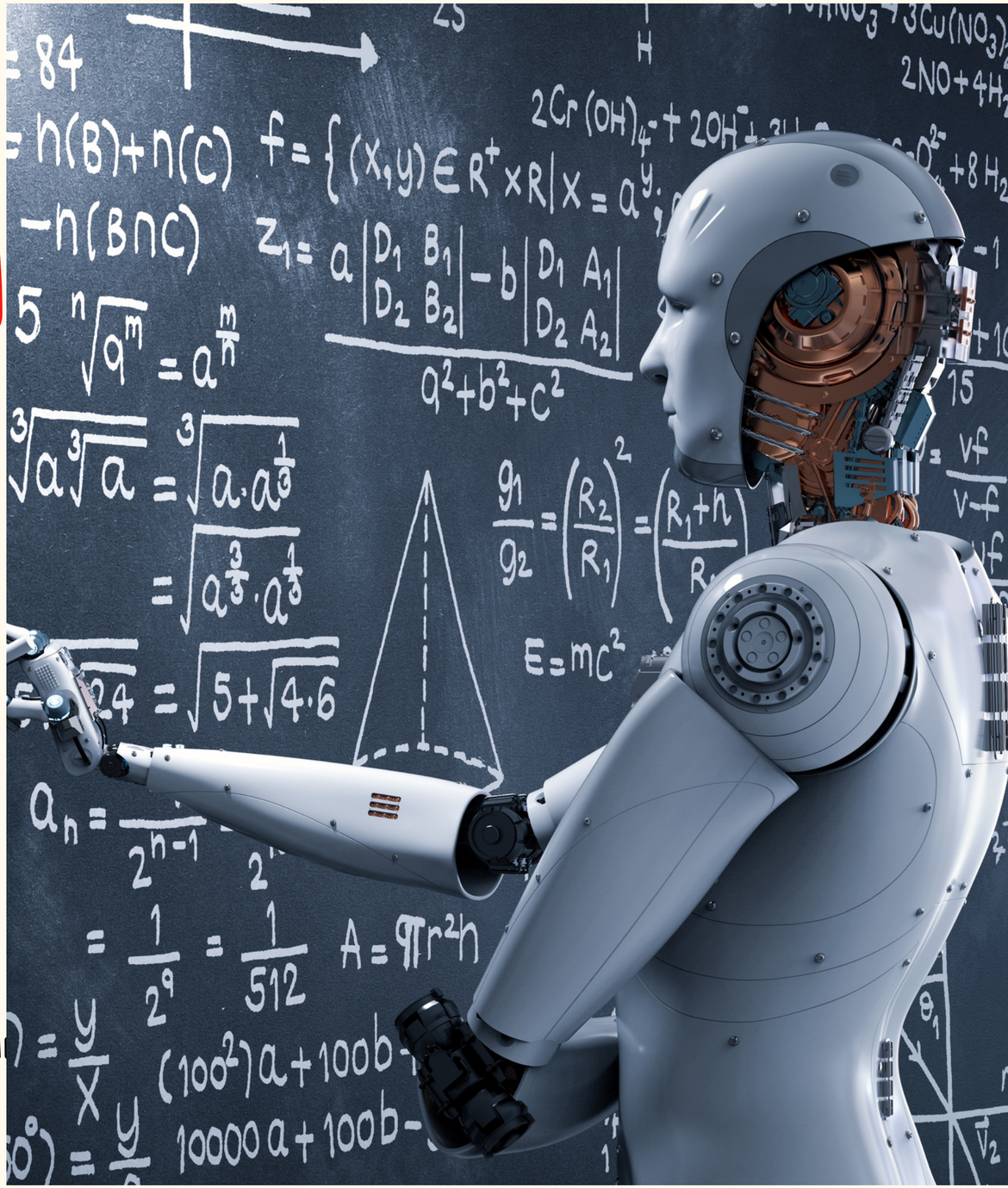
HTTPS://GITHUB.COM/INSIGHTBUILDER

# MASTERING HUGGINGFACE MODEL EVALUATION

## IN-DETAIL WALKTHROUGH OF MEASUREMENT, METRIC & COMPARISON



# TRAINING A NLP MODEL : 5 STEPS

**STEP 1 :** TRAINING DATA TO BE PREPROCESSED IN THE FORMAT MODEL EXPECTS

DEPENDING ON MEDIA, A TOKENIZER, FEATURE EXTRACTOR OR IMAGE PROCESSOR WILL BE USED

**STEP 2:** LOAD A PRE-TRAINED MODEL CHECKPOINT SPECIFIC TO YOUR TASK FOR FINE-TUNING PROCESS

**STEP 3:** INSTANTIATE THE TRAINING ARGUMENTS WHICH CONTAINS THE HYPERPARAMETERS

**STEP 4:** INSTANTIATE EVALUATE INSTANCE THAT WILL **HELP TRAINER TO CALCULATE METRICS**

**STEP 5:** CREATE THE TRAINER INSTANCE THAT CONTAINS

- MODEL
- TRAINING ARGS + **EVALUATION METRIC**
- TRAIN + EVAL  DATASETS

# CHALLENGE SOLVED : EVALUATE

1. HELPS IN EVALUATING THE MACHINE LEARNING PIPELINE.
    A. METRIC: EVALUATE MODEL PERFORMANCE W.R.T PREDICTIONS AND GROUND TRUTH
    B. COMPARISON: EVALUATE TWO MODELS FOR THEIR METRICS AND PERFORMANCE
    C. MEASURE: INVESTIGATE THE DATASET PROPERTIES AND FEATURES
2. EACH SOLUTION IS SEPERATE MODULES INSIDE EVALUATE LIBRARY, WITH ITS OWN DESCRIPTION, FEATURES AND INPUTS
3. COMPUTING THE METRICS CAN BE DONE EITHER ONE BY ONE, BATCH WISE OR AS A WHOLE
4. ML PIPELINES WILL REQUIRE MULTIPLE EVALUATIONS, FOR THAT EVALUATE HAS COMBINE METHOD
5. EVALUATOR MODULE AUTOMATES THE MODEL EVALUATION WITH MODEL + DATASET + METRIC. (THIS MAKES THINGS VERY EASY)

HTTPS://GITHUB.COM/INSIGHTBUILDER

# WHAT ARE THE METRICS/ MEASUREMENT

## MODEL IS TRAINED. LETS SEE HOW WELL THE MODEL IS DOING

1. GENERIC METRICS : CAN BE APPLIED TO MULTIPLE SITUATION, LIKE ACCURACY & PRECISION
2. TASK SPECIFIC METRICS : USEFUL FOR EVALUATING A PARTICULAR TASKS, EX : SEQEVAL
3. DATASET SPECIFIC METRICS : EVALUATE ON SPECIFIC BENCHMARKS LIKE GLUE

## HOW TO FIND THE METRICS FOR THE TASKS?

1. TASK PAGES IN HUGGINGFACE PROVIDES THE OUTPUT
2. USE THE LEADER BOARDS LIKE HTTPS://PAPERSWITHCODE.COM/
3. EACH OF THE DATASET WILL HAVE ITS ASSOCIATED METRICS AND TASKS IT WILL SUPPORT

## CUSTOM METRICS CAN ALSO BE CREATED

## EVALUATOR CAN BE USED TO EVALUATE THE TRIPLETS OF MODEL, DATASET AND METRIC. ONLY 9 TASKS ARE SUPPORTED BY EVALUATOR AT THIS MOMENT.

## EVALUATE LIBRARY ALSO HAS METHODS TO VISUALIZE THE DATA GENERATED

# EXAMPLES OF METRICS/ COMPARISON/ MEASUREMENT

## METRICS:

**53 DIFFERENT METRICS**

- METEOR
- SEQEVAL
- PEARSON CORR
- INDICGLUE
- RL RELIABILITY
- NIST_MT
- MANY MORE...

## COMPARISON:

**3 DIFFERENT COMPARERS**

- MCNEMAR
- EXACT MATCH
- WILCOXON

## MEASUREMENT:

**8 DIFFERENT MEASURES**

- LABEL DISTRIBUTION
- WORD COUNT
- WORD LENGTH
- REGARD
- PERPLEXITY
- HONEST
- TOXICITY
- TEXT DUPLICATES

HTTPS://GITHUB.COM/INSIGHTBUILDER

# PRACTICING METRICS WITH EVALUATE

## DIVING INTO THE COLAB NOTEBOOK TO FIND OUT HOW THE METRICS WORK IS KEY FOR SUCCESS

### CONSIDERATIONS FOR MODEL EVALUATION

- PROPERLY SPLITTING YOUR DATA
- THE IMPACT OF CLASS IMBALANCE
- OFFLINE VS. ONLINE MODEL EVALUATION
- TRADE-OFFS IN MODEL EVALUATION
- INTERPRETABILITY
- INFERENCE SPEED AND MEMORY FOOTPRINT
- LIMITATIONS AND BIAS

HTTPS://GITHUB.COM/INSIGHTBUILDER

# THANKS FOR WATCHING

HTTPS://GITHUB.COM/INSIGHTBUILDER