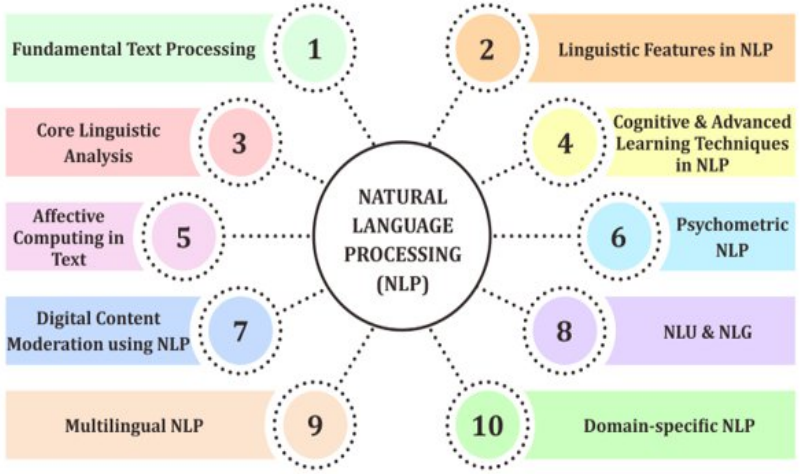
Evaluation Metrics

Dr. Dipika Jain

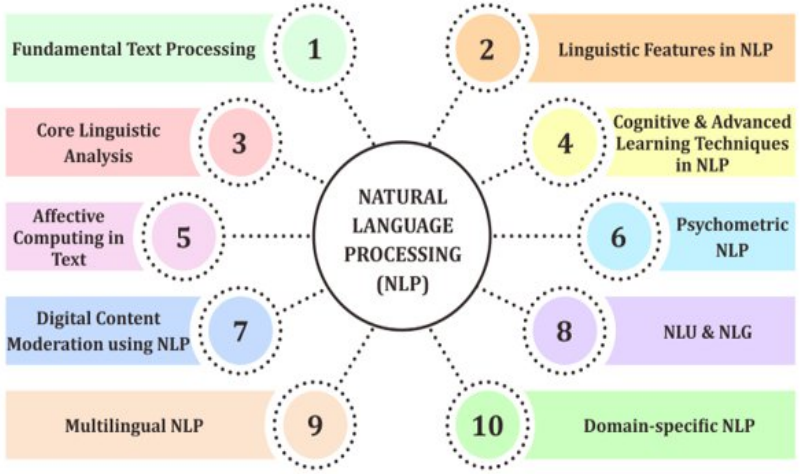
Lecture Outline

I. Intrinsic Evaluation vs. Extrinsic Evaluation
II. Qualitative vs. Quantitative Evaluation
III. Accuracy
IV. Precision
V. Recall
VI. F1 Score
VII. AUC
VIII. MRR: Mean Reciprocal Rank.
IX. MAP: Mean average precision,
X. RMSE: Root mean squared error



Lecture Outline

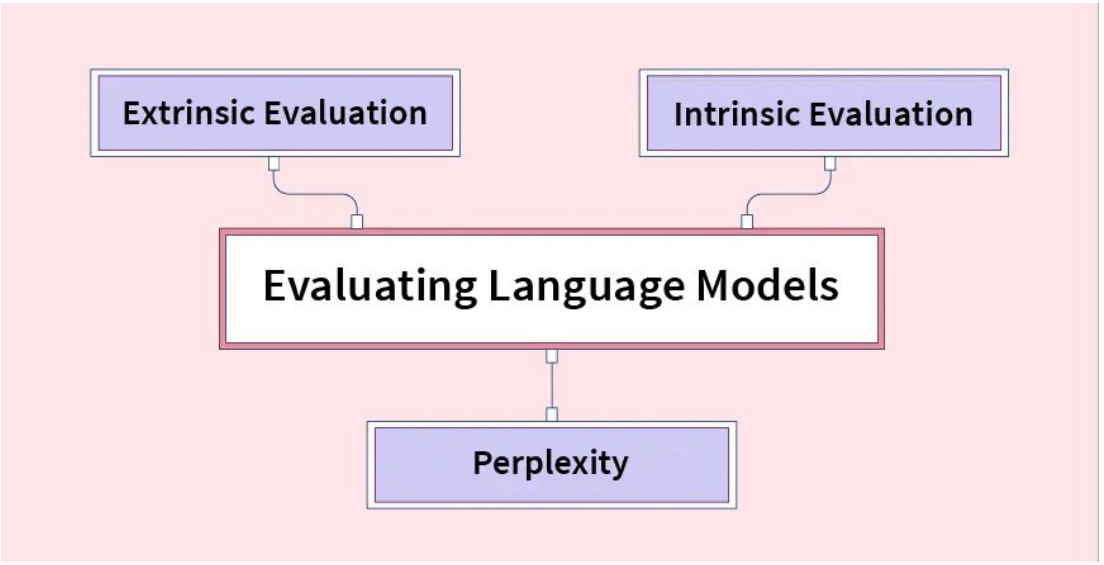
I. MAPE: Mean absolute percentage error.
II. BLEU, bilingual evaluation
III. METEOR: Precision-based metric
IV. ROUGE
V. Perplexity
VI. Log-likelihood
VII. Human Evaluation



Evaluation

Whenever we build Machine Learning models, we need some form of metric to measure the goodness of the model. Determining whether the model being used for a specific task is successful depends on 2 key factors:

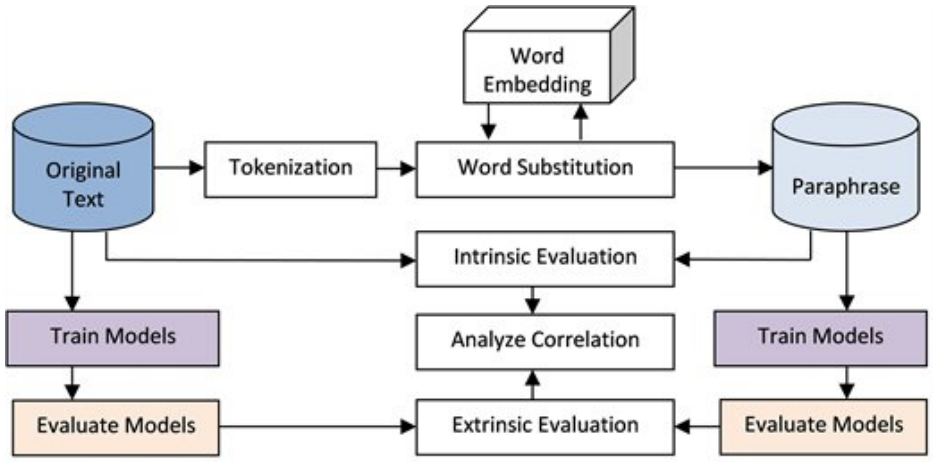
- Whether the evaluation metric we have selected is the correct one for our problem
- If we are following the correct evaluation process



Evaluation Types

Intrinsic evaluation - aims to measure the quality of embeddings by assessing their performance on specific NLP tasks that are related to the embedding space itself, such as word similarity, analogy, and classification.

Extrinsic evaluation - aims to measure the quality of embeddings by assessing their performance on downstream NLP tasks, such as machine translation or text classification, that are not directly related to the embedding space itself.



Intrinsic Evaluation Metrics

- ***Cosine similarity*** measures the similarity between two vectors by computing the cosine of the angle between them. In the context of embeddings, cosine similarity is often used to measure the similarity between two words, or between a word and its context. The formula for cosine similarity is as follows:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

where **A** and **B** are the embeddings of two words, and $|\cdot|$ denotes the Euclidean norm.

The Three Documents and Similarity Metrics

Considering only the 3 words from the above documents: 'sachin', 'dhoni', 'cricket'

Doc Sachin: Wiki page on Sachin Tendulkar	Doc Dhoni: Wiki page on Dhoni	Doc Dhoni_Small: Subsection of wiki on Dhoni
Dhoni - 10	Dhoni - 400	Dhoni - 10
Cricket - 50	Cricket - 100	Cricket - 5
Sachin - 200	Sachin - 20	Sachin - 1

Document - Term Matrix (Word Counts)

Word Counts	*"Dhoni"*	*"Cricket"*	*"Sachin"*
Doc Sachin	10	50	200
Doc Dhoni	400	100	20
Doc Dhoni_Small	10	5	1

Similarity Metrics

Similarity or Distance Metrics	*Total Common Words*	*Euclidean distance*	*Cosine Similarity*
Doc Sachin & Doc Dhoni	10 + 50 + 10 = **70**	432.4	0.15
Doc Dhoni & Doc Dhoni_Small	20 + 10 + 7 = **37**	204.0	0.23
Doc Sachin & Doc Dhoni_Small	10 + 10 + 7 = **27**	401.85	0.77

Intrinsic Evaluation Metrics

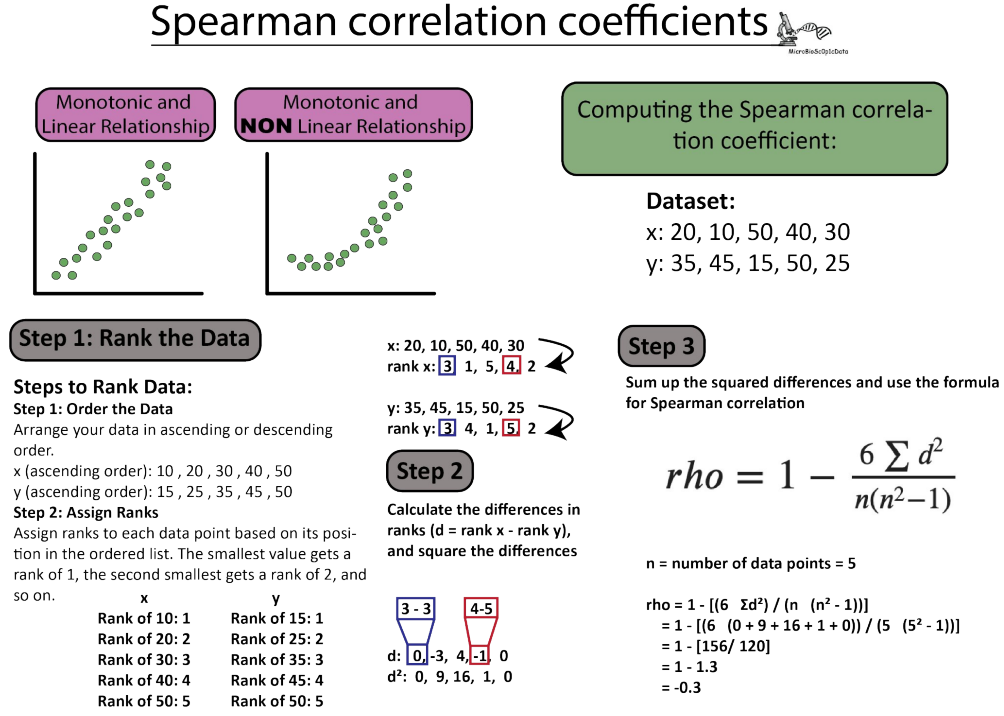
- ***Spearman Correlation*** measures the monotonic relationship between two variables, which can be the similarity scores of two sets of words or phrases computed by humans and by embeddings. The formula for Spearman correlation is as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

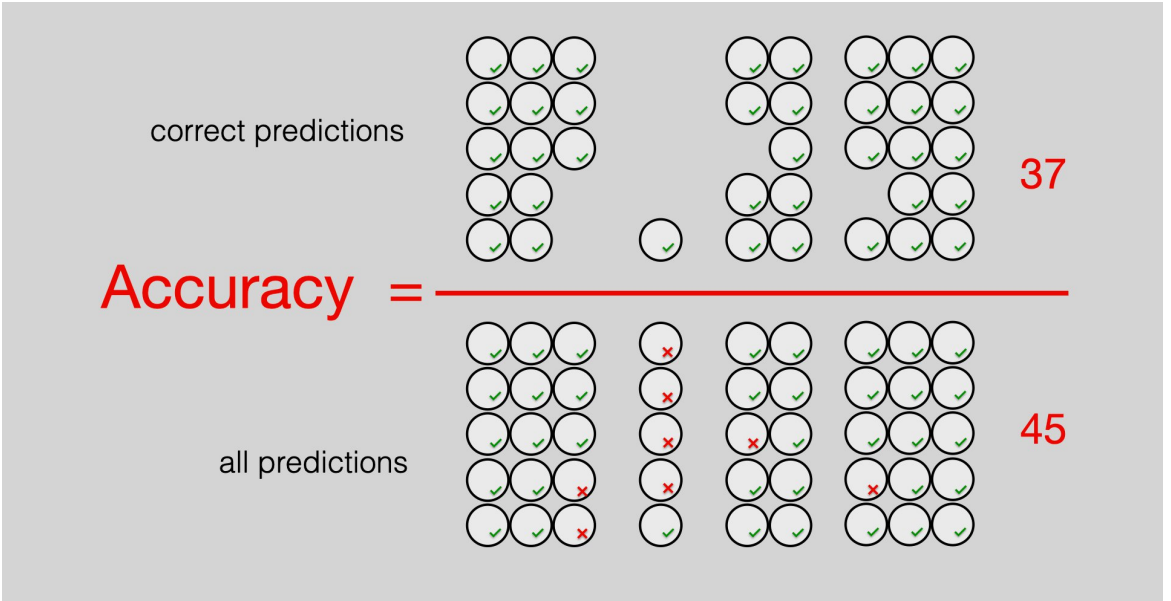
d_i = difference between the two ranks of each observation

n = number of observations



Intrinsic Evaluation Metrics

- ***Accuracy*** measures the performance of embeddings on classification tasks, such as sentiment analysis or topic classification. Given a dataset of labeled examples, the embeddings are used to represent each example, and a classifier is trained on these representations..



Advantages of Intrinsic Evaluation

- ***Task-Specific:*** Intrinsic evaluations are task-focused, providing insights into how well the model performs on a particular NLP task.
- ***Quick Feedback:*** Results are obtained relatively quickly, allowing for rapid model iterations and improvements.
- ***Benchmarking:*** Intrinsic evaluations often involve widely accepted benchmarks, making it easier to compare models and track progress.
- ***Focused Metrics:*** Metrics such as accuracy, precision, recall, and F1-score provide detailed insights into model capabilities.
- ***Controlled Environment:*** Researchers can control and manipulate evaluation conditions to gather precise data.

Extrinsic Evaluation Metrics

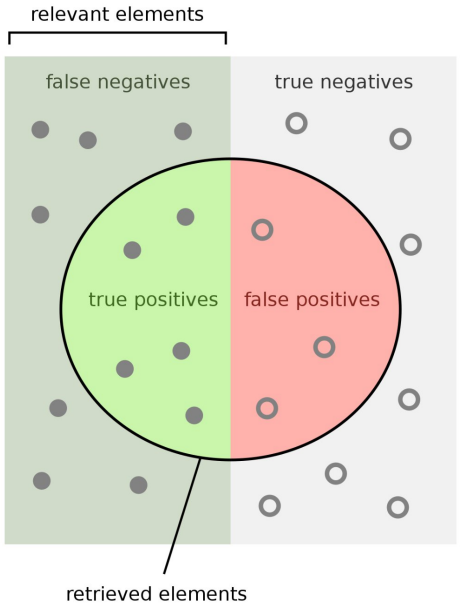
- ***F1 score*** is a metric commonly used in binary classification problems, such as sentiment analysis or named entity recognition. It combines precision and recall into a single score that ranges from 0 to 1. A high F1 score indicates that the embeddings are able to capture the relevant features of the input data. The formula for F1 score is as follows:

ACTUAL VALUES	POSITIVE	NEGATIVE
POSITIVE	TP	FN
NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$



Extrinsic Evaluation Metrics

- ***Perplexity*** is a metric commonly used in language modeling tasks, such as machine translation or text generation. It measures how well a language model can predict a held-out test set of text, given the embeddings as input. A low perplexity indicates that the embeddings are able to capture the semantic and syntactic structures of the language. The formula for perplexity is as follows:

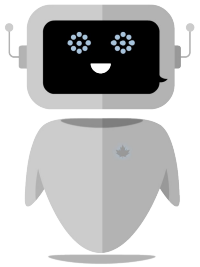
$$\text{perplexity} = \prod_{t=1}^{T} \left(\frac{1}{P_{\text{LM}}(\boldsymbol{x}^{(t+1)} | \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})} \right)^{1/T}$$

Normalized by number of words

Inverse probability of corpus, according to Language Model

Disadvantages of Intrinsic Evaluation

- ***Complexity:*** Designing and conducting extrinsic evaluations can be more resource-intensive and time-consuming than intrinsic evaluations.
- ***Subjectivity:*** Extrinsic evaluations may involve human judgment, introducing subjectivity in assessing the model's performance.
- ***Difficulty in Isolation:*** Isolating the model's contribution from other factors in a real-world application can be challenging.
- ***Dependent on the Application:*** The effectiveness of extrinsic evaluation heavily depends on the quality and complexity of the application.



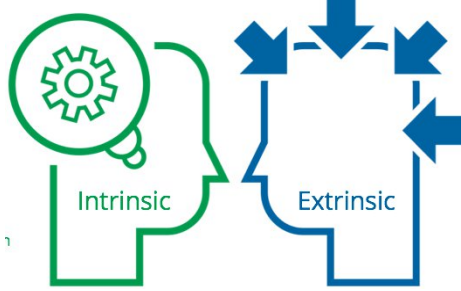
Use Case Example

Imagine you are developing a chatbot for a customer support service in an e-commerce company. The primary goal is to enhance user satisfaction and resolve customer queries efficiently.

- **Intrinsic Evaluation:** In this case, intrinsic evaluation might involve assessing the chatbot's language understanding capabilities, response time, and sentiment analysis accuracy. These metrics provide insights into how well the chatbot performs individual NLP tasks.

- **Extrinsic Evaluation:** Extrinsic evaluation would assess the chatbot's overall impact on customer satisfaction, response time reduction, and query resolution rate. This evaluation method considers the chatbot's real-world performance in the context of improving customer support.

When to use which approach?



- ***Intrinsic Evaluation:*** Use intrinsic evaluation when you want to fine-tune and assess the performance of individual NLP components or when benchmarking against specific tasks. It helps identify areas for improvement within the model.

- ***Extrinsic Evaluation:*** Choose extrinsic evaluation when you need to measure the model's effectiveness in real-world applications or when assessing its contribution to achieving broader goals. This approach provides insights into how well the model performs in practical scenarios.

Quantitative Evaluation

1. **Objective Measurement:**

- ***Focus:*** Quantitative evaluation deals with measurable and numerical aspects of a model's performance.
- ***Metrics:*** It often involves metrics like accuracy, precision, recall, F1 score, mean squared error, etc., depending on the nature of the task.

1. **Reproducibility:**

- ***Repeatability:*** Results are typically reproducible, as the metrics are based on objective and standardized measures.

1. **Statistical Rigor:**

- ***Statistical Analysis:*** Statistical methods are commonly used to analyze the significance of the observed metrics.

Quantitative Evaluation

4. **Scalability:**

- ***Applicability:*** Well-suited for scenarios where the primary concern is the efficiency and scalability of a model, especially in large-scale applications.

4. **Examples:**

- ***Accuracy:*** Percentage of correctly predicted instances.
- ***Precision:*** Proportion of true positives among all predicted positives.
- ***Recall (Sensitivity):*** Proportion of true positives among all actual positives.

Quantitative Evaluation Methods

- ***Descriptive:*** A descriptive analysis is a summarization of data points that describe the data. Statisticians like to describe these in terms of central tendency or spread, while non-statisticians look for narratives to describe what the data means.
  - A simple example of a narrative this could be a descriptive column on a report that reviews sales 12 months ago and last month and states:

    "Sales has grown from $1.2M in FM2022-12 to $1.4M in FM2023-12 based largely on growth in appliances."

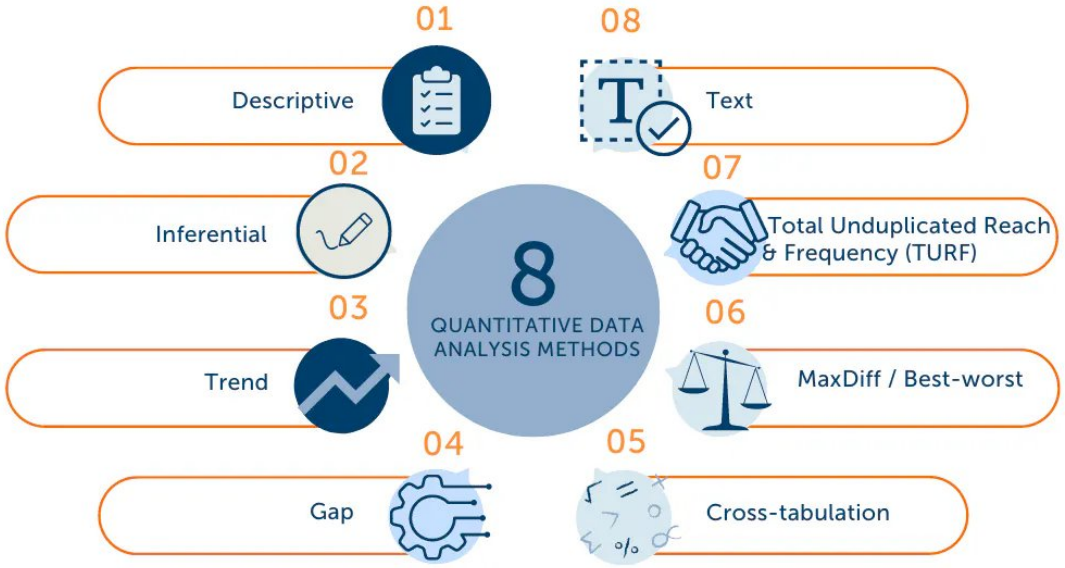


Table 1: Descriptive Statistics

	count	mean	sd	min	max
Age (years)	100	52.6	28.0	1	99
Gender (1=male)	100	0.71	0.46	0	1
Monthly Income (dollars)	100	2605.1	1403.0	82	4984

Source: our sample

Quantitative Evaluation Methods

- ***Inferential:*** Inferential analysis is the ability to draw conclusions based on reviewing samples of data
  - Example, when interest rates rise, house sales decrease. Analysts infer a connection between the two variables:

    "Because the cost of a mortgage is higher, fewer people are buying homes."
  - Two of the most common types of inferential statistics are:
    - **Regression analysis:** This is the act of evaluating across a population how one variable will change with respect to another. Linear regression is most common and is based on changes to an independent variable based on the values of its dependent variable.
    - **Hypothesis testing** is one type of inferential statistics that is used to ask a question and test the answers.

Inferential Statistics	
Hypothesis Testing	Regression Analysis
Z test	Linear Regression
F test	Nominal Regression
T test	Logistic Regression
ANOVA Test	Ordinal Regression
Wilcoxon Signed Rank Test	
Mann-Whitney U Test	

Quantitative Evaluation Methods

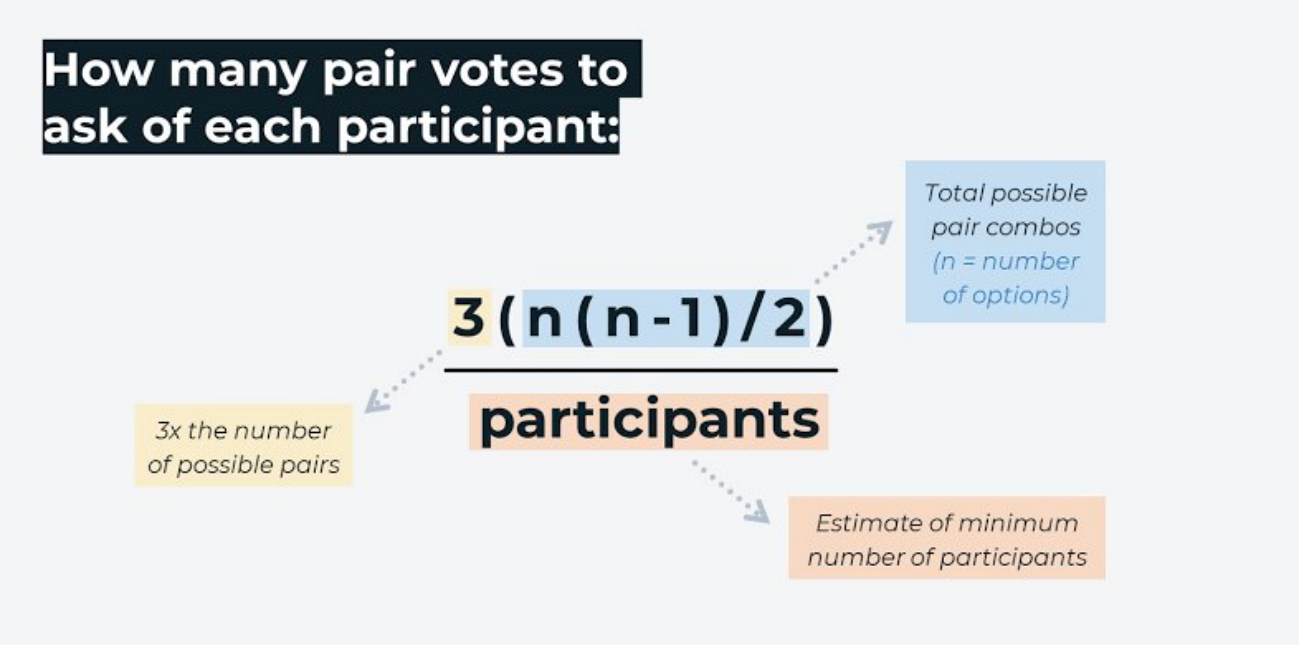
- ***Gap:*** In gap analysis, past and current state data is compared to evaluate performance or make decisions about what needs to be done to fix a problem. For example, past weekly sales employee hours can be compared to current weekly sales employee hours to determine whether more employees are needed.

- ***Cross-tabulation:*** A cross-tabulation or contingency table groups multiple variables together so that mathematical correlations are easier to find. For example, data about a person's age and the time of year when they purchase the most items can be aggregated to gain visibility into how people shop for the holidays.

	Below 30 years	Above 30 years	Total
Preference			
Denim	14	8	**22**
Cotton Trousers	6	12	**18**
Total	**20**	**20**	**40**

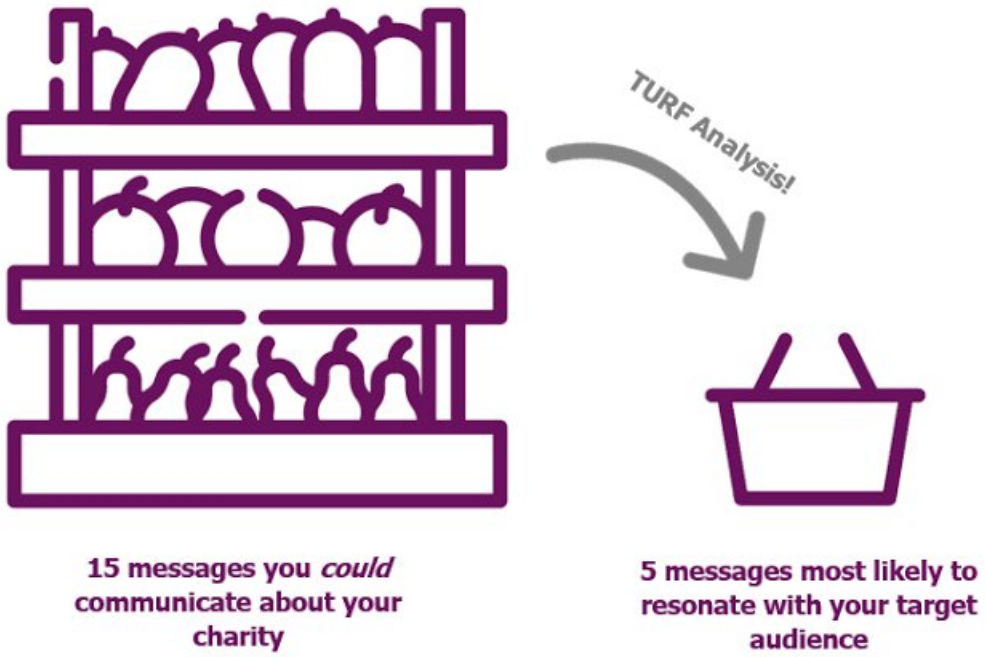
Quantitative Evaluation Methods

- ***MaxDiff / Best-worst:*** With a MaxDiff, you analyze how people responded to a survey's "most important to least important" scale question by creating a mean score for each point on the scale to determine the order of preference. This can also be used when doing market research to understand how important a new feature would be to customers.



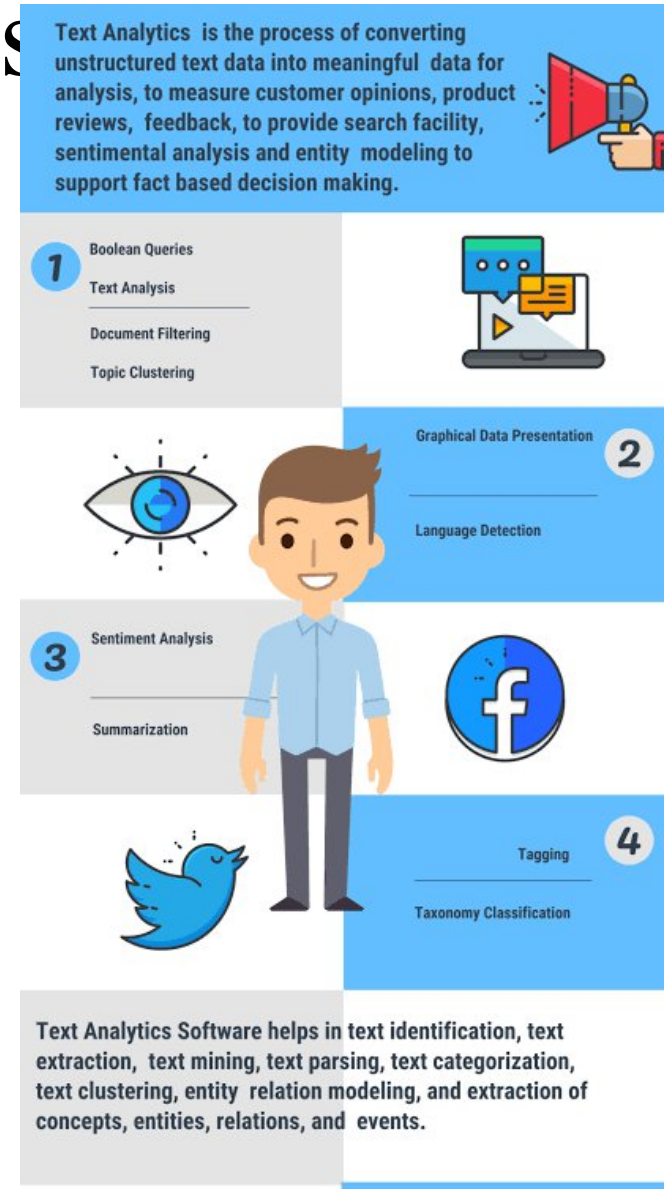
Quantitative Evaluation Methods

- ***Total Unduplicated Reach and Frequency (TURF):*** The TURF analysis methodology helps assess a combination of products and services by reviewing the number of customers reached in conjunction with how often the communication source reaches them. It's usually used in market research and often combined with a MaxDiff analysis.



Quantitative Evaluation Methods

- ***Text Analysis:*** A text analysis uses statistics and automation to draw inferences by looking at the number of responses containing a word or phrase, the respondents' grammar, or themes within responses. For example, a text analysis can be used to identify key emotional themes across customer satisfaction surveys.



Qualitative Evaluation

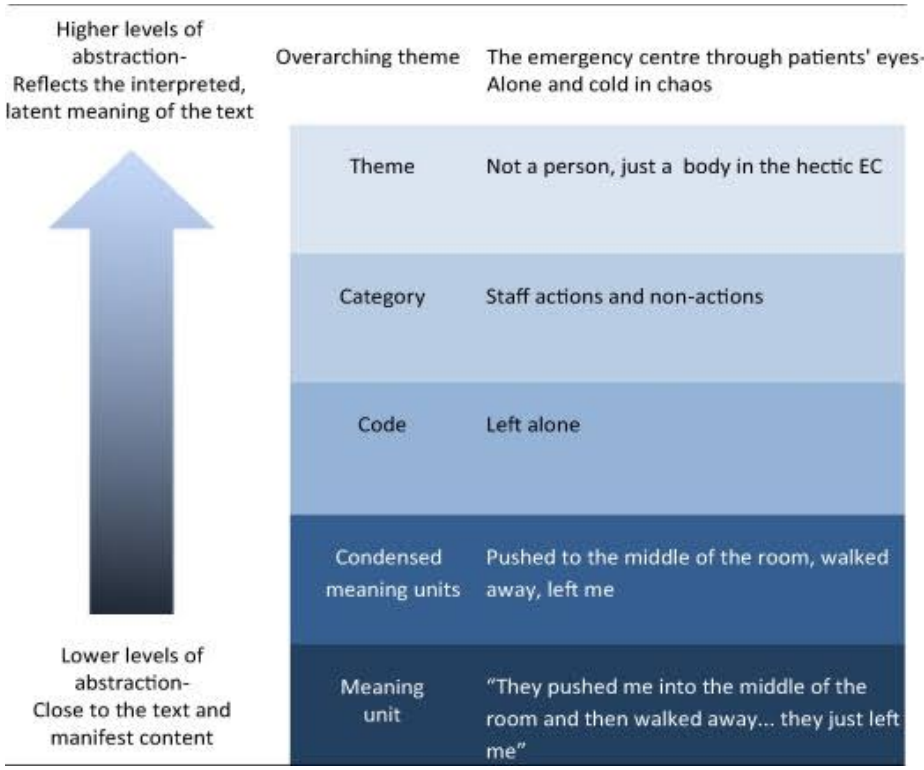
1. **Subjective Assessment:**
   - ***Focus:*** Qualitative evaluation emphasizes the subjective characteristics and the overall behavior of the model.
   - ***Metrics:*** It involves more subjective judgment, often captured through human interpretation.
1. **Contextual Understanding:**
   - ***Context Awareness:*** Qualitative evaluation is sensitive to the specific context and the nuances of the problem domain.
2. **Human Interpretation:**
   - ***User Feedback:*** Incorporates human feedback and interpretation, considering aspects like user experience and interpretability.

Qualitative Evaluation

4. **Insights and Patterns:**
   - ***Discoverability:*** Suitable for uncovering insights, patterns, or anomalies that might not be apparent through quantitative metrics alone.
5. **Examples:**
   - ***User Satisfaction:*** How satisfied users are with the system.
   - ***Interpretability:*** How easily the model's decisions can be understood by humans.
   - ***Robustness:*** How well the model performs in edge cases or under novel conditions.

Qualitative Evaluation Methods

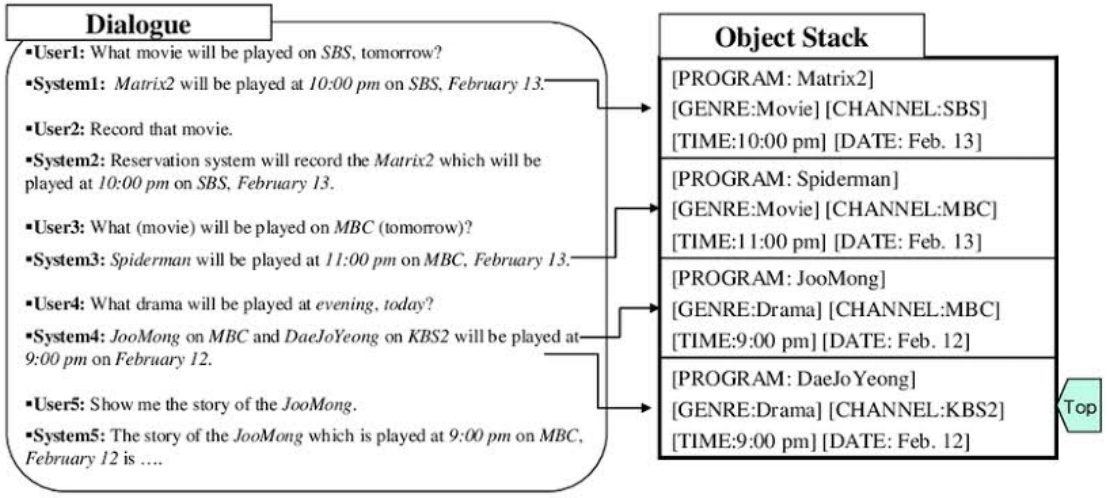
- ***Content analysis*** is a subjective interpretation of data that includes the following steps:
  - Preparing data
  - Defining the unit of analysis
  - Creating categories
  - Establishing a coding scheme
  - Testing the coding scheme
  - Coding the text
  - Reviewing for consistency
  - Drawing conclusions
  - Reporting findings



For example, content analysis can be used to find correlations and patterns across focus group answers to make decisions about product development.

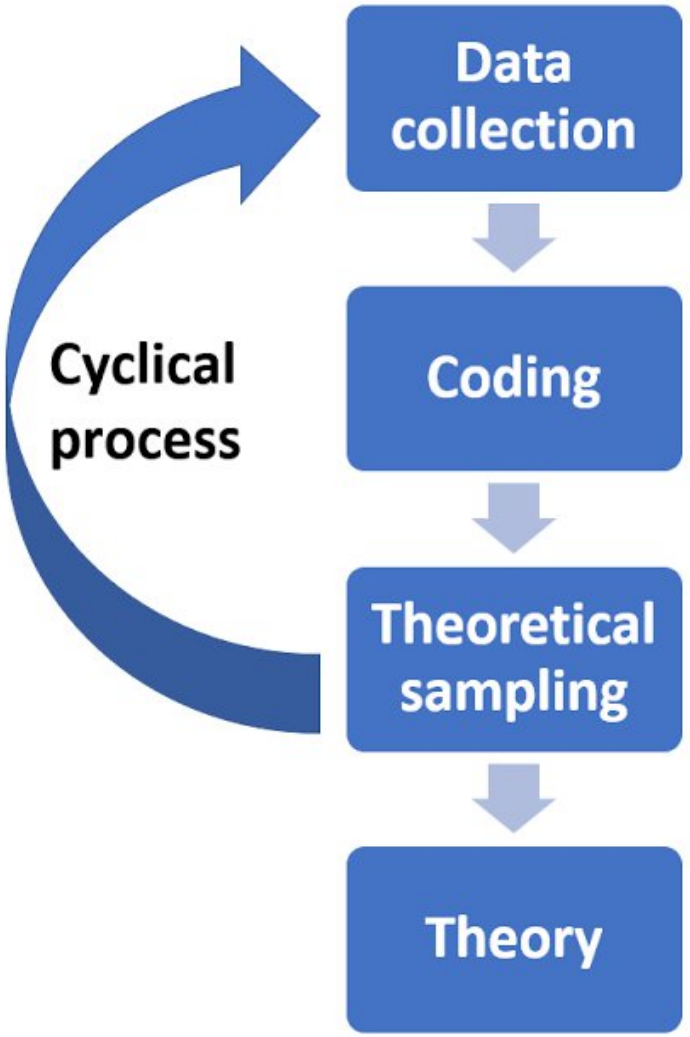
Qualitative Evaluation Methods

- ***Discourse***: Less systematic than content analysis, discourse analysis enables interpretation by exploring the meanings that language produces. This includes the details within the text and contextual knowledge about how people use language. For example, analysts use discourse analysis to understand how people communicate in interviews; this helps them obtain insights on how to write compelling marketing materials for a particular audience.



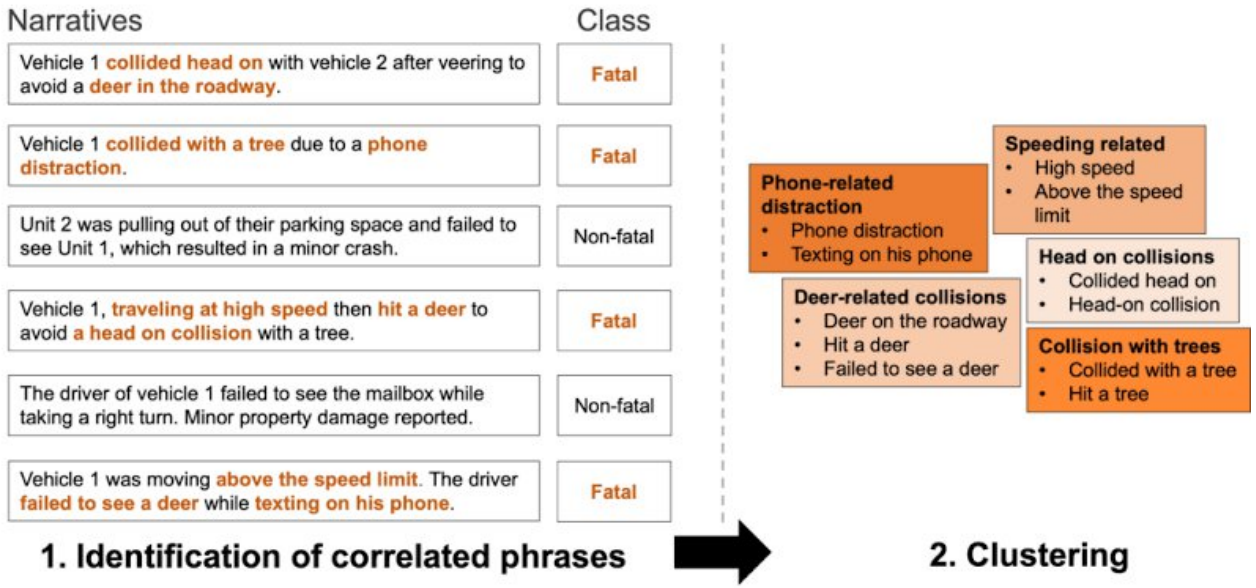
Qualitative Evaluation Methods

- ***Grounded Theory***: Grounded theory analysis uses transcripts of interviews to look for repeated themes by coding them with keywords and phrases to create a concept hierarchy. For example, grounded theory analysis can be used to correlate two different populations, like age and geographic demographics, to understand a new market.



Qualitative Evaluation Methods

- ***Narrative***: Narrative analysis revolves around the premise that stories are functional and purposeful. People use stories to organize their ideas and understand their lives. The four narrative analysis frameworks are:
  - Structural
  - Functional
  - Thematic
  - Dialogic/performance

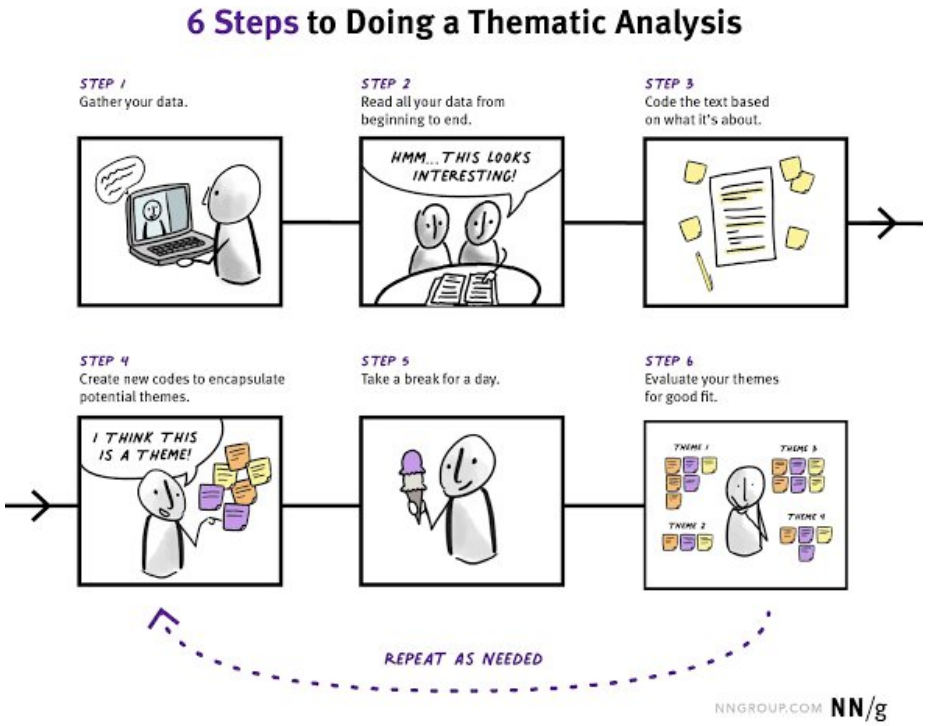


Qualitative Evaluation Methods

- ***Thematic*** : A thematic analysis examines themes or patterns in data. It requires less theoretical and technical knowledge, and as thus more accessible. The three thematic analysis types are:

  - Coding reliability
  - Codebook
  - Reflexive

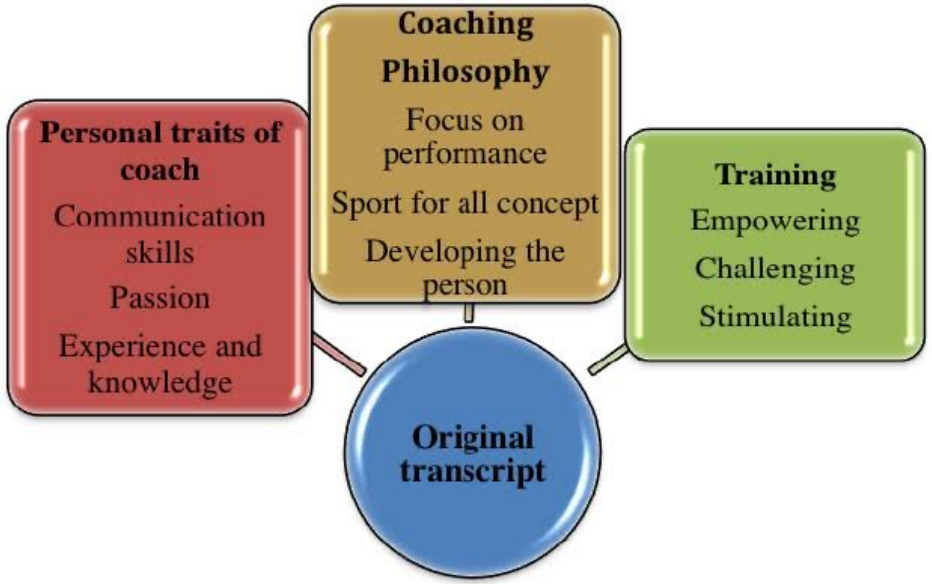
  For example, you can use thematic analysis of social media users to understand how a customer segment feels about a competitor.



Thematic Analysis

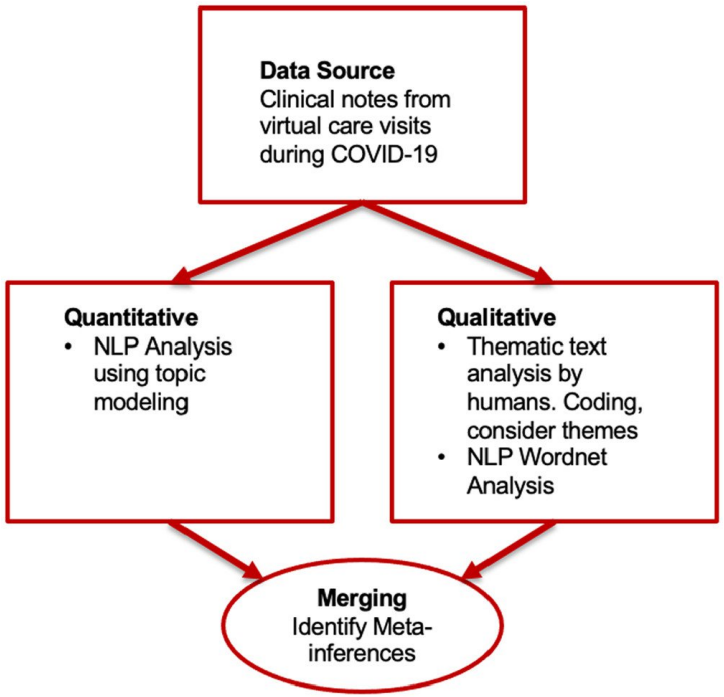
Interpretative Phenomenological Analysis (IPA)

An IPA explores people's responses in relation to their lived experience; it seeks insights into how someone would understand an event based on a given context. For example, IPA can be used for insights into how customers at a specific restaurant location felt about the service provided.

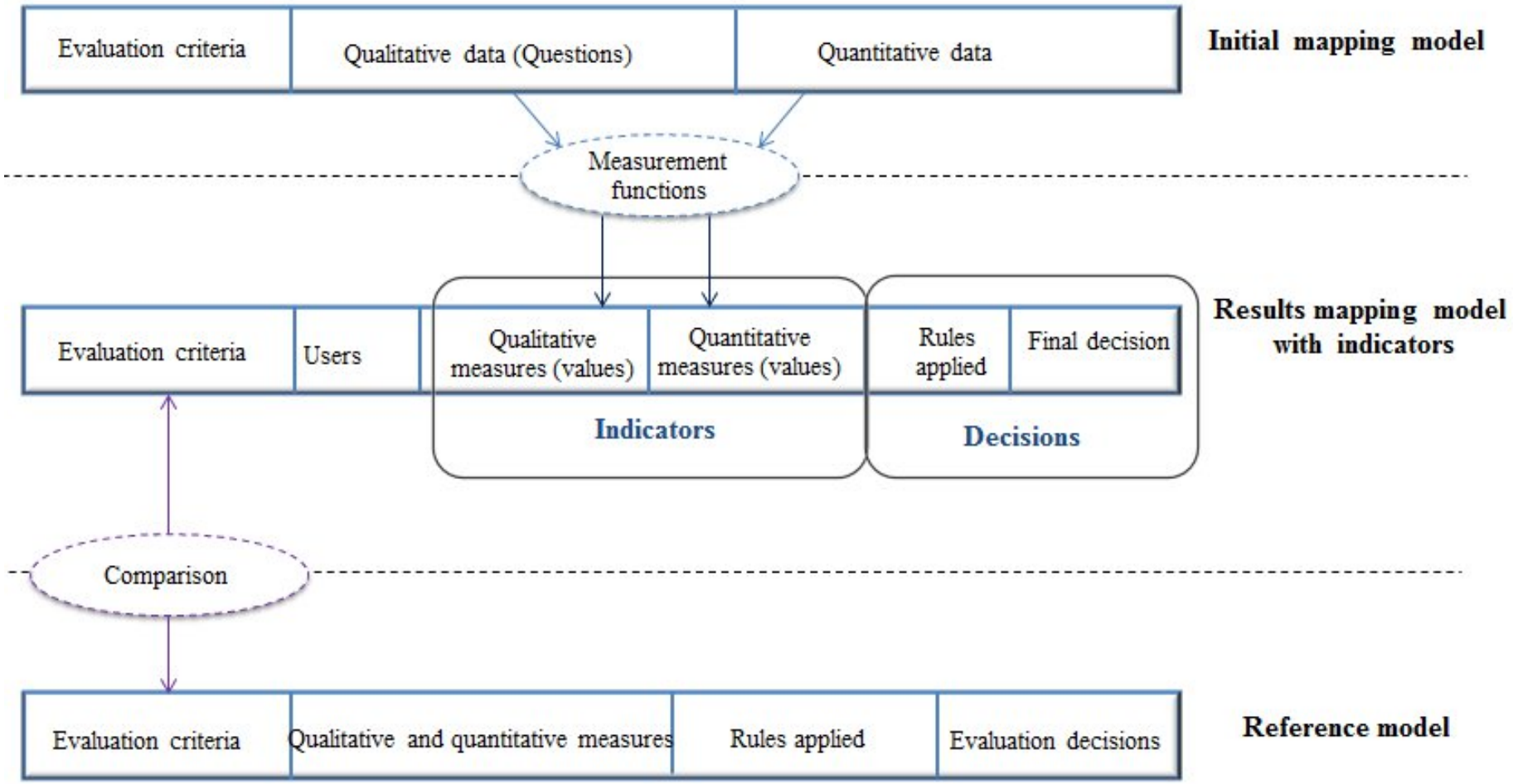


Integrated Evaluation

1. **Balanced Approach:**
   - ***Complementary:*** Often, a combination of both quantitative and qualitative evaluations provides a more comprehensive understanding of a model's performance.
1. **Iterative Process:**
   - ***Feedback Loop:*** Qualitative insights can guide improvements in the model, which can then be quantitatively measured for validation.
1. **Use Case Dependency:**
   - ***Application-Specific:*** The choice between quantitative and qualitative evaluation often depends on the specific goals and requirements of the application.

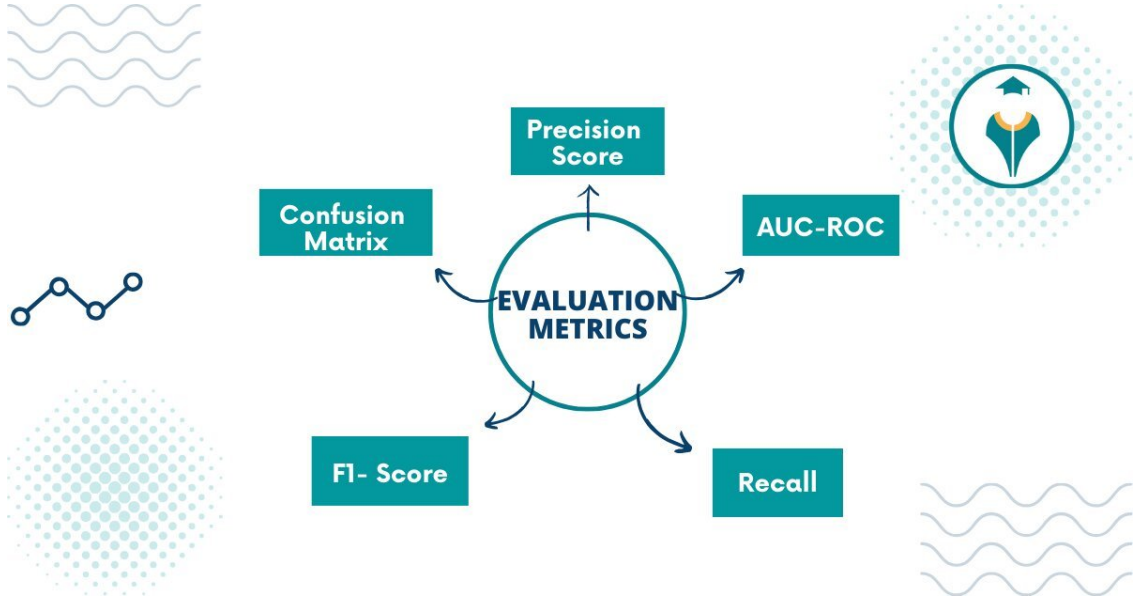


Integrated Evaluation



Evaluation Metrics

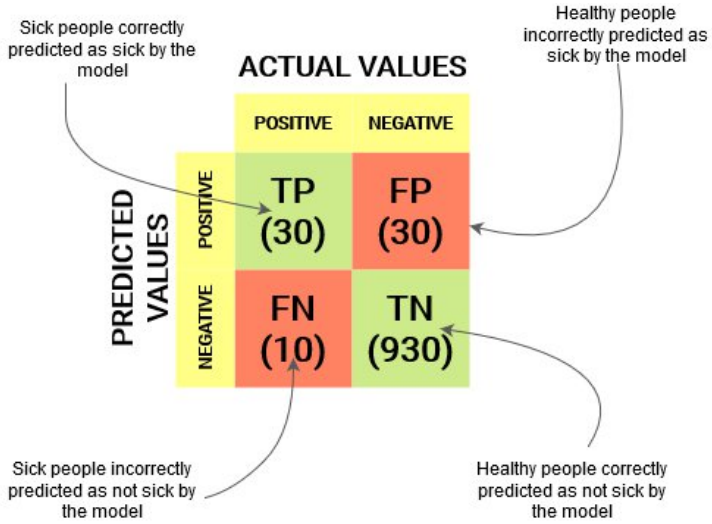
Evaluation is typically performed using metrics that reflect the accuracy or effectiveness of the model. These metrics may vary depending on the task and the specific goals of the evaluation. For example, accuracy, precision, recall, and F1-score are common metrics for evaluating text classification and information retrieval models, while BLEU and ROUGE are metrics used in machine translation evaluation, Perplexity and WER metrics is used for Automatic speech recognition and text generation.



Confusion Matrix

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is a means of displaying the number of accurate and inaccurate instances based on the model's predictions. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance. The matrix displays the number of instances produced by the model on the test data.

- ***True positives (TP):*** occur when the model accurately predicts a positive data point.
- ***True negatives (TN):*** occur when the model accurately predicts a negative data point.
- ***True positives (FP):*** occur when the model predicts a positive data point incorrectly.
- ***False negatives (FN):*** occur when the model mis predicts a negative data point.



Evaluation Metrics

- ***Accuracy:*** Accuracy is the proportion of correctly classified instances out of the total number of instances. In NLP tasks, accuracy is often used to measure the overall performance of a model.

$$\textbf{\textit{Accuracy}} = \frac{TP + TN}{TP + TN + FP + FN}$$

- ***Precision:*** Precision is the proportion of true positives (correctly identified instances) to the total number of instances identified as positive. In NLP tasks, precision is used to measure how many of the instances identified as positive are actually positive.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Evaluation Metrics

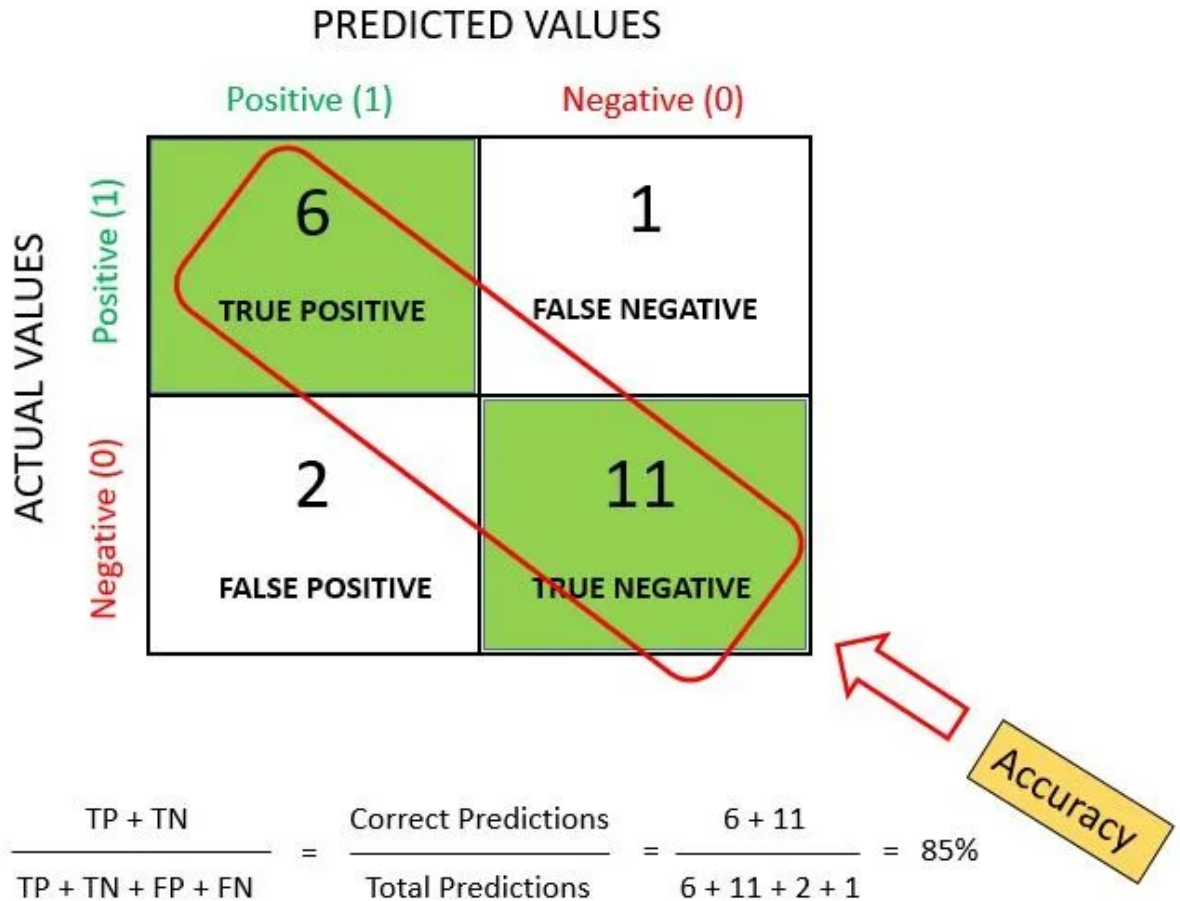
- ***Recall:*** Recall is the proportion of true positives to the total number of instances that are actually positive. In NLP tasks, recall is used to measure how many of the positive instances were correctly identified by the model.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

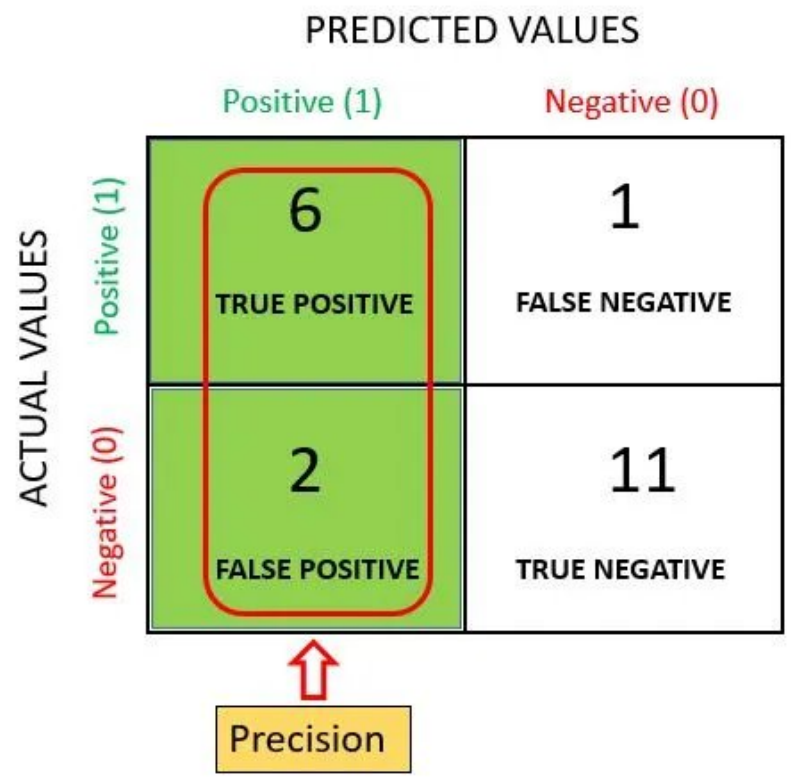
- ***F1-score:*** F1-score is the harmonic mean of precision and recall, and is used to balance the trade-off between precision and recall. It ranges from 0 to 1, with 1 being the best possible score. In NLP tasks, F1-score is often used to evaluate the overall performance of a model.

$$F_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Example



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} = \frac{6 + 11}{6 + 11 + 2 + 1} = 85\%$$



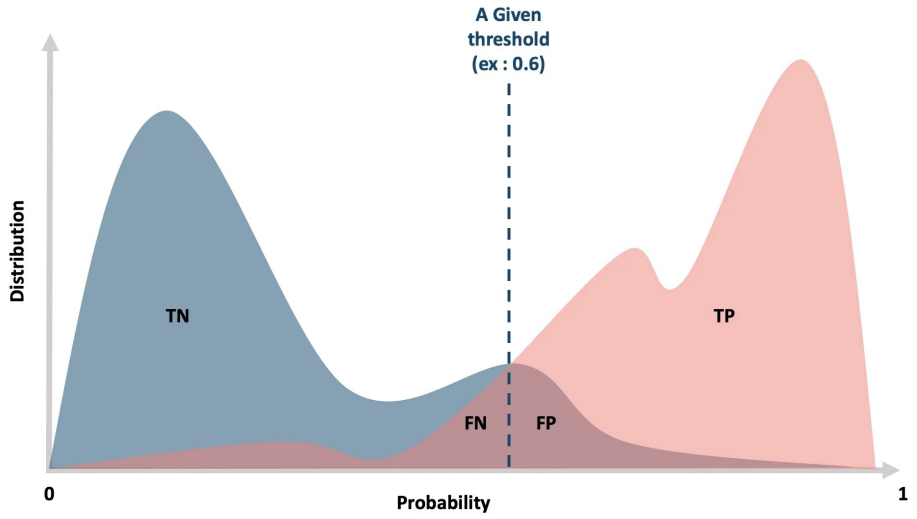
$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{Predictions Actually Positive}}{\text{Total Predicted positive}} = \frac{6}{6 + 2} = 0.75$$

ROC (Receiver Operating Characteristic)

ROC is one of the most important evaluation metrics for checking any classification model's performance. It's plotted with two metrics against each other. TPR (True Positive Rate or Recall) and FPR (False Positive Rate) where the former is on y-axis and the latter is on x-axis.

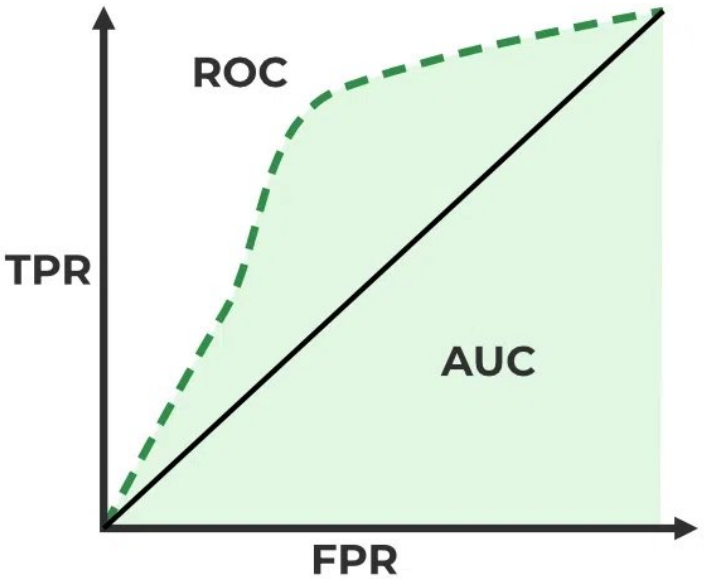
- ***TPR:*** is the recall which is, out of all positive cases, how many we predicted correctly.
- ***FPR:*** out of all negatives cases how many we didn't predict correctly.

$$TPR = \frac{TP}{TP + FN} \quad \text{and} \quad FPR = \frac{FP}{FP + TN}$$



AUC (Area Under Curve)

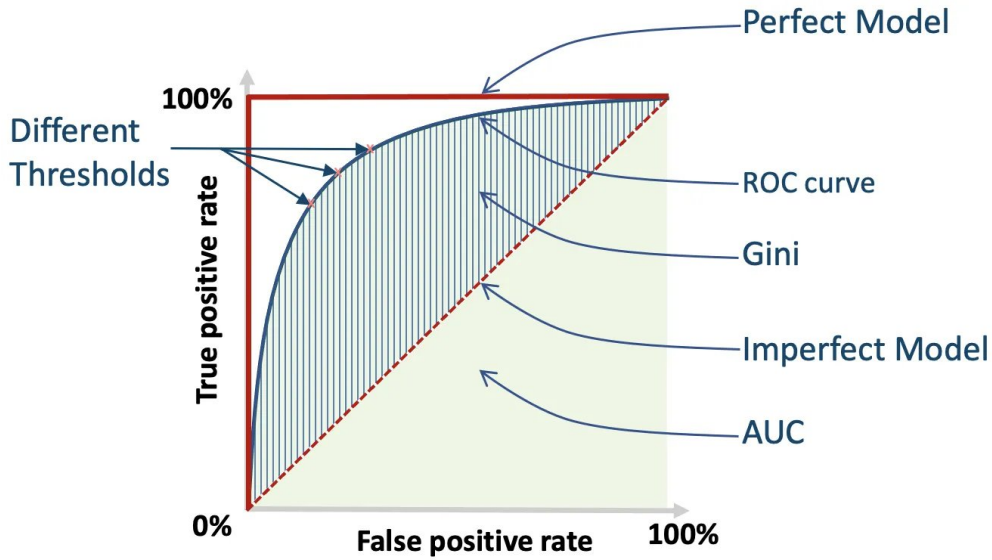
The ROC curve on its own is not a metric to compute because it's just a curve where we want to see in every threshold TPR and FPR metrics against each other. So, to quantify this curve and compare two models we need a more explicit metric. AUC measures the entire two-dimensional area underneath the ROC curve.

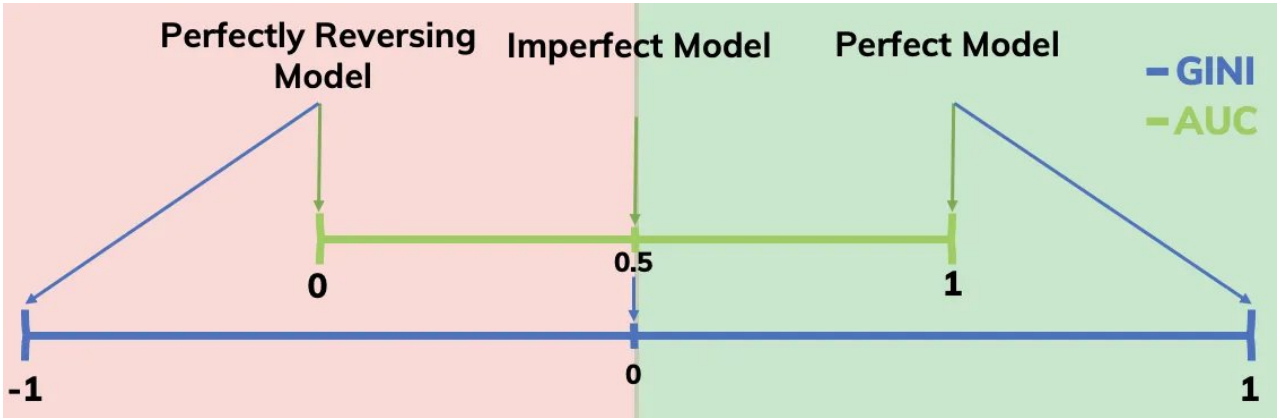


AUC tells how much our model, regardless of our chosen threshold, is able to distinguish between the two classes. The higher it is the better the model is. It has a value between 1 and 0.

GINI Index

The Gini index or coefficient is a way to adjust the AUC so that it can be clearer and more meaningful. It's more natural for us to see a perfectly random model having 0, reversing models with a negative sign and the perfect model having 1. The range of values now is [-1, 1].





GINI Index

- **Perfectly reversing model :** This model is doing the exact opposite of a perfect model. It's predicting every positive observation as a negative one and vice-versa. This means if we invert all the outputs we'll have a perfect model. It has a Gini=-1 and AUC=0.

- **Imperfect model:** It means this model has no discrimination ability to distinguish between the two classes. It's a perfectly random model. It has a Gini=0 and AUC=0.5.

- **Perfect model:** The perfect model is the model that predicts every observation correctly for positive and negative classes. It means in every threshold at least one of FPR and TPR is equal to zero. This model has an AUC=1 and a Gini=1.

MRR (Mean Reciprocal Rank)

- Mean Reciprocal Rank (MRR) is a ranking quality metric. It considers the position of the first relevant item in the ranked list.
- A Reciprocal Rank is the inverse of the position of the first relevant item. If the first relevant item is in position 2, the reciprocal rank is 1/2.
- MRR values range from 0 to 1, where "1" indicates that the first relevant item is always at the top.
- Higher MRR means better system performance.

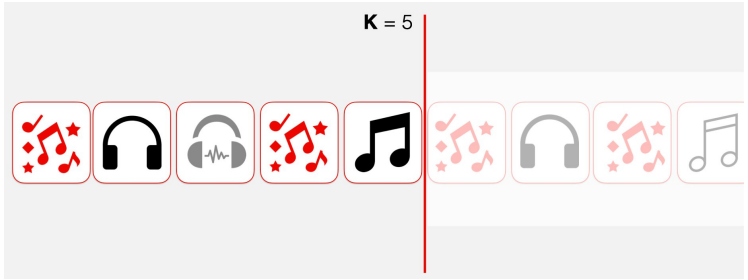
$$\text{MRR} = \frac{1}{U}\sum_{u=1}^{U}\frac{1}{rank_i}$$

where,

- ***U*** is the total number of users (in case of recommendations) or queries (in case of information retrieval) in the evaluated dataset.
- ***Rank i*** is the position of the first relevant item for user u in the top-K results.

Computing MRR

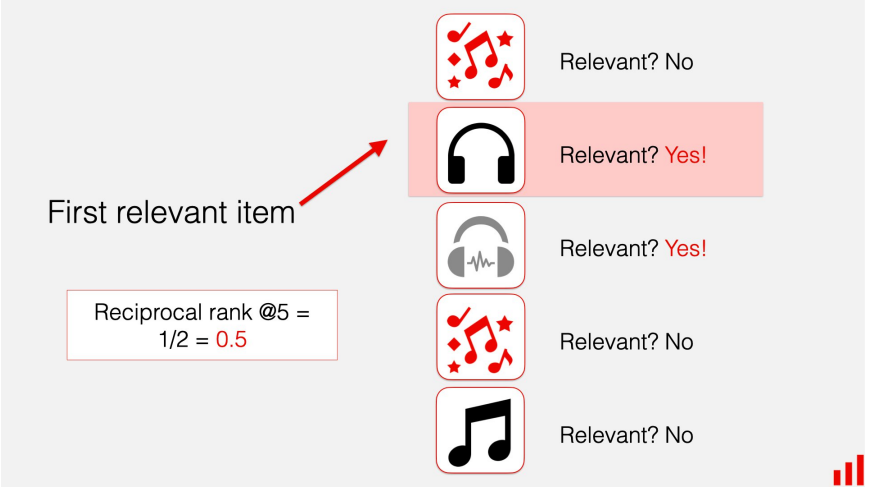
1. To compute MRR, you need to prepare the dataset and decide on the K parameter, which is the number of top-ranked items you will use in your evaluation.



1. Identify relevant items

Computing MRR

3. Identify the first relevant rank

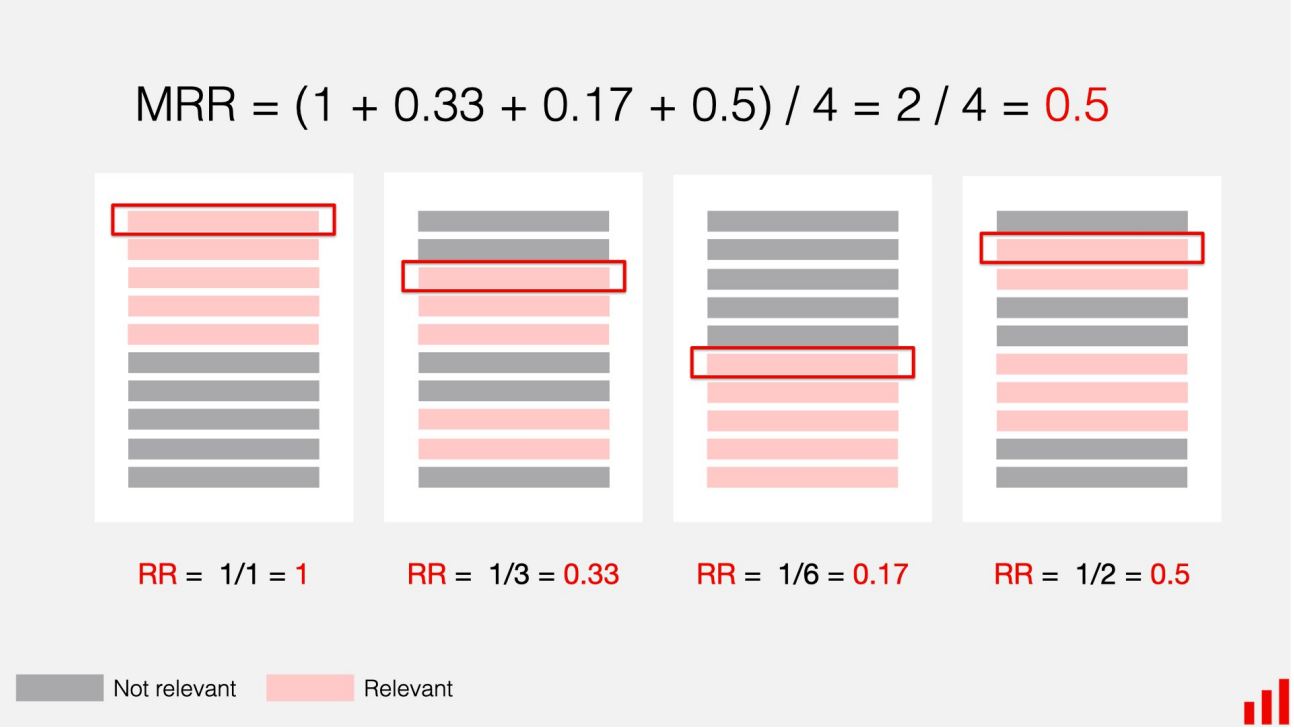


3. Calculate the Reciprocal Rank

$$RR = \frac{1}{\text{rank of the first correct result}}$$

Computing MRR

5. Compute the MRR



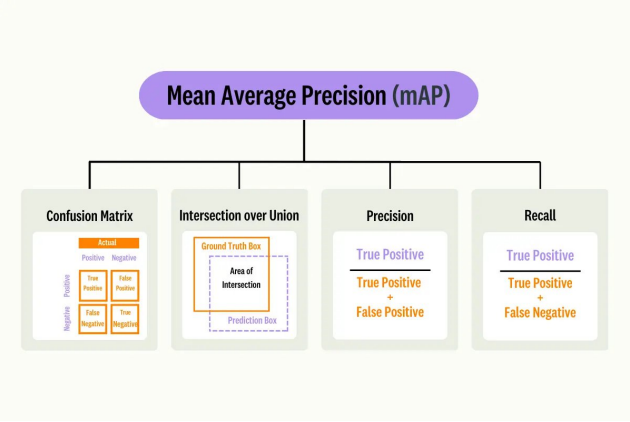
MRR can take values from 0 to 1.

- MRR equals 1 when the first recommendation is always relevant.
- MRR equals 0 when there are no relevant recommendations in top-K.
- MRR can be between 0 and 1 in all other cases. The higher the MRR, the better.

What is a good MRR? This depends on the use case. For example, if you have a recommender system that suggests a set of five items out of many thousand possibilities, an MRR of 0.2 might be acceptable. This indicates that, on average, users find a relevant item at position 5.

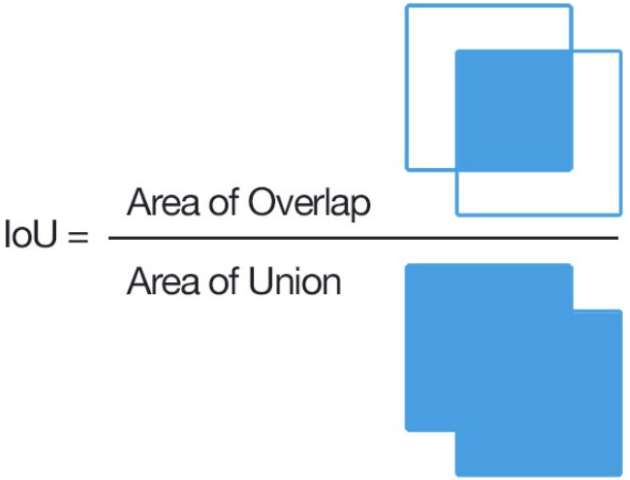
MAP (Mean Average Precision)

Mean average precision, calculated across each retrieved result. It is a metric used to evaluate object detection models such as Fast R-CNN, YOLO, Mask R-CNN, etc. The mean of average precision(AP) values are calculated over recall values from 0 to 1.



- ***Intersection over Union (IoU)***

  Intersection over Union indicates the overlap of the predicted bounding box coordinates to the ground truth box. Higher IoU indicates the predicted bounding box coordinates closely resembles the ground truth box coordinates.



Calculating MAP

Mean Average Precision is the average of AP of each class. Here is a summary of the steps to calculate the AP:

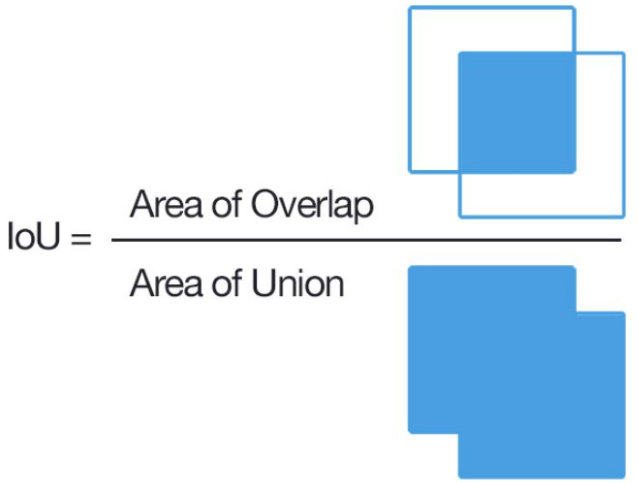
- Generate the prediction scores using the model.
- Convert the prediction scores to class labels.
- Calculate the confusion matrix—TP, FP, TN, FN.
- Calculate the precision and recall metrics.
- Calculate the area under the precision-recall curve.
- Measure the average precision.

The mAP is calculated by finding Average Precision(AP) for each class and then average over a number of classes.

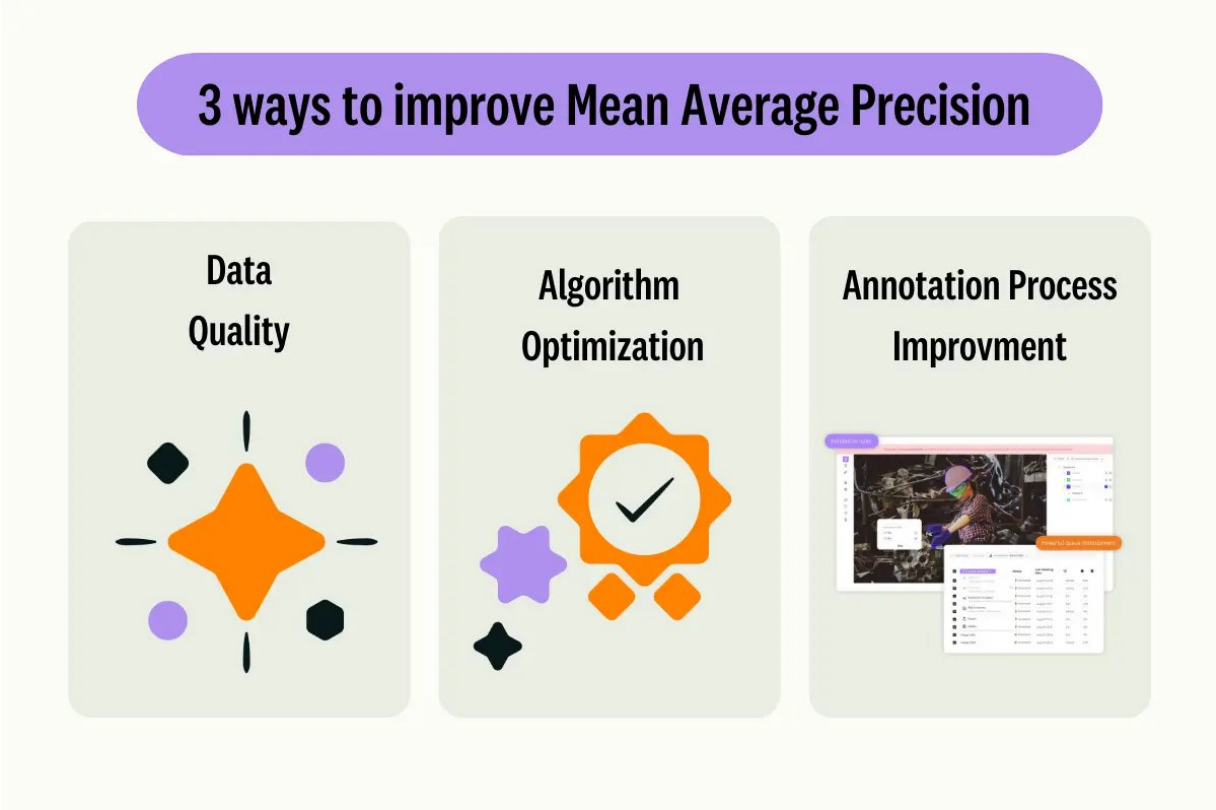
$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \textbf{the AP of class } k$$

$$n = \textbf{the number of classes}$$



Improving MAP



1. *Data Quality*

Increasing the quality of your training data is imperative to a machine learning model's performance. Quality data means data that it is representative of the data that will be found when the model is deployed in production: the image attributes should be similar (brightness, contrast, zoom level etc.), should contain the same background elements, and all the

Improving MAP

2. ***Optimizing the object detection algorithm***

The state-of-the-art object detection algorithms, such as Convolutional Neural Networks Fast R-CNN, and YOLO (You Only Look Once) are becoming more and more popular and keep improving.

If you primarily focus on working with real-time object detection, you will typically be using YOLO-type algorithms, as shown in this Real-Time Object Detection leaderboard. The TOP 3 models are:

- 1.YOLOv7-E6E(1280)
- 1.YOLO
- 1.YOLOX

Improving MAP

3. *Improving the annotation process*

Data annotations are typically manual tasks that become tedious over time. Especially when the dataset becomes more complex and large, there's a lot of room for error. To prevent this from happening, you can follow these strategies:

- Ensure your annotation instructions are user-friendly but comprehensive.
- Ensure that the annotators have been quality-screened.
- Add a review and evaluation stage to ensure the benchmark is met.

RMSE (Root Mean Squared Error)

Root mean square error or root mean square deviation shows how far predictions fall from measured true values using Euclidean distance.

To compute RMSE, calculate the residual (difference between prediction and truth) for each data point, compute the norm of residual for each data point, compute the mean of residuals and take the square root of that mean.

Root mean square error can be expressed as

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}}$$

RMSD = root-mean-square deviation

i = variable i

N = number of non-missing data points

x_i = actual observations time series

$\hat{x}_i$ = estimated time series

MAPE (Mean Absolute Percentage Error)

Mean absolute percentage error is an evaluation metric often used as the loss function in regression problems and forecasting models due to the intuitive interpretation in terms of relative error for evaluation. Also known as the mean absolute percentage deviation (MAPD), MAPE is defined as the average absolute percentage difference between predicted values and actual values.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i}$$

A_i is the actual value

F_i is the forecast value

n is total number of observations

BLEU (Bilingual Evaluation Understudy)

Captures the amount of n-gram overlap between the output sentence and the reference ground truth sentence. Has many variants, and mainly used in machine translation tasks. Has also been adapted to text to text tasks such as paraphrase generation and summarization.

Target Sentence: *The guard arrived late because it was raining*

Predicted Sentence: *The guard arrived late because of the rain*

The first step is to compute Precision scores for 1-grams through 4-grams.

- ***Precision 1-gram***: We use the Clipped Precision method.

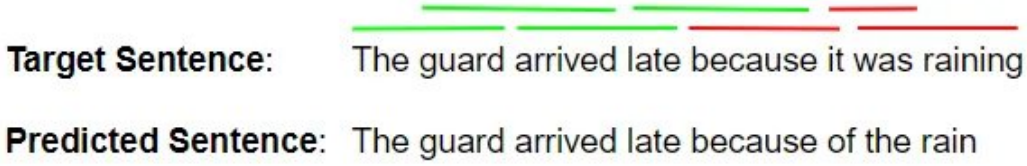
  Precision 1-gram = Number of correct predicted 1-grams / Number of total predicted 1-grams

So, Precision 1-gram **(p_1) = 5 / 8**

BLEU (Bilingual Evaluation Understudy)

- **Precision 2-gram:**

  Precision 2-gram = Number of correct predicted 2-grams / Number of total predicted 2-grams



So, Precision 2-gram **(p_2) = 4 / 7**

- **Precision 3-gram**

  Similarly, Precision 3-gram **(p_3) = 3 / 6**

BLEU (Bilingual Evaluation Understudy)

- **Precision 4-gram**

  And, Precision 4-gram (p_4) = 2 / 5

- **Geometric Average Precision Scores**

  Next, we combine these Precision Scores using the formula below. This can be computed for different values of N and using different weight values. Typically, we use N = 4 and uniform weights w_n = N / 4

$$
\begin{aligned}
Geometric\ Average\ Precision\ (N) &= exp(\sum_{n=1}^{N} w_n\ log\ p_n) \\
&= \prod_{n=1}^{N} p_n^{w_n} \\
&= (p_1)^{\frac{1}{4}} \cdot (p_2)^{\frac{1}{4}} \cdot (p_3)^{\frac{1}{4}} \cdot (p_4)^{\frac{1}{4}}
\end{aligned}
$$

BLEU (Bilingual Evaluation Understudy)

- **Brevity Penalty**

The third step is to compute a 'Brevity Penalty'. If you notice how Precision is calculated, we could have output a predicted sentence consisting of a single word like "The' or "late".

For this, the 1-gram Precision would have been 1/1 = 1, indicating a perfect score. This is obviously misleading because it encourages the model to output fewer words and get a high score.

To offset this, the Brevity Penalty penalizes sentences that are too short.

$$Brevity\ Penalty = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c <= r \end{cases}$$

c is predicted length = number of words in the predicted sentence and
r is target length = number of words in the target sentence

In this example, c = 8 and r = 8, which means Brevity Penalty = 1

BLEU (Bilingual Evaluation Understudy)

- **Bleu Score**

Finally, to calculate the Bleu Score, we multiply the Brevity Penalty with the Geometric Average of the Precision Scores.

$$Bleu\ (N) = Brevity\ Penalty \cdot Geometric\ Average\ Precision\ Scores\ (N)$$

Bleu Score can be computed for different values of N. Typically, we use N = 4.

BLEU-1 uses the unigram Precision score

BLEU-2 uses the geometric average of unigram and bigram precision

BLEU-3 uses the geometric average of unigram, bigram, and trigram precision

and so on.

$$\begin{aligned} log\ Bleu &= min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{4} \frac{log\ p_n}{4} \\ &= min(1 - \frac{r}{c}, 0) + \frac{log\ p_1 + log\ p_2 + log\ p_3 + log\ p_4}{4} \end{aligned}$$

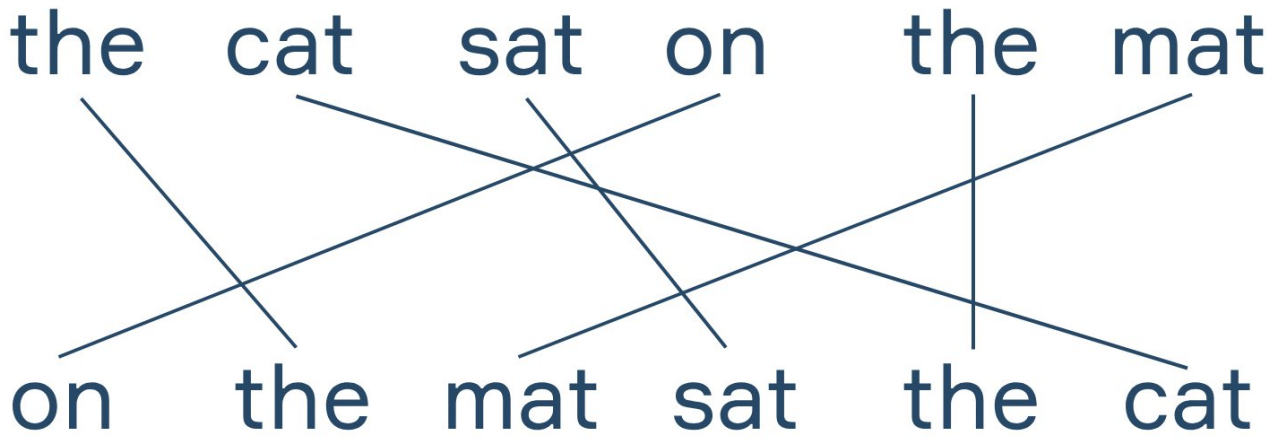
METEOR: Precision based metric

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a precision-based metric to measure quality of generated text.

F-score: METEOR uses a weighted F-score (with recall being weighted higher than precision) based on mapped unigrams and a penalty function (fragmentation measure) for incorrect word order.

- First, you search for the largest subset of mappings (called alignment) between the candidate and the reference, so exact matches are considered first, followed by matches after Porter stemming, and then WordNet synonyms.
- Alignment is a mapping between unigrams, such that every unigram in each string maps to zero or one unigram in the other string, and to no unigrams in the same string.

METEOR: Precision based metric



After the largest alignment is found, let m be the number of mapped unigrams between the reference and the candidate. Calculate the weighted F-Score (with α controlling the relative weight for precision and recall):

$$F_{\text{mean}} = \frac{\text{PR}}{\alpha\text{P} + (1-\alpha)\text{R}}$$

METEOR: Precision based metric

- **Penalty function:** The penalty function is responsible for the correct word order in the candidate and is defined as follows:

$$\text{Penalty} = \gamma\left(\frac{c_m}{m}\right)^{\beta},$$

  - c_m – the number of matching chunks (a **chunk** is the set of unigrams that are positioned next to each other in the reference and in the candidate),
  - m – the number of mapped unigrams between the reference and the candidate,
  - β – a parameter responsible for the penalty's shape, and
  - γ – the relative weight for the fragmentation penalty, $\gamma \in [0, 1]$.

The penalty decreases with the lower number of chunks present (the longest matches are rewarded, while the more fragmented matches are penalized).

METEOR: Precision based metric

METEOR uses a weighted F-score (with recall being weighted higher than precision) based on mapped unigrams and a penalty function (fragmentation measure) for incorrect word order, and is calculated as:

The Meteor score is then calculated:

$$Score = (1 - Pen) \cdot F_{mean}$$

METEOR lies in the [0, 1] range, with higher values indicating better scores.

ROUGE

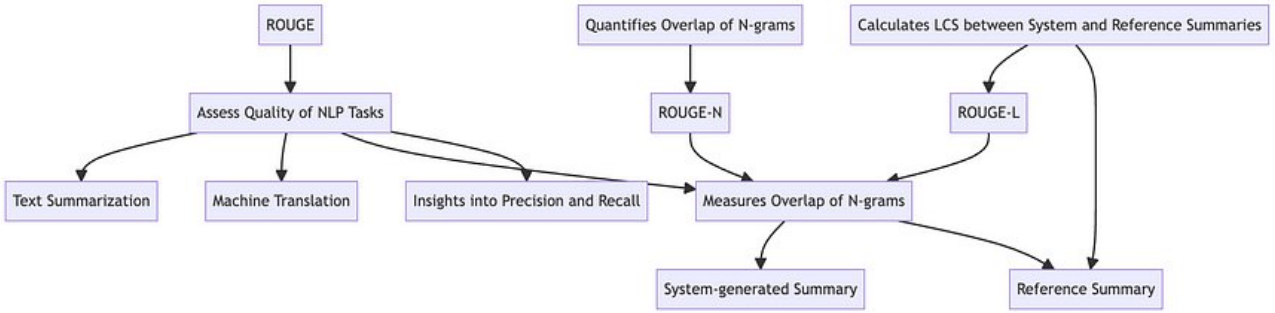
Like BLEU and METEOR, compares quality of generated to reference text. Measures recall. Mainly used for summarization tasks where it's important to evaluate how many words a model can recall (recall = % of true positives versus both true and false positives).

ROUGE-N measures the number of matching n-grams between the model-generated text and a human-produced reference.

Consider the reference R and the candidate summary C:

R: The cat is on the mat.

C: The cat and the dog.



ROUGE

ROUGE-1

Using R and C, we are going to compute the precision, recall, and F1-score of the matching n-grams. Let's start computing ROUGE-1 by considering 1-grams only.

- ROUGE-1 precision can be computed as the ratio of the number of unigrams in C that appear also in R (that are the words "the", "cat", and "the"), over the number of unigrams in C.
  - ROUGE-1 precision = 3/5 = 0.6
- ROUGE-1 recall can be computed as the ratio of the number of unigrams in R that appear also in C (that are the words "the", "cat", and "the"), over the number of unigrams in R.
  - ROUGE-1 recall = 3/6 = 0.5
- Then, ROUGE-1 F1-score can be directly obtained from the ROUGE-1 precision and recall using the standard F1-score formula.
  - ROUGE-1 F1-score = 2 * (precision * recall) / (precision + recall) = 0.54

ROUGE

ROUGE-L

ROUGE-L is based on the longest common subsequence (LCS) between our model output and reference, i.e. the longest sequence of words (not necessarily consecutive, but still in order) that is shared between both.

We can compute ROUGE-L recall, precision, and F1-score just like we did with ROUGE-N, but this time we replace each n-gram match with the LCS.

Remember our reference R and candidate summary C:

R: The cat is on the mat.

C: The cat and the dog.

The LCS is the 3-gram “the cat the” (remember that the words are not necessarily consecutive), which appears in both R and C.

ROUGE

- ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in C.
  - ROUGE-L precision = 3/5 = 0.6
- ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in R.
  - ROUGE-L recall = 3/6 = 0.5
- Therefore, the F1-score is:
  - ROUGE-L F1-score = 2 * (precision * recall) / (precision + recall) = 0.55

ROUGE

ROUGE-S

ROUGE-S allows us to add a degree of leniency to the n-gram matching performed with ROUGE-N and ROUGE-L.

ROUGE-S is a skip-gram concurrence metric: this allows to search for consecutive words from the reference text that appear in the model output but are separated by one-or-more other words.

If we consider the 2-gram “the cat”, the ROUGE-2 metric would match it only if it appears in C exactly, but this is not the case since C contains “the gray cat”. However, using ROUGE-S with unigram skipping, “the cat” would match “the gray cat” too.

Perplexity

Measures how confused an NLP model is, derived from cross-entropy in a next word prediction task. Used to evaluate language models, and in language-generation tasks, such as dialog generation.

Perplexity is calculated as exponent of the loss obtained from the model.The formula for perplexity is the exponent of mean of log likelihood of all the words in an input sequence.

$$\mathrm{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p_\theta(x_i|x_{<i})\right\}$$

Log Likelihood

Sentiments probability calculation requires the multiplication of many numbers with values between 0 and 1. Carrying out such multiplications on a computer runs the risk of numerical underflow when the number returned is so small if can't be stored on your device.

The trick is to use a log of the score instead of the raw score. This allows you to write the previous expression as the sum of the log prior and the log likelihood, which is a sum of the logarithms of the conditional probability ratio of all unique words in your corpus.

Log Likelihood

$$\frac{P(pos)}{P(neg)}\prod_{i=1}^{m}\frac{P(w_i|pos)}{P(w_i|neg)}$$

- Products bring risk of underflow
- $log(a*b) = log(a) + log(b)$
- $log(\frac{P(pos)}{P(neg)}\prod_{i=1}^{n}\frac{P(w_i|pos)}{P(w_i|neg)}) \Rightarrow log\frac{P(pos)}{P(neg)} + \sum_{i=1}^{n} log\frac{P(w_i|pos)}{P(w_i|neg)}$

log prior + log likelihood

Conducting Human Evaluation

- **Crowdsourcing**, which involves hiring online workers to rate, rank, or select the best output among several options, is a fast and cheap way to collect data but may suffer from low quality, inconsistency, and bias.
- **Expert review** is a more reliable and rigorous method of evaluation but is more expensive and time-consuming.
- **User testing** involves recruiting real or potential users of the NLP system to use it in a natural or simulated setting and collecting their feedback.

Human Evaluation Challenges

- Quality control and reliability must be taken into account, as evaluator skills, knowledge, motivation, and attention may vary.
- The results of human evaluation can be complex, noisy, or contradictory due to subjective opinions. Thus, statistical and qualitative methods must be used to analyze and interpret the results and compare them with automatic metrics.
- Quality control measures like screening, training, monitoring, and filtering of evaluators should be implemented to ensure reliability. Additionally, inter-rater agreement and reliability scores should be calculated.