for more accurate
repersentation of original
words
⇓

LLM terms cont:

- Lemmatization: a text normalization technique in NLP where we reduce words to thier dictionary form known as Lemma. insted of chopping suffix/prefix it considers the context of the whole word and applies Morphological analysis to find the base dictonary form

Ex: running → run, ate → eat, children → child, worst → bad

- Stemming: a text normalization technique in NLP to reduce words to base or root form known as "stem" this is done by removing suffixes Like ("ing", "-ed", "s") from words the goal is to treat words with simallar meaning as the same (for efficiency in NLP tasks)

Ex: running → runn, happily → happi, cars → car

- Semi Supervised learning: a type of ML where training data has both Labeled and unlabeled data. the goal is to leverage the information from the unlabeled data to improve the preformance of a model that would otherwise be trained solely on labeled data. Good when labed data is hard to find but unlabeled data is redily avalible

Self supervised learning: a type of ML where the model learns from unlabeled data by creating its own supervisory signals this is achived by masking parts of the input data and training the model to predict the masked portion based on the remaning data in essence the data itself provides the labels allowing model to learn without human anotation for data.

1) Self training: for classification + regression, Model labels its own unlabeled data confidently

METHODS 2) Co-training: mainly for classification, 96 two models label data for each other improves learning from limmited data, gives two views, a better acuracy (low bias)