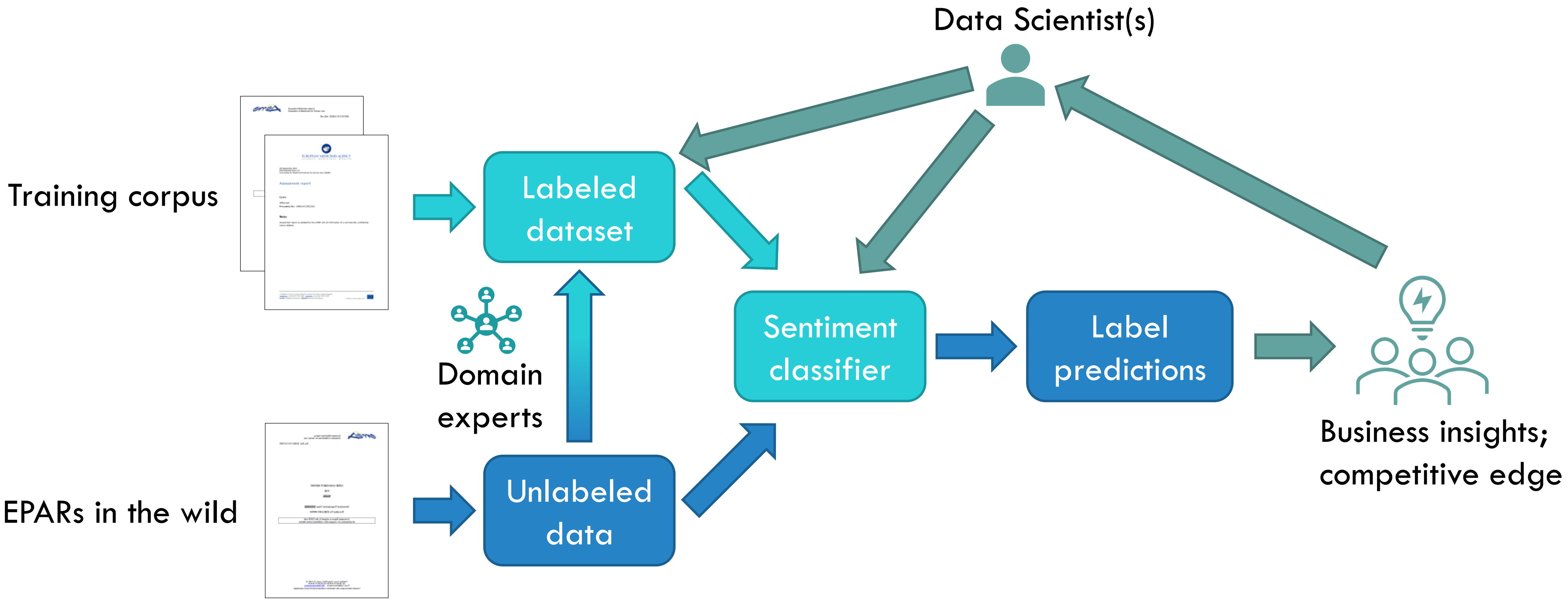
SENTIMENT ANALYSIS FROM EPAR DOCUMENTS

Panu Aho, Data Scientist
17.1.2021

AGENDA

- Business objective of research
- Training data overview
- Sentiment classification model pipeline
  - Data cleaning
  - Data preprocessing
  - Classifier training
- Results
- Discussion
  - Achievements
  - Limitations
  - Future research

BUSINESS OBJECTIVES



TRAINING DATASET

- An annotated dataset was provided in a 266 × 5 matrix containing an integer ID, Sentence and one-hot-encoded sentiment labels
- Out of the 266 rows, 236 were unique while 30 were at least once duplicated (not taking in to account the ID column)

	ID	Sentence	Positive	Negative	Neutral
0	1	[redacted]	1	0	0
1	2		1	0	0
2	3		0	0	1
3	4		1	0	0
4	5		1	0	0

DUPLICATED SENTENCES

- 30 rows contained sentences that were identical replicas of other rows (excluding the first occurence)
  - In these instances, also the labels were identical
- In addition, some near-identically replicated sentences (~spelling mistakes) were found by examining the data visually
- Without further knowledge of the data gathering/annotation process, the decision was made to retain the replicated in data.

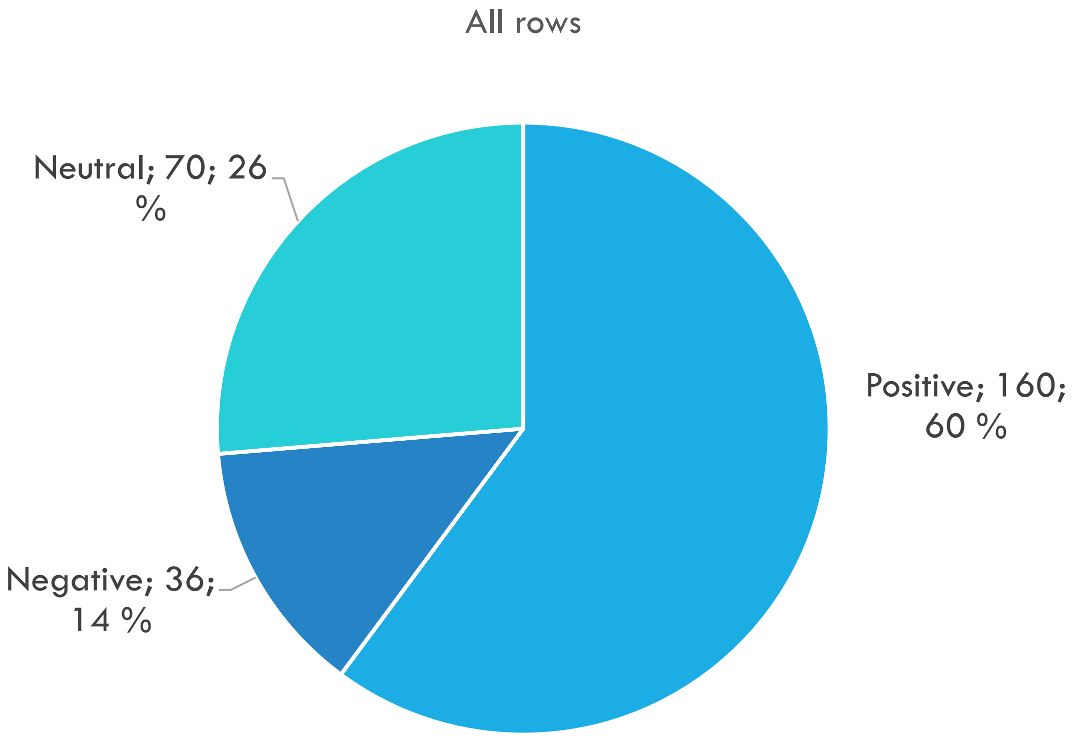
Identical training examples

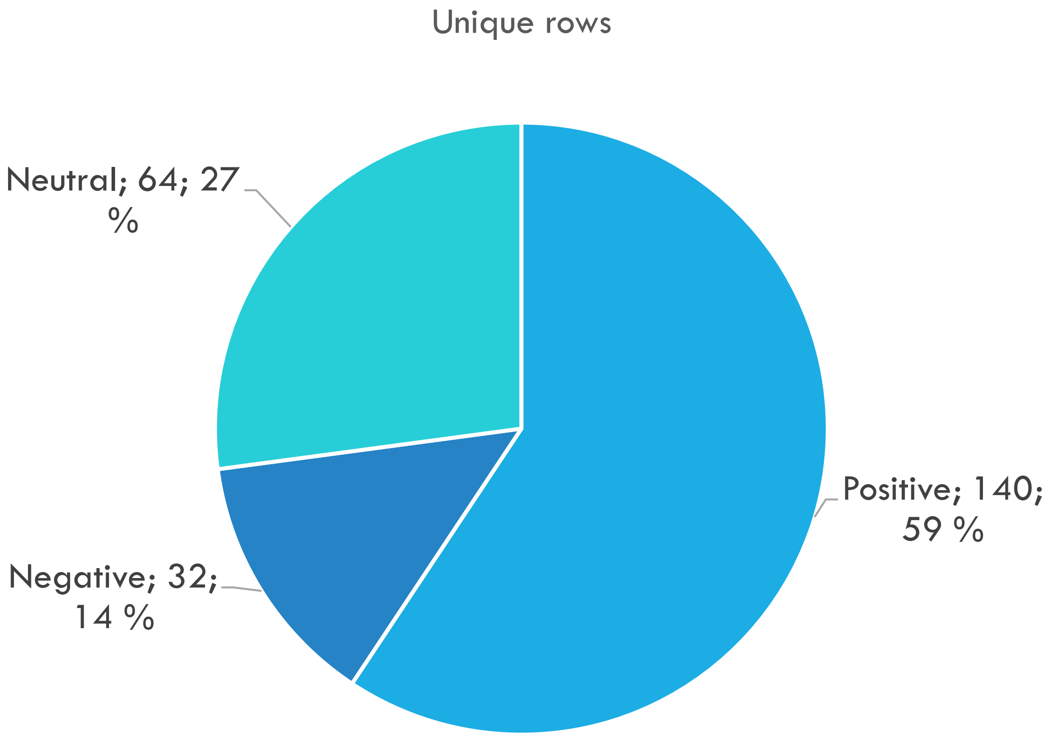
21	[redacted]	based on the efficacy data.
137	[redacted]	based on the efficacy data.
148	[redacted]	based on the efficacy data.

Near-identical training examples

30	[redacted]	follow-up period should be provided	[redacted]
157	[redacted]	Follow-up Period should be provide	[redacted]
146	[redacted]	Follow-up Period should be provide	[redacted]

LABEL DISTRIBUTION

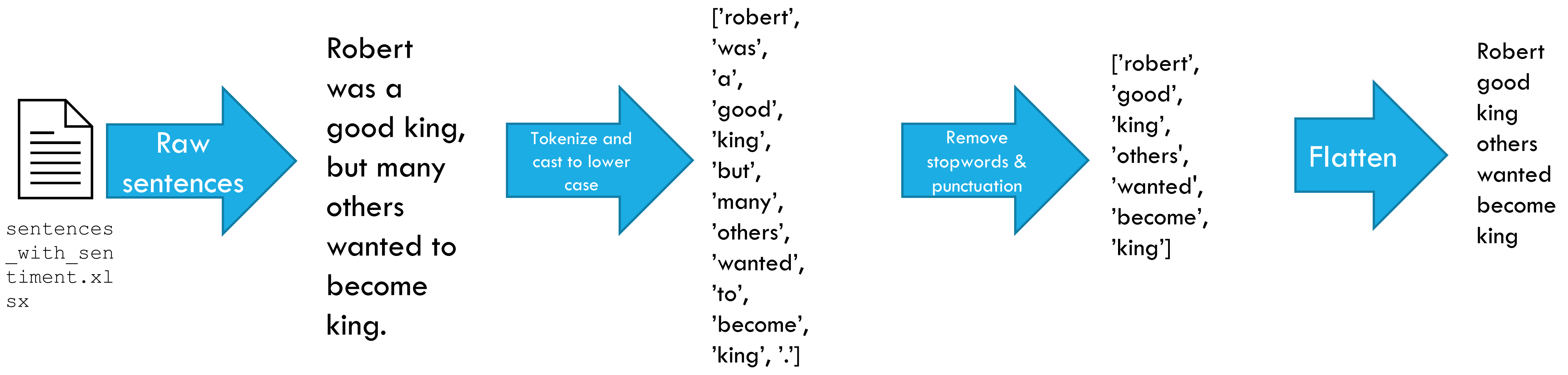




DATA EXPLORATION

- Frequent (top-ten) 1-grams, 2-grams and 3-grams were exctracted from the cleaned data
- Top-ten 1-grams have many terms replicated in both positive and negative classes, but the occurence rates are different
- Negative class frequently has terms with obvious negative connotation

DATA CLEANING



Illustrative example – not actual training data

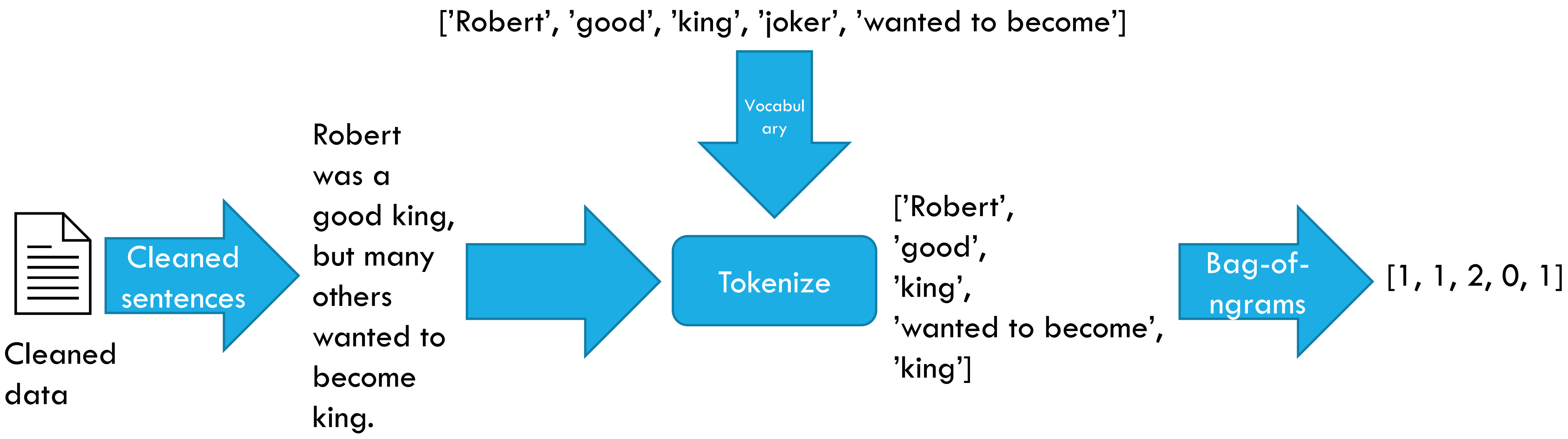
STOP WORDS

- A starting point for stop words list was obtained from the `nltk` Python library
- `nltk.corpus.stopwords.words('english')`
- The list was further hand-tweaked in an attempt to reduce the noise present by meaningless words such as 'a', 'the', 'it' etc. while keeping acceptable discriminative power between classes.

```
1 print(STOP_WORDS)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves',
'you', "you're", "you've", "you'll", "you'd", 'your', 'your
s', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself',
'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'it
s', 'itself', 'they', 'them', 'their', 'theirs', 'themselve
s', 'what', 'which', 'who', 'whom', 'this', 'that', "that'l
l", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be',
'been', 'being', 'have', 'has', 'had', 'having', 'do', 'doe
s', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'o
r', 'as', 'of', 'at', 'by', 'for', 'with', 'about', 'into',
'through', 'during', 'to', 'from', 'in', 'out', 'on', 'off',
'then', 'once', 'here', 'there', 'when', 'where', 'why', 'ho
w', 'both', 'each', 'other', 'such', 'own', 'so', 's', 't',
'can', 'will', 'just', 'now', 'd', 'll', 'm', 'o', 're', 'v
e', 'y']
```

FEATURE EXTRACTION



Illustrative example – not actual training data

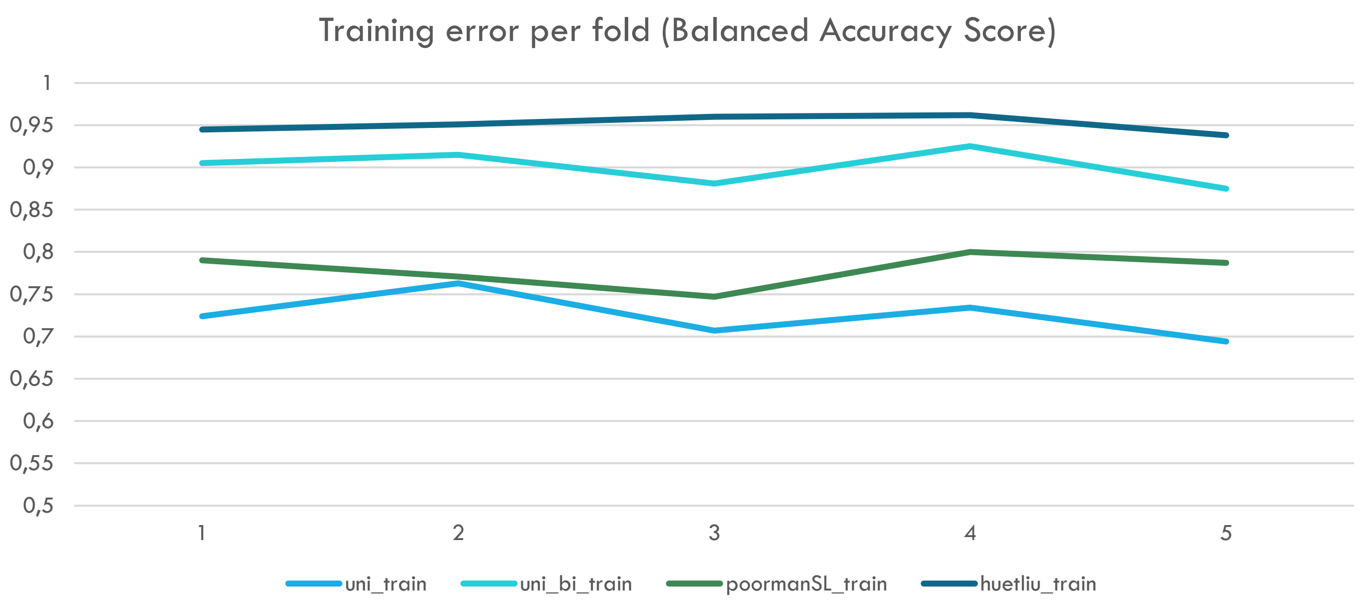
VOCABULARY?

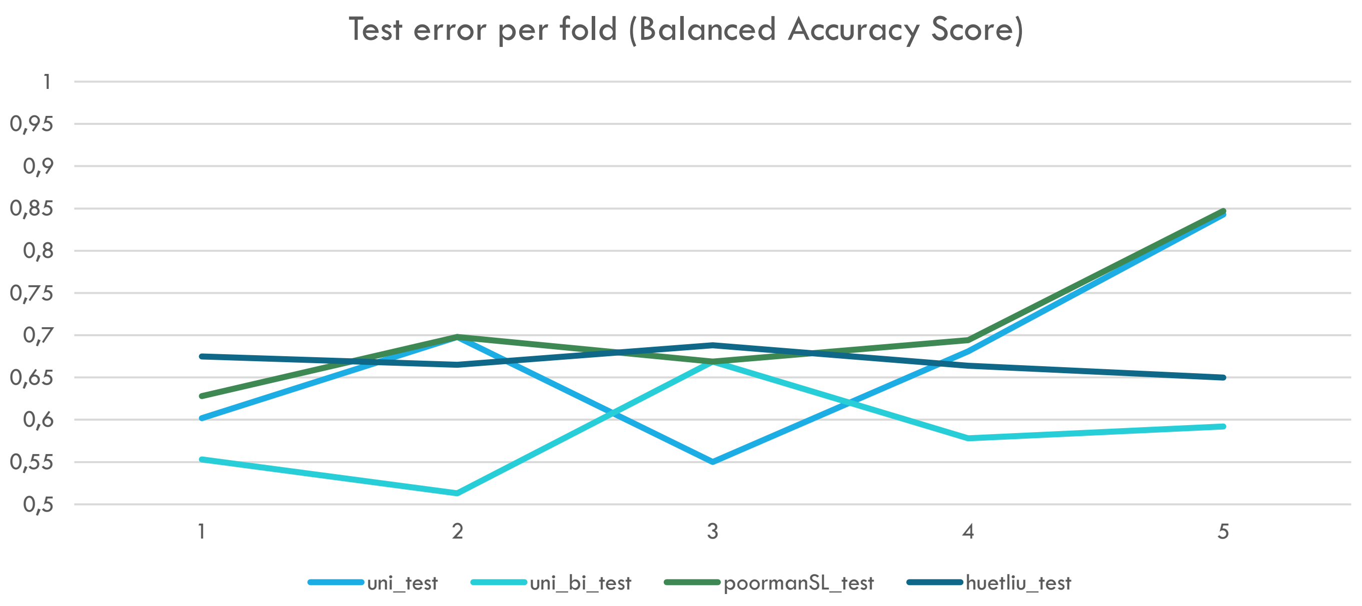
- Configurations tried:
  - Top ten 1-grams
  - Top ten 1-grams + 2-grams
  - Top ten 1-grams + hand-engineered ”poor man’s” *Sentiment Lexicon (SL)*
    - Contains 11 ”low-hanging fruit” phrases obviously associated[1] with one of the classes and repeated multiple times in dataset, e.g.:
      - ’These objectives have been met’
      - ’Data are considered very limited’
    - The phrases were gathered by arranging sentences alphabetically and visually examining them without looking at the labels
  - Top ten 1-grams + poor man’s SL + ~6800 word open source English SL (Hu & Liu 2004) [1]

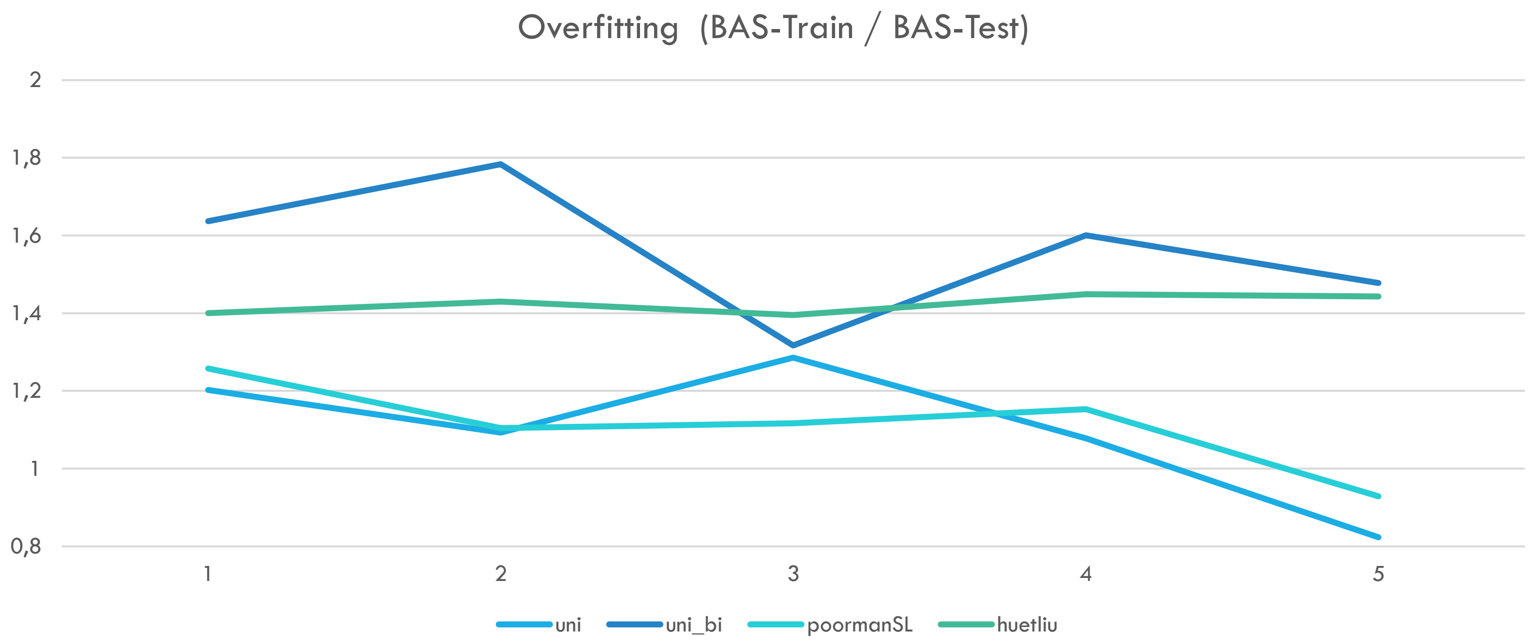
1) According to my subjective perception as a non –domain expert

TRAINING

- A supervised classification model was trained using the scikit-learn library with the different vocabularies
- A support vector machine with linear kernel was chosen as initial approach due to its well-reported performance in similar tasks [2, 3, 4]
- Default parameters
- Random train/test split, 5-fold cross validation → 20 % used for testing per fold

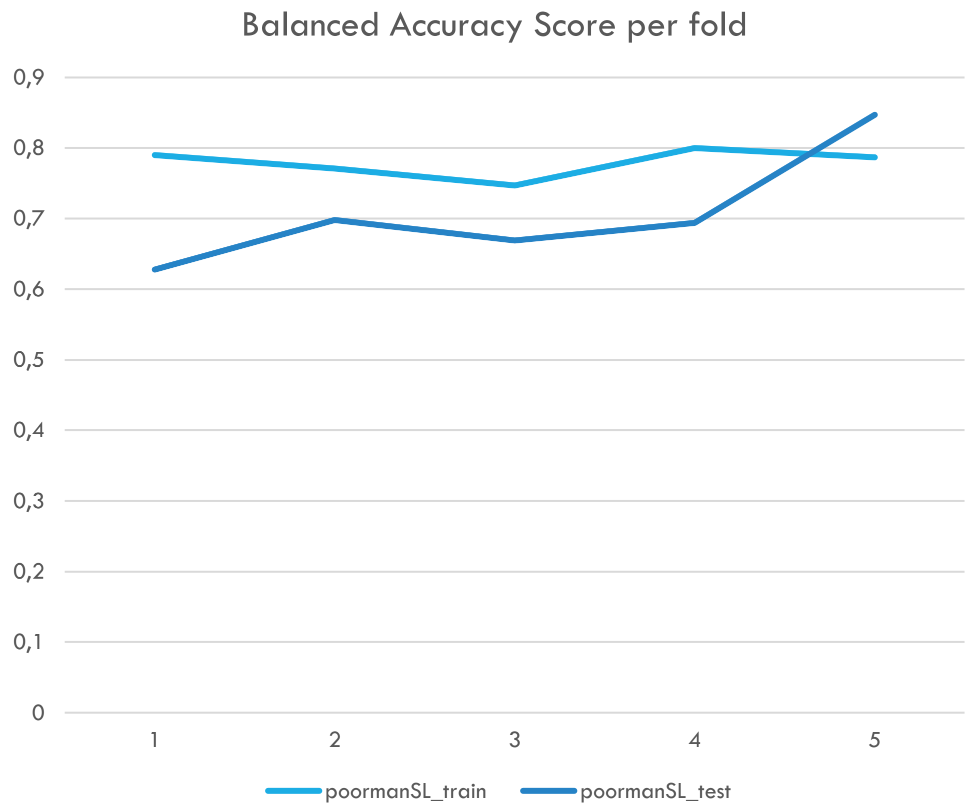

Training error per fold (Balanced Accuracy Score)
1
0,95
0,9
0,85
0,8
0,75
0,7
0,65
0,6
0,55
0,5
1
2
3
4
5
uni_train
uni_bi_train
poormanSL_train
huetliu_train


Test error per fold (Balanced Accuracy Score)
1
0,95
0,9
0,85
0,8
0,75
0,7
0,65
0,6
0,55
0,5
1
2
3
4
5
uni_test
uni_bi_test
poormanSL_test
huetliu_test


Overfitting (BAS-Train / BAS-Test)
2
1,8
1,6
1,4
1,2
1
0,8
1
2
3
4
5
uni
uni_bi
poormanSL
huetliu

RESULTS

- Hand-engineered domain specific sentiment lexicon together with most popular 1-grams seems like a promising approach w.r.t.:
  - Good model robustness across folds
  - Not grossly overfitting
- Using a generic open-source SL also yielded surprisingly good results in terms of robustness, but with significant model overfitting
- In other approaches tried, robustness was questionable (~widely varying performance accross folds in test set)



DISCUSSION

- Achievements
  - A simple bag-of-ngrams approach with hand-selected features and linear SVN shows promising results
  - Naive approaches (e.g. guessing the largest class) are outperformed by a comfortable margin
- Limitations
  - Dataset is too small to effectively make use of complex features such as 2-grams and beyond (→ Risk overfitting)
  - Class proportions slightly skewed
  - BoW does not take word ordering and other context in to account, so model does not probably misses important information

DISCUSSION

- Suggestions for future work
  - Build a CI/CD pipeline, deploy the model to production and see how it performs in the wild!
  - Improve the Sentiment Lexicon vocabulary
  - Additional feature engineering – look into *Scientific Citation Sentiment Analysis (SCSA)* [5, 6]
  - If possible to get a LOT of data, could try more complex approaches e.g. bidirectional LSTM

Example		*Our data are generally consistent with that of other studies [TC], but not with studies where a single dose of paracetamol was administered [OC].*	*These values are lower than those reported by French et al. [20].*
Step 1		*[Our data]*CITINGWORK *are generally [consistent with]*POSITIVE *that of other studies [TC]*TC *, [but]*CONTRAST *[not]*NEGATION *with studies where a single dose of paracetamol was administered [OC]*OC	*[These values]*CITINGWORK *are [lower]*COMPARATIVE *[than]*THAN *those reported by French et al. [TC]*TC
Step 2	Unigram	'CITINGWORK', 'POSITIVE', 'TC', 'CONTRAST', 'NEGATION', 'OC'	'CITINGWORK', 'COMPARATIVE','THAN', 'TC'
	Bigram	'CITINGWORK_POSITIVE', 'POSITIVE_TC','TC_CONTRAST', 'CONTRAST_NEGATION', 'NEGATION_OC'	'CITINGWORK_COMPARATIVE', 'COMPARATIVE_THAN', 'THAN_TC'
	Trigram	'CITINGWORK_POSITIVE_TC', 'POSITIVE_TC_CONTRAST', 'TC_CONTRAST_NEGATION', 'CONTRAST_NEGATION_OC'	'CITINGWORK_COMPARATIVE_THAN', "COMPARATIVE_THAN_TC'
	Direction	'CITINGWORK_CONTRAST_DIR', 'TC_CONTRAST_DIR', 'CONTRAST_OC_DIR'	

Examples of extracting *structure features* from citation data related to clinical trials, as proposed in [6]

REFERENCES

[1] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge. Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

[2] Walaa Medhat, Ahmed Hassan, Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. Volume 5, Issue 4. 2014 https://doi.org/10.1016/j.asej.2014.04.011

[3] Mullen, Tony & Collier, Nigel. (2004). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. 412-418.

[4] Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." International Journal 2.6 (2012): 282-292.

[5] Yousif, A., Niu, Z., Tarus, J.K. et al. A survey on sentiment analysis of scientific citations. Artif Intell Rev 52, 1805–1838 (2019). https://doi.org/10.1007/s10462-017-9597-8

[6] Xu J, Zhang Y, Wu Y, Wang J, Dong X, Xu H. Citation Sentiment Analysis in Clinical Trial Papers. AMIA Annu Symp Proc. 2015;2015:1334-1341. Published 2015 Nov 5.

THANK YOU!

Panu Aho

Data Scientist
+358 50 4129 150
panu.aho@gmail.com
linkedin.com/in/panuaho/