Raw dialog

Utterance-level preprocessing

Tokenization

Apply `TweetTokenizer`

Padding at message level

Pad on max message length

Numericalization

Mapping tokens to FastText vectors indexes

Processed utterances

Dialog-level preprocessing

Padding at dialog level

Pad on max dialog length

Processed dialog
as a list of processed utterances

Raw utterances

Dialog and labels formatting

Text formatting

Join text of utterances with a [SEP] token

Padding at dialog level

Pad on max dialog length

Dialog-level preprocessing

Tokenization

Apply a BERT tokenizer
Padding included (max length strategy)

Processed dialog
as a single object