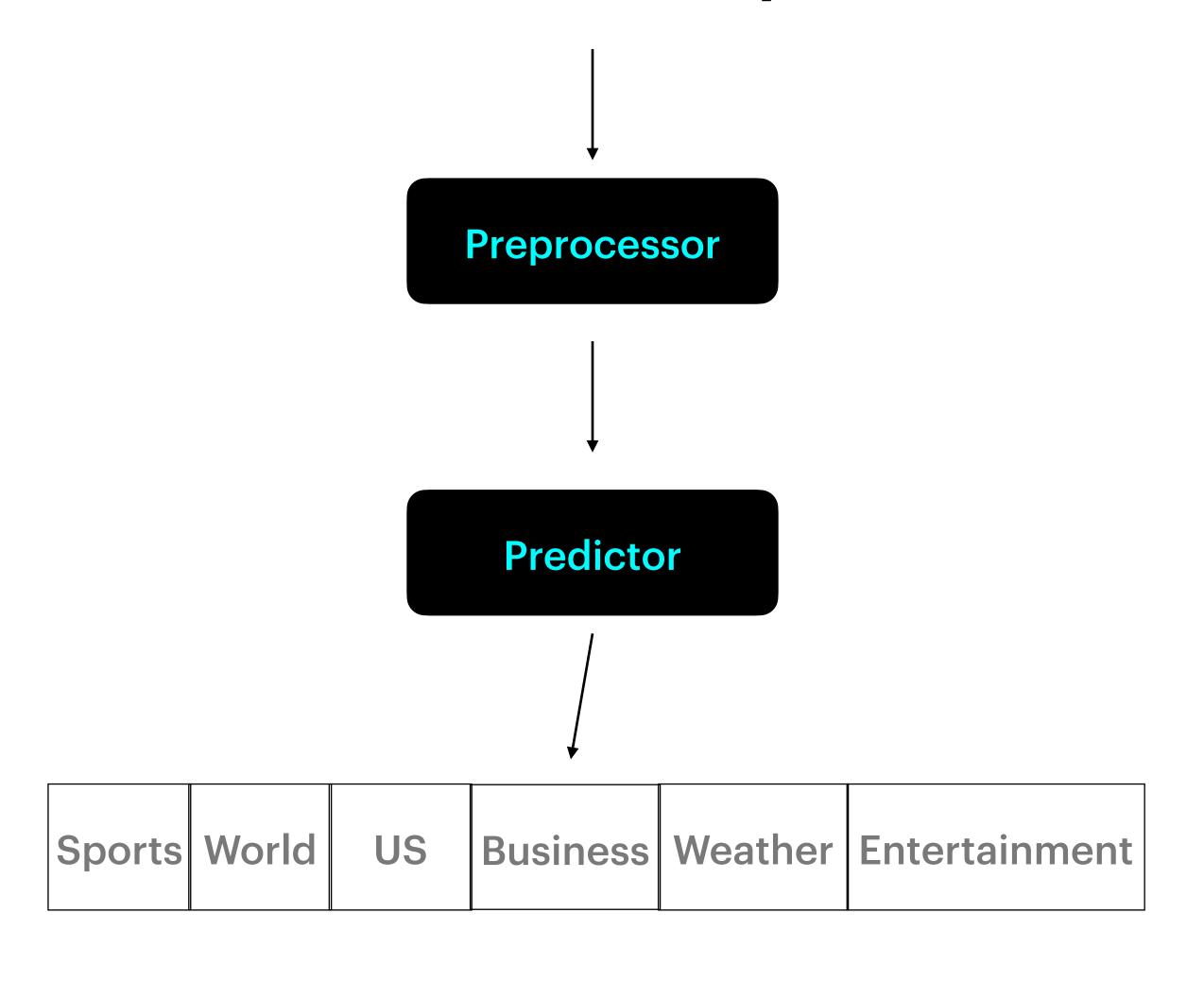# Typical text classification pipeline

"Amazon warehouse workers in the US are required to wear masks again"

Preprocessor

Predictor

Sports	World	US	Business	Weather	Entertainment

16

Text normalization

- Contraction expansion

  ```
  aren't  -> are not
  isn't   -> is not
  they'll -> they will
  ```

- Punctuation & whitespace stripping

  ```
  … had, for various reasons, went broke.  … -> had for various reasons went broke …
  ```

- Capitalization

  ```
  Sara is gregarious -> sara is gregarious
  ```

- Stop word removal (applies to phrases, too)

  ```
  Determiners:  For, an, nor, but
  Conjunctions: the, a, an, another
  Prepositions: in, under, towards, before
  ```

- These normalizations are often performed using regular expressions