# Hypothetical effects on Processing changes:

**Handling Punctuation and Special Characters:** Depending on your dataset, you might want to remove or retain certain punctuation marks or special characters. In some cases, punctuation can change the meaning of a word (e.g., "not" vs. "not!") or indicate strong emotions, which can be relevant for detecting hate speech.

**Removing Rare Words:** Words that appear very infrequently in your dataset might not contribute much to the classification and could even lead to overfitting. You can consider removing these rare words.

**Dealing with Negations:** Negations can significantly alter the sentiment or meaning of a phrase (e.g., "not good" vs. "good"). You might want to ensure that negations are properly handled, perhaps by combining negated words with the words that follow them (e.g., "not\_good").

**Character Filtering**: Filtering out unnecessary characters or correcting repeated character usage (e.g., "loooove" to "love") can help standardize your text data.

**Lemming or Stemming:** Lemming sometimes removes relvant semtantic aspects of words, therefore it makes sense to compare an approach between the two.

## Vorschlag Adrian:

**# Lemmatization (default behavior, without stemming)**

data\_train['text\_lemmatization'] = data\_train['comment\_text'].apply(lambda x: preprocess\_text(x, use\_stemming=False))

**# Stemming (enabling stemming, no lemmatization)**

data\_train['text\_stemming'] = data\_train['comment\_text'].apply(lambda x: preprocess\_text(x, use\_stemming=True))

**# Keeping semantic punctuation (keeping ! and ?)**

data\_train['text\_punctuation'] = data\_train['comment\_text'].apply(lambda x: preprocess\_text(x, use\_stemming=False, keep\_semantic\_punctuation=True))

**# Removing all punctuation**

data\_train['text\_no\_punctuation'] = data\_train['comment\_text'].apply(lambda x: preprocess\_text(x, use\_stemming=False, keep\_semantic\_punctuation=False))

Diese 4 mit dem gesamten processing testen und beste accuracy dann nehmen. Und dann die anderen aufbauen vergleichen, siehe unten.

data\_text\_1: Only lowercase applied.

data\_text\_2: Lowercase and stopwords removal.

data\_text\_3: Add punctuation handling to the above.

data\_text\_4: Incorporate lemmatization.

data\_text\_5: Incorporate stemming instead of lemmatization.

Poster: A0

Intro:

Pipeline:

Folien anschauen Vorlesung 4 Seite 6-7