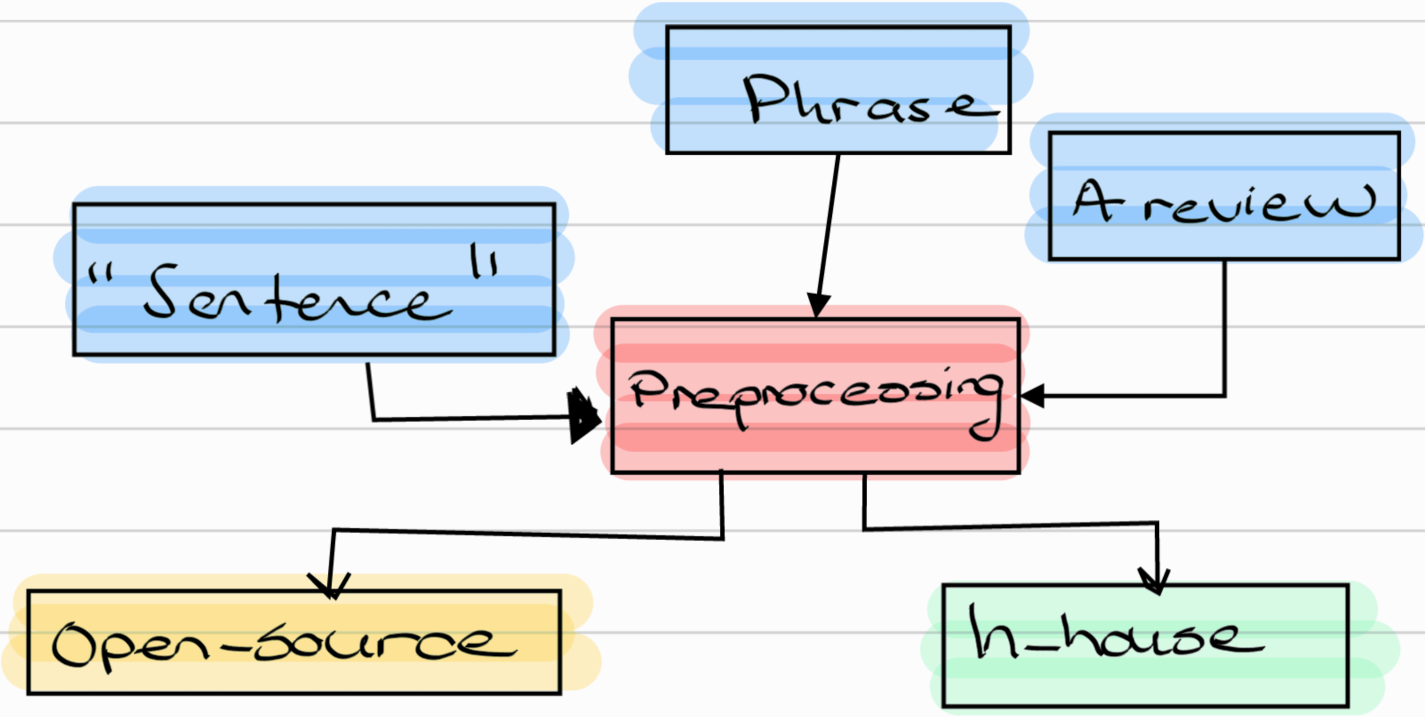


Open-source

- Punctiation removed ✓
- Stopwords removed ✓
- Lower case ✓
- Word split ✓
- Vectorizer: Count Vectorizer
- Feature Number: 10000
- target vector by ⇒ VADER

In-house

- Punctuation ✓
- Stopwords ✓
- Lower case ✓
- Word split — No, tokens lemmatizations
- Vectorizer: In house function
- Feature Number: 209 (?)
- target vector by ⇒ As given

NB →

SVM →

K-Means →

DT →

KNN →

LR →

supervised	Open Source	In-house
Naive Bayes	0.784	
Gaussia Bayes	0.527	
Desicion Tree	0.757	
Random Desicion Tree	0.795	
K-nn	?	
SVM	? 0.833	
Logistic Regression	0.798	

- Naive Bayes, Logistic Regression, SVM, Desicion Tree look good.
- We have a small "test" data, labelled - by "HUMAN".

supervised	Open Source	In-house
Naive Bayes	0.653	
Gaussia Bayes	0.525	
Desicion Tree	0.633	
Random Desicion Tree	0.643	
K-nn	?	
SVM	?	
Logistic Regression	0.634	

- Human-labelled data show lower results → data is limited (small-sized)
  ↳ our-models cannot make prediction as good as human!

OR!

The prediction of human is NOT consistent?

We made an experiment. We chose a sentences and regularly asked students (native / not native balanced) to label the sentence "1" or "0" pos/neg.

Result 50% say 1
50% say 0

1) The failure / the issue with human data is well-known. (Give ref). Needs to be done by people multiple-times to decrease the error.

2) What if we don't push into 2 cathegories.

Wh we start with binary-classification?

→ Neutral scored data is 20 times more than pozitive or negative scored data.

We want to focus on "polarity" of sentiment.

→ Human-error is a common problem, people who uses human annotators uses following sequence:

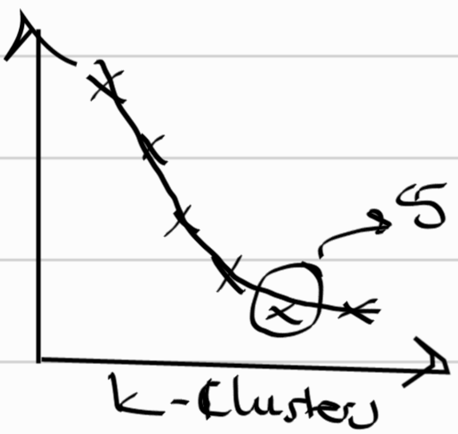
each sentence → 2 times by same annotator

↳ 2 differt by different annotator

↳ get a CI at the end.

→ $\boxed{\text{K-Means}}$

↓



PCA analysis (not look good

Plans for Competitions