

# Commonly Used Text Features

- Character n-grams: can be mixed with different n's
  - General and robust to spelling/recognition errors, but less discriminative than words
- Word n-grams: can be mixed with different n's
  - Unigrams are often very effective, but not for sentiment analysis (e.g. , “it’s not good” or “it’s not as good as”)
  - Long n-grams are discriminative, but may cause overfitting
- POS tag n-grams: mixed n-gram with words and POS tags
  - E.g., “ADJECTIVE NOUN” or “great NOUN”