platforms.

LGBM input Year Genre (Wikipedia) Number of Votes Duration Awards (GPT) **BERT Input** Movie

... Generate Awards

(GPT)

ChatGPT Prompt:

"Did the Movie {PrimaryTitle} from {StartYear} that has a duration of {runtimeMinutes} minutes won Any Awards? Answer in : "Yes" or "No"

...BERT Predict Rating Probabilities on 'Plots'

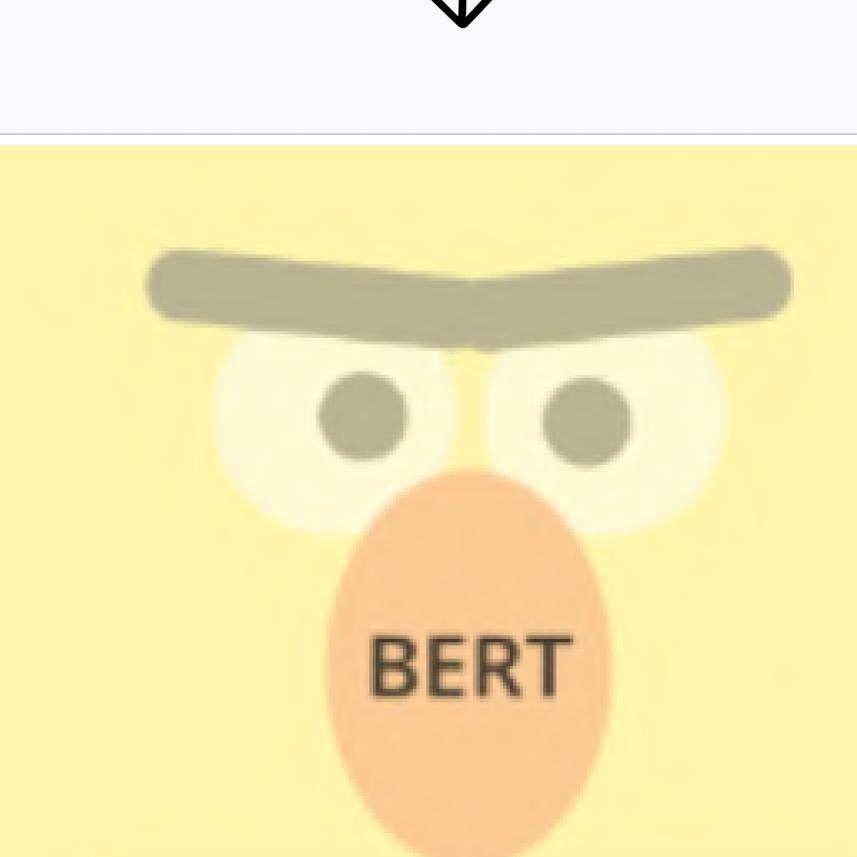
The beverly hillbillies is a 1993 comedy film . The movie follows the clampett family, a poor backwoods family from the ozarks, as they move to beverly hills after striking it rich with an oil well. led by jed clampett (jim varney), the family must adjust to their newfound wealth and life in the big city. along the way, they encounter a variety of characters, including a conniving banker and a snooty socialite ...

Al Generated Train File Test file Data PySpark Transformers: 1. Concatenate trains 2. Impute NULLs, remove

- \N in year columns
- 3. Lowercase titles
- 4. Cast to INT numeric columns
- 5. Binning Years
- 6. Impute and scale number of votes
- 7. Rename Wikipedia
- columns to add genres
- 8. One hot encode genres
- 9. Drop columns
- Call to OpenAI API
- Generate Movie plot Generate Awards

Train and Infer

BERT model



SOCIA

... Preproc. with Spark

- User Defined Functions(UDF)
- Spark Transformers and Estimators
- · Created re-utilizable PySpark pipeline object



- Enhancing traditional ML with Large language models
- ... Generate Movie plots

ChatGPT Prompt: "Give me a roughly 100 words plot of the {PrimaryTitle} from {StartYear}that has a duration of {runtimeMinutes}

temperature=0, max_tokens=128, top_p=1, frequency_penalty=0.5, presence_penalty=0

Juliet of the spirits," is a 1965 italian comedydrama film directed by federico fellini. the film follows the story of juliet, a middle-aged housewife living in rome with her husband, giorgio. juliet begins to suspect that giorgio is having an affair and decides to confront him. in the process, she discovers her own inner strength and begins to explore her own spirituality. she learns more about herself and discovers that she has the power to make changes"

p=0.961

p = 0.048

...LGBM Predict Rating Probabilities on other features

p = 0.12



• Plugin number of Votes, Awards, Duration etc.

Train and infer LGBM

Model

p = 0.88

...Ensemble Probabilities!

Classify Rating

