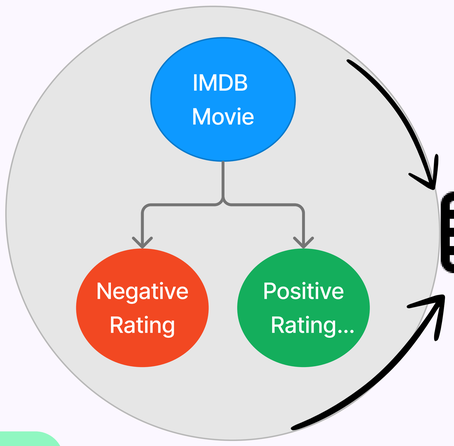
1 Data & Novelty

Mission

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

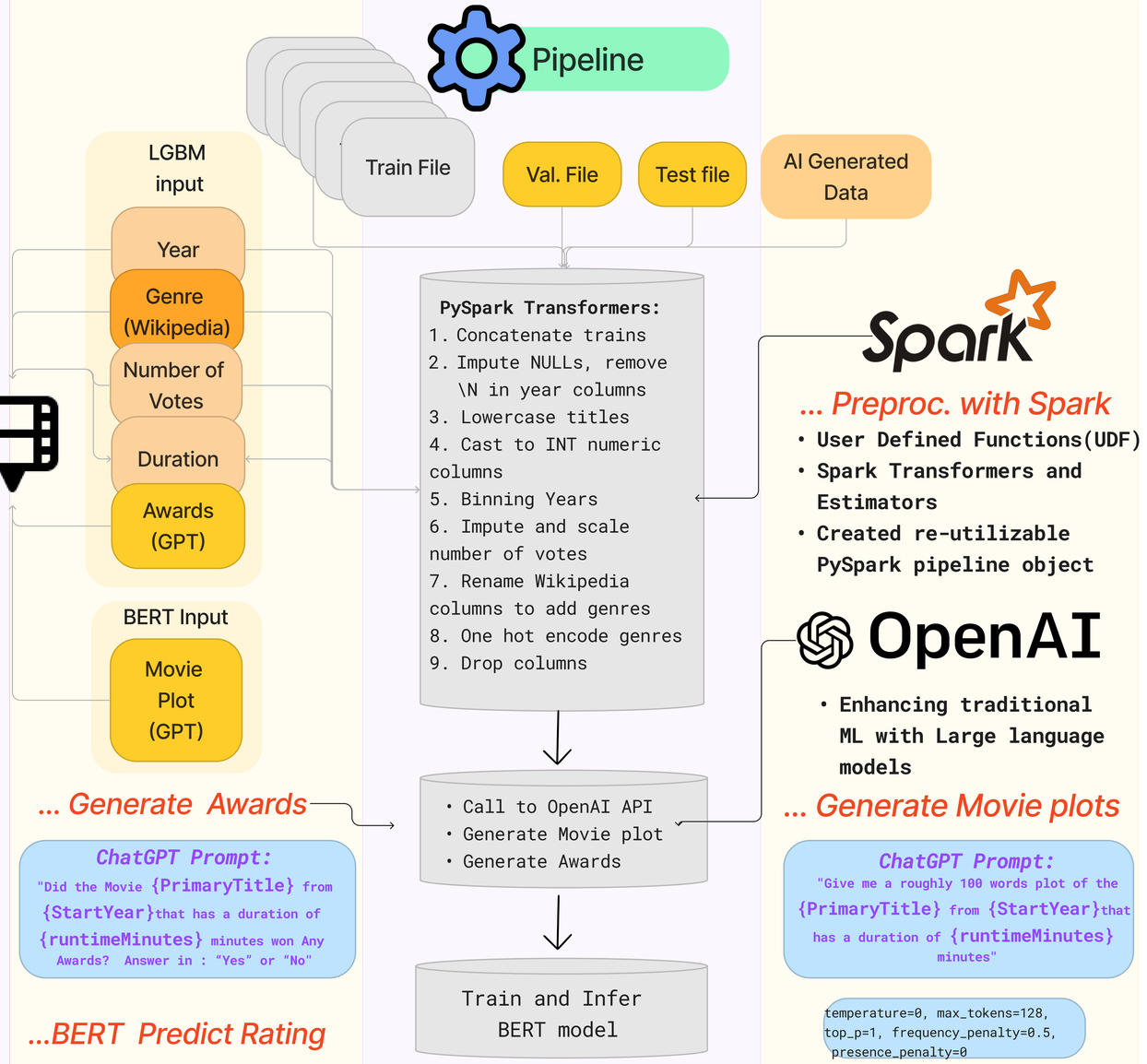


Novelty

- Generated novel contextual features by **generating 10K** descriptions of the **Movie Plots** (max 128 tokens) through OpenAI's GPT3 API, prompting movie name, year and duration.
- Generated the variable **Awards** won, through OpenAI 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.

2 Pre-processing & Pipelining

Pipeline



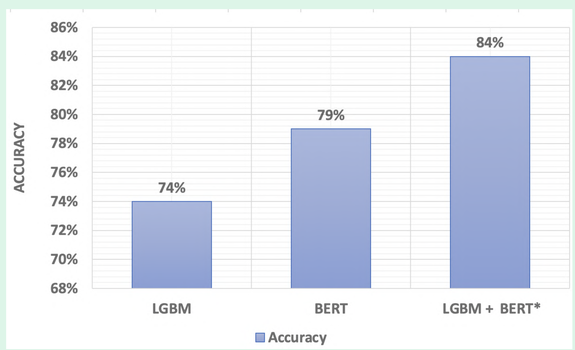
3 Learning performance

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



**LGBM+ Bert is an ensemble model composed by 4 BERTS with weight of 0.125 and one LGBM model with weight of 0.5*

4 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.*