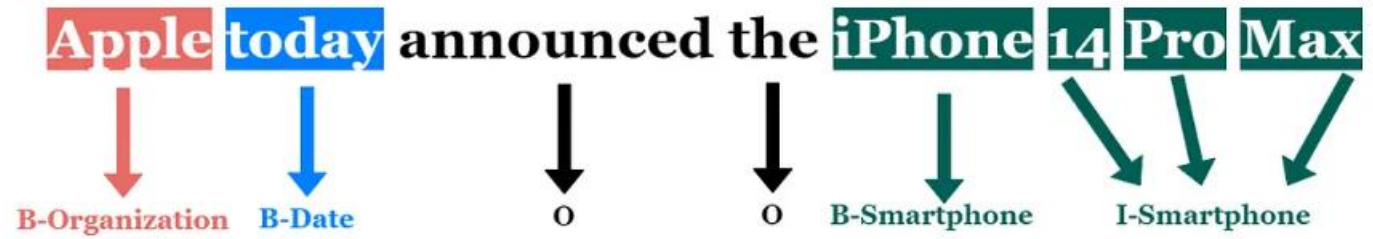
GLOBAL METHODOLOGY DOCUMENT (GMD)

FINETUNING NER MODEL

COLLECT AND ANNOTATE DATA

- financial reports with structured data
- structured financial data
- To annotate data:
  i. Prodigy (commercial)
  ii. spaCy's doccano (open-source)
  iii. Label Studio (open-source)
  iv. Use LLM

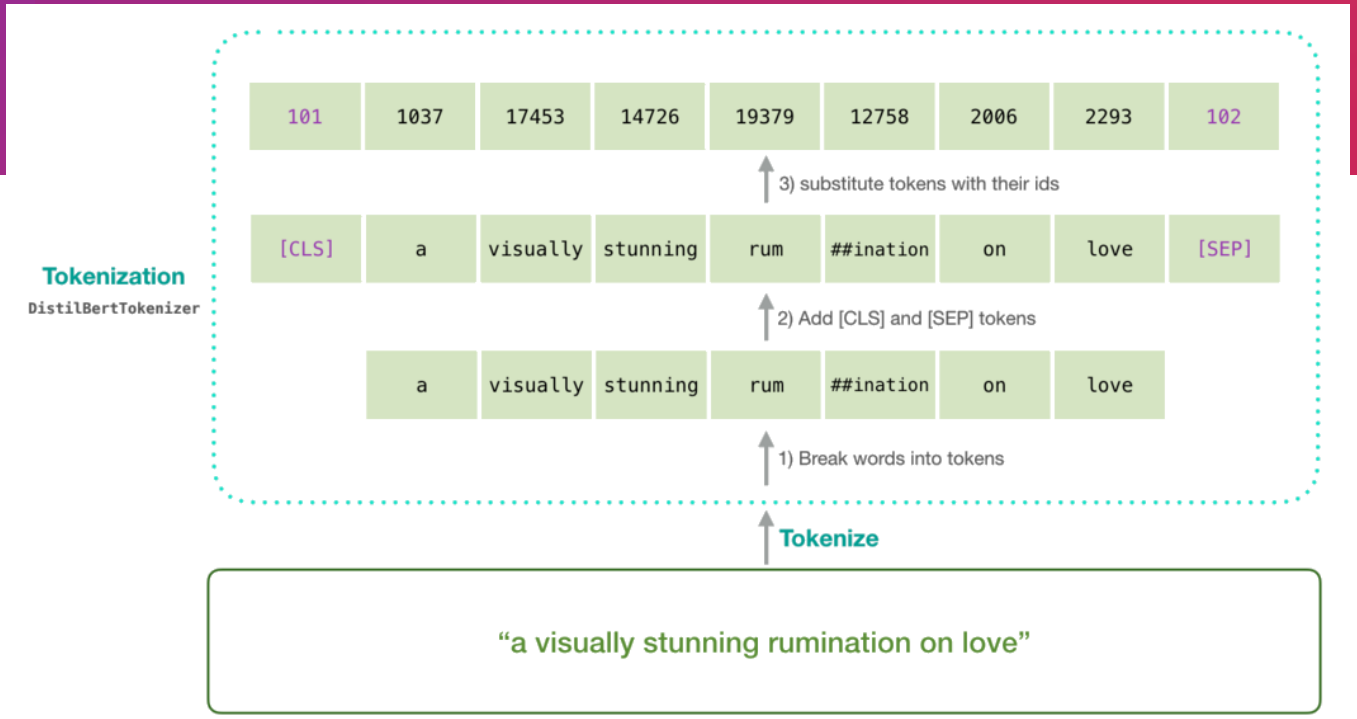
```
I'll revert regarding BANK ABC to try to do another 200 mio at 2Y
0                    B-ORG             O  O  O  O    O  O O  B-MONEY O O B-DURATION

FR001400QV82 AVMAFC FLOAT 06/30/28
B-ISIN       O      B-INSTRUMENT B-DATE

offer 2Y EVG estr+45bps
O     B-DURATION B-RATE B-RATE
```

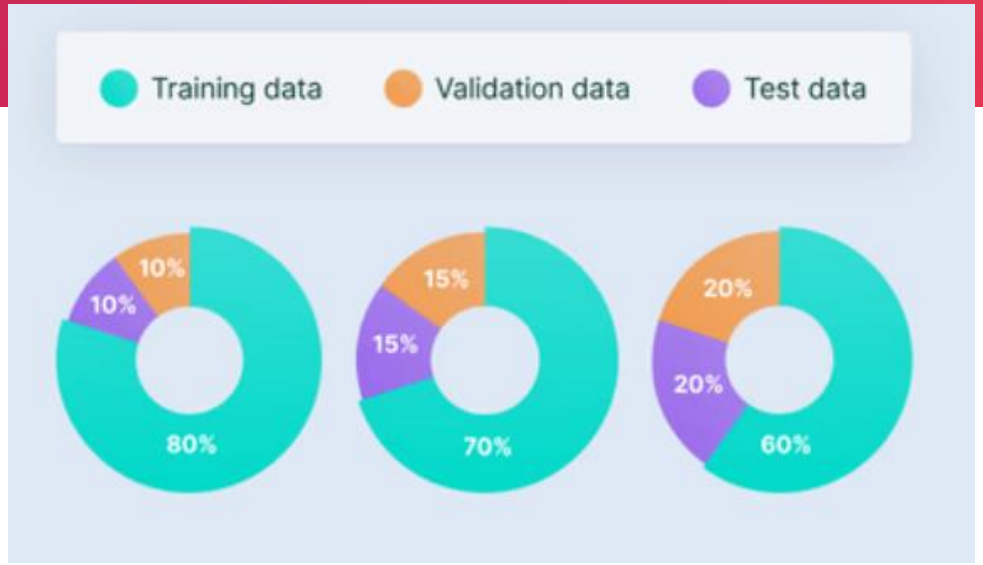


- `bert-base-uncased` (general-purpose)
- `roberta-base` (more contextualized)
- `flair/ner-english-ontonotes` (pre-trained on general NER)
- `xlm-roberta-base` (for multilingual finance)
- `FinBERT` (if you're working with finance-specific language)

CHOOSE A BASE MODEL

PREPROCESS AND TOKENIZE DATA





- **Tokenize using WordPiece (BERT) or Byte-Pair Encoding (RoBERTa)**
- **Split dataset into train, validation, and test sets (e.g., 80%/10%/10%)**

FINE-TUNE THE MODEL USING HUGGING FACE'S TRAINER API

1. Load Pre-trained Model & Tokenizer
2. Load and Prepare Dataset
3. Tokenize & Align Labels
4. Define Training Arguments
5. Define Trainer and Train Model
   Use Hugging Face's Trainer API for training.
6. Evaluate Model Performance
   Use SeqEval to compute F1-score.