Praktikum Smart Data Analytics

Übungsblatt 4

25.07.2022

Gruppe 4: Ola Sharfeldin, Xin Tian, Dmitrii Seletkov, Danilo Rosenthal, Dan Jia

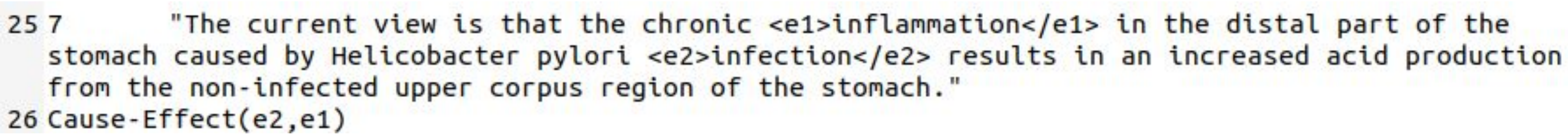
Agenda

- Causal Extraction
  - Task
  - Datasets
  - Metrics
- Baseline
  - Naive Bayes
  - Other models
- One More Model: CNN
- Summary

Datasets

Causal Extraction

- Task:
  - Input: sentence and tags(e1, e2) with cause-effect
  - Output: multi-class classification
    - Other: no causality
    - Cause-Effect(e1,e2)
    - Cause-Effect(e2,e1)

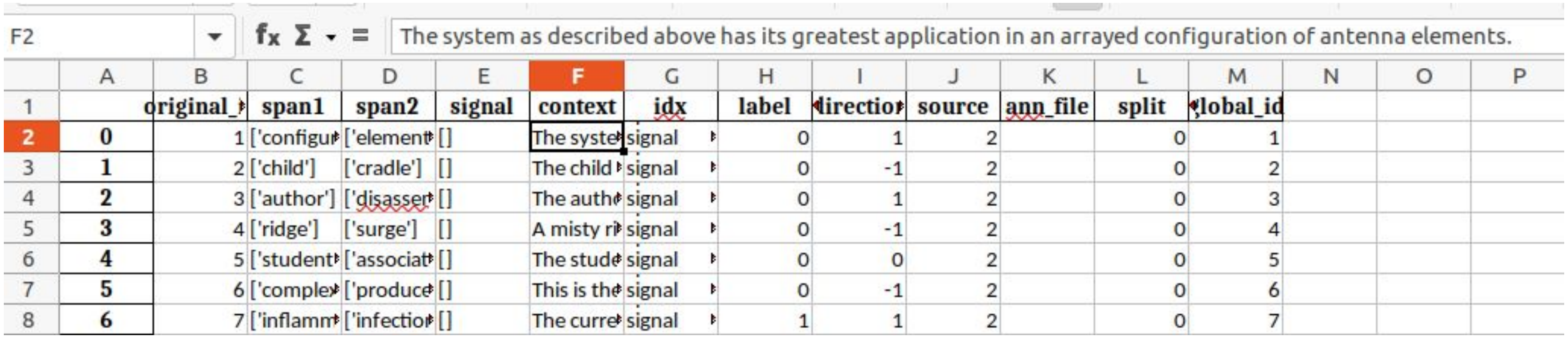


```
7	"The current view is that the chronic <e1>inflammation</e1> in the distal part of the
stomach caused by Helicobacter pylori <e2>infection</e2> results in an increased acid production
from the non-infected upper corpus region of the stomach."
Cause-Effect(e2,e1)
```

Example of the Sentence in SimEval2010

Datasets And Preprocessing

- Picked datasets:
  - SimEval2007
  - SimEval2010
- Problem 1: more classes than in CE
  - e.g. SimEval2010 has 19 classes
    - Other
    - Cause-Effect(e1,e2), Cause-Effect(e2,e1)
    - Component-Whole(e1,e2), Component-Whole(e2,e1)
    - etc.
- Use CREST for common format and getting only causality labels, i.e.
  - Other, Cause-Effect(e1,e2), Cause-Effect(e2,e1)



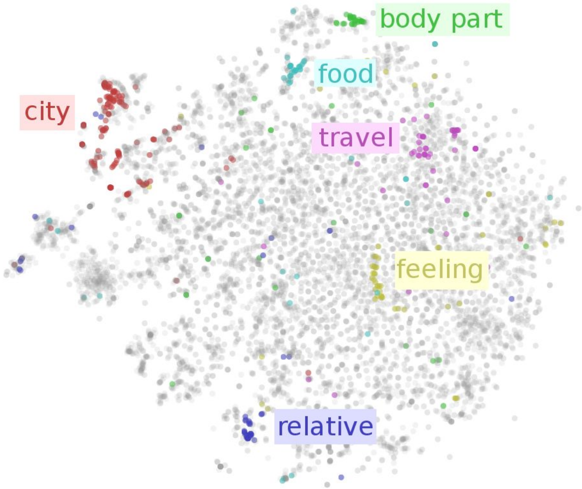
F2 fx Σ = The system as described above has its greatest application in an arrayed configuration of antenna elements.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		original_	span1	span2	signal	context	idx	label	directior	source	ann_file	split	global_id			
2	0	1	['configur	['element	[]	The syste	signal	0	1	2		0	1			
3	1	2	['child']	['cradle']	[]	The child	signal	0	-1	2		0	2			
4	2	3	['author']	['disasser	[]	The auth	signal	0	1	2		0	3			
5	3	4	['ridge']	['surge']	[]	A misty ri	signal	0	-1	2		0	4			
6	4	5	['student	['associat	[]	The stude	signal	0	0	2		0	5			
7	5	6	['complex	['produce	[]	This is the	signal	0	-1	2		0	6			
8	6	7	['inflamm	['infectior	[]	The curre	signal	1	1	2		0	7			

Example of the Sentence in SimEval2010 after CREST

Datasets And Preprocessing

- Problem 2: special characters
  - Remove
- Problem 3: Input of words into model
  - Word Embedding
- Preprocessing pipeline:
  - Convert to CREST to obtain only CE labels
  - Remove special characters
  - Obtain feature vectors with Word Embedding



Word Embedding Visualisation Map

Metrics

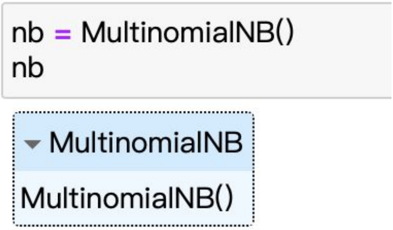
- Based on the survey [Yang21]
- Common metrics
  - Precision
  - Recall
  - F1 Score
- Problems: overoptimistic, misleading results on imbalanced datasets

- Also recommended:
  - Matthews Correlation Coefficient (MCC):
    - Only high (->1) if the classifier does both positives and negatives well
  - Geometric Mean (G-Mean):
    - Poor performance in positive examples prediction lead to a low G-mean value, even if negative instances are correctly classified by the classifier

Baseline

Baseline

- Naive Bayes [Sorgente13]
  - simple and easy to implement
  - not sensitive to irrelevant features
  - fast and can be used to make real-time predictions
  - easy to explain



```
nb = MultinomialNB()
nb
```

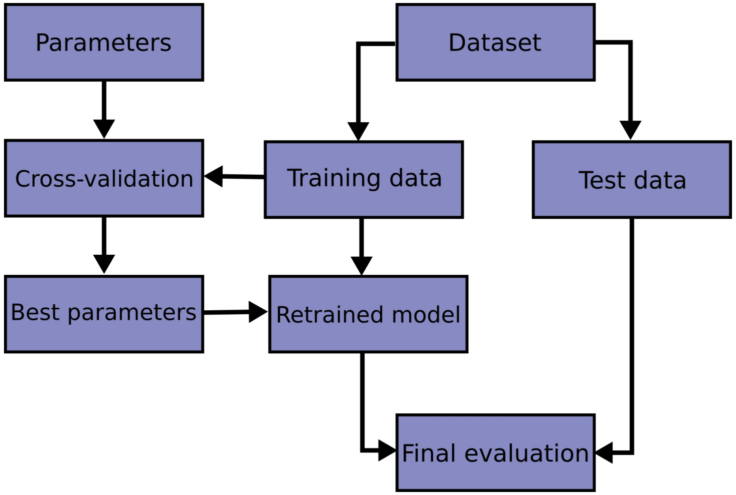
```
MultinomialNB
MultinomialNB()
```

Train it on following datasets

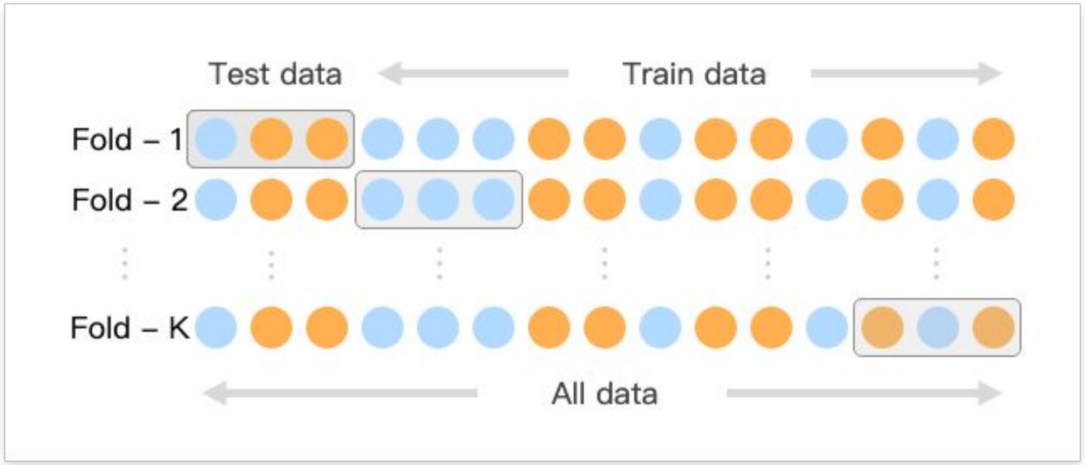
- SimEval2007
- SimEval2010
  - possible reason: the data does not obey gaussian distribution.

Datasets	Accuracy	Precision	Recall	F1	MCC	G-Mean
SimEval2007	0.67	0.88	0.67	0.75	0.07	0.0
SimEval2010	0.38	0.78	0.38	0.49	0.01	0.35

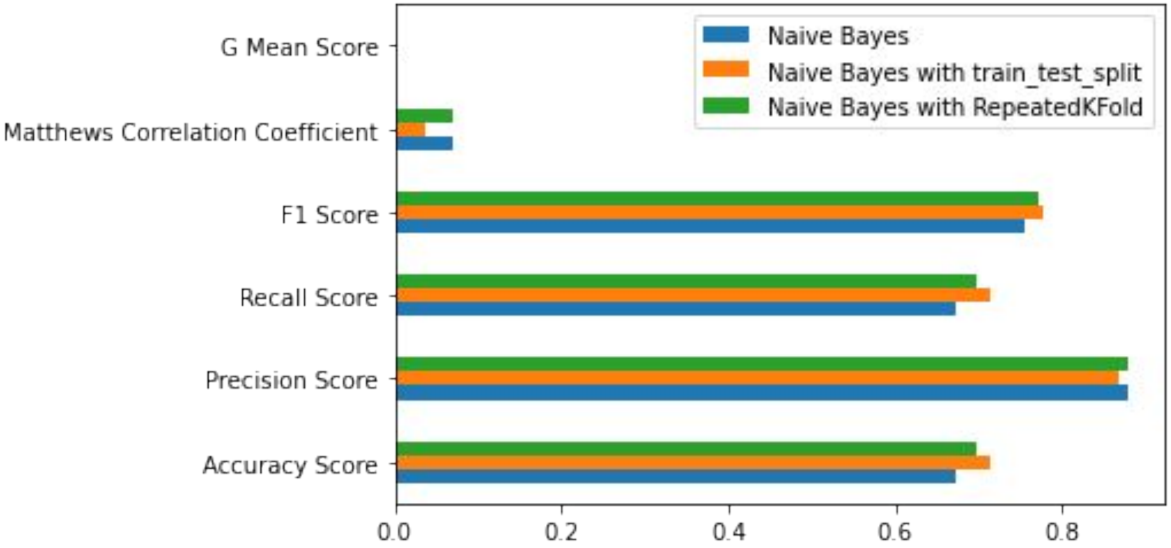
Cross validation

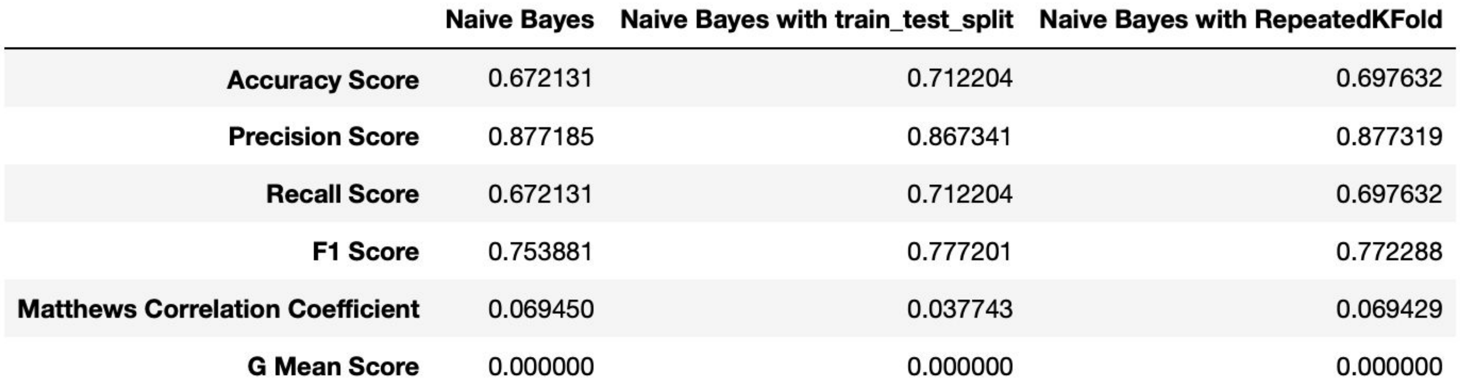


- train_test_split
- repeatedKFold



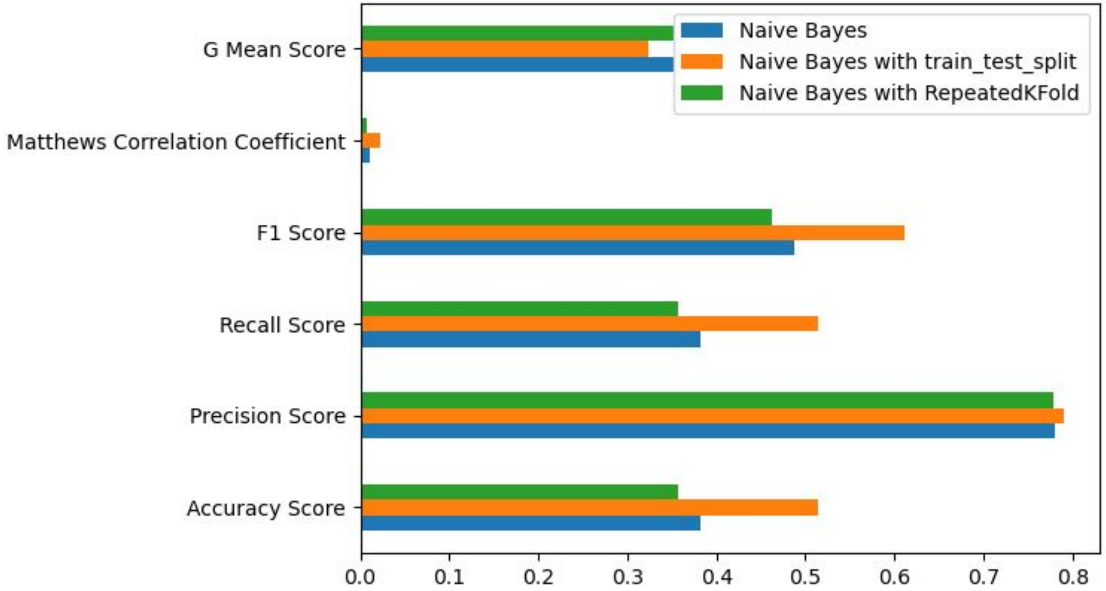
Cross validation results 1 (SimEval2007)

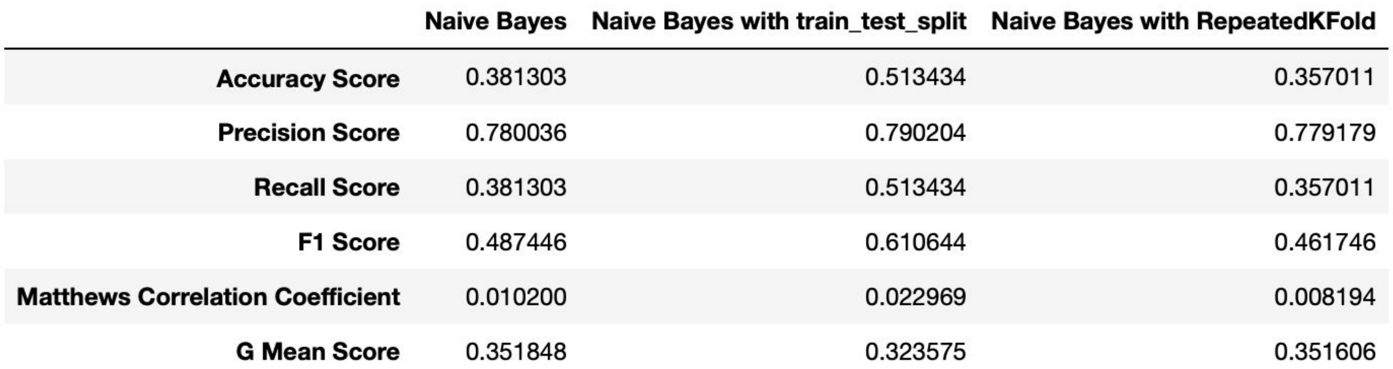




	Naive Bayes	Naive Bayes with train_test_split	Naive Bayes with RepeatedKFold
Accuracy Score	0.672131	0.712204	0.697632
Precision Score	0.877185	0.867341	0.877319
Recall Score	0.672131	0.712204	0.697632
F1 Score	0.753881	0.777201	0.772288
Matthews Correlation Coefficient	0.069450	0.037743	0.069429
G Mean Score	0.000000	0.000000	0.000000

Cross validation results 2 (SimEval2010)





	Naive Bayes	Naive Bayes with train_test_split	Naive Bayes with RepeatedKFold
Accuracy Score	0.381303	0.513434	0.357011
Precision Score	0.780036	0.790204	0.779179
Recall Score	0.381303	0.513434	0.357011
F1 Score	0.487446	0.610644	0.461746
Matthews Correlation Coefficient	0.010200	0.022969	0.008194
G Mean Score	0.351848	0.323575	0.351606

Other Experiments

- SimEval2007

Model	Accuracy	Precision	Recall	F1	MCC	G-Mean
Bayes (Baseline)	0.67	0.88	0.67	0.75	0.07	0.0
Logistic Regression	0.90 (0.000)	0.854 (0.000)	0.90 (0.000)	0.88 (0.000)	0.04 (0.000)	0.00 (0.000)
SVM	0.92 (0.000)	0.85 (0.000)	0.92 (0.000)	0.89 (0.000)	0.00 (0.000)	0.00 (0.000)
Decision Tree	0.85 (0.009)	0.86 (0.004)	0.86 (0.008)	0.86 (0.006)	0.013 (0.025)	0.00 (0.000)
Random Forest	**0.93 (0.001)**	**0.89 (0.037)**	**0.93 (0.0001)**	**0.89 (0.002)**	**0.06 (0.074)**	**0.06 (0.076)**

Other Experiments

- SimEval2010

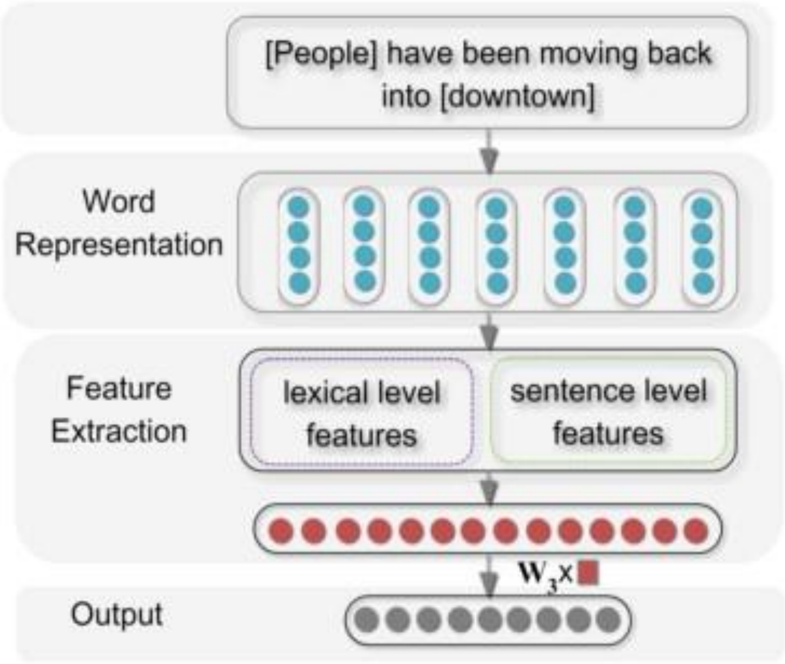
Model	Accuracy	Precision	Recall	F1	MCC	G-Mean
Bayes (Baseline)	0.38	0.78	0.38	0.49	0.01	0.35
Logistic Regression	0.87 (0.000)	0.77 (0.000)	0.87 (0.000)	0.82 (0.000)	0.02 (0.000)	0.00 (0.000)
SVM	0.87 (0.000)	0.77 (0.000)	0.87 (0.000)	0.82 (0.000)	0.00 (0.000)	0.00 (0.000)
Decision Tree	0.78 (0.001)	0.81 (0.002)	0.79 (0.001)	0.80 (0.001)	0.14 (0.009)	0.33 (0.015)
Random Forest	**0.88** **(0.001)**	**0.85** **(0.035)**	**0.88** **(0.000)**	**0.83** **(0.000)**	**0.10** **(0.010)**	**0.02** **(0.027)**

One More Model

CNN with Max Pooling and word embedding

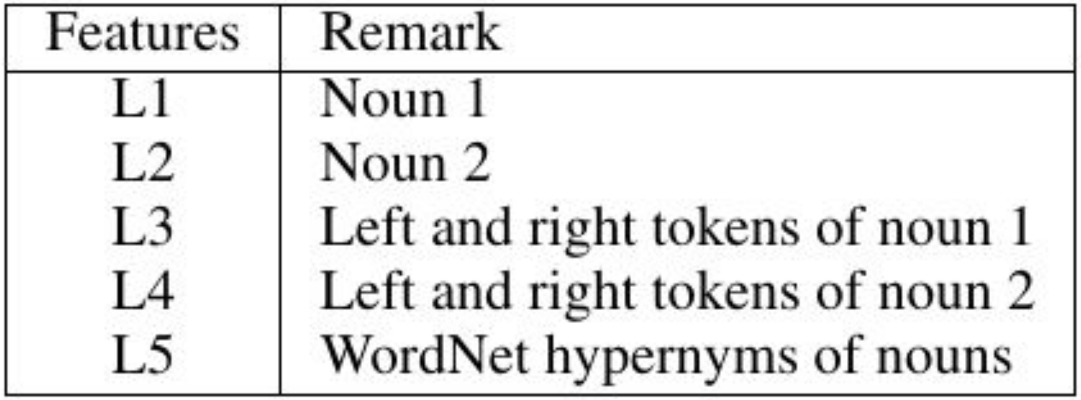
[zeng2014]

- Embedding Layer
  - Word Representation[turian2010]
- Feature Extraction
  - Lexical Level Features
  - Sentence Level Features
- Fully-connected Layer
- Softmax Classifier



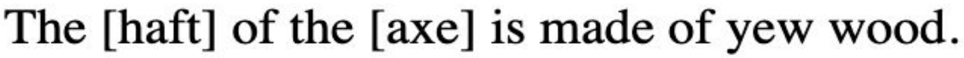
Lexical Level Features Extraction[zeng2014]

Lexical Feature

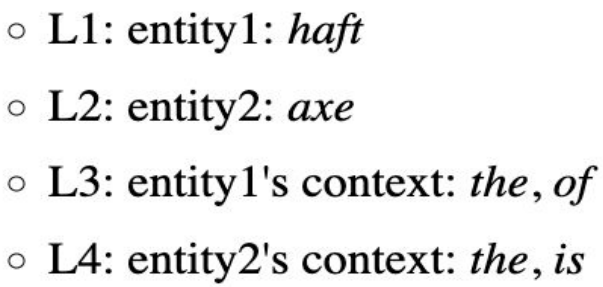


Features	Remark
L1	Noun 1
L2	Noun 2
L3	Left and right tokens of noun 1
L4	Left and right tokens of noun 2
L5	WordNet hypernyms of nouns

Example



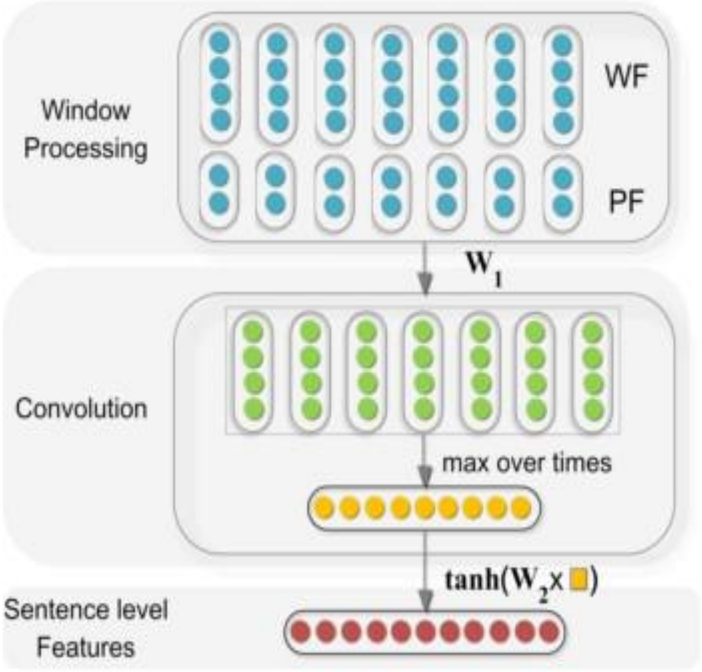
The [haft] of the [axe] is made of yew wood.



- L1: entity1: *haft*
- L2: entity2: *axe*
- L3: entity1's context: *the*, *of*
- L4: entity2's context: *the*, *is*

Sentence Level Feature Extraction[zeng2014]

- Word Features (WF)
- Position Features (PF)
- Convolution with Max Pooling
- Fully-connected Layer



Experiments Results

- SimEval2010

Model	Accuracy	Precision	Recall	F1	MCC	G-Mean
Bayes (Baseline)	0.38	0.78	0.38	0.49	0.01	0.35
CNN	**0.97 (0.002)**	**0.97 (0.002)**	**0.97 (0.002)**	**0.97 (0.002)**	**0.86 (0.010)**	**0.89 (0.009)**

- SimEval2007

Model	Accuracy	Precision	Recall	F1	MCC	G-Mean
Bayes (Baseline)	0.67	0.88	0.67	0.75	0.07	0.0
CNN	**0.94 (0.004)**	**0.93 (0.005)**	**0.94 (0.004)**	**0.93 (0.003)**	**0.47 (0.028)**	**0.63 (0.014)**

Summary

Summary

- Two datasets:
  - SimEval2007, SimEval2010

- Baseline:
  - Naive Bayes
  - Other models e.g. Random Forest

- One more model: CNN
  - Best performance by all datasets

Karlsruhe Institute of Technology

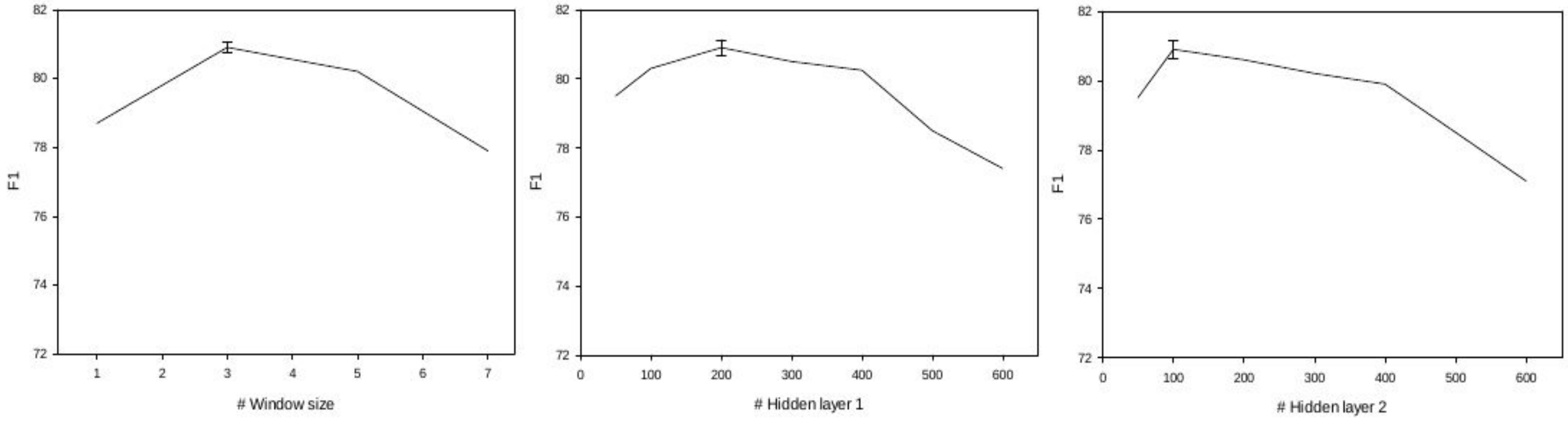
Questions?

Appendix

Experiments

Parameter Settings

Hyperparameter	Window size	Window dim.	Distance dimension	Hidden layer 1	hidden layer 2	learning rate
value	w=3	n=50	d=5	n1=200	n2=100	λ=0.01



CNN: Results (micro)

- SimEval2010

Model	Accuracy	Precision	Recall	F1	MCC	G-Mean
Bayes	0.38	0.38	0.38	0.38	0.01	0.35
CNN	**0.97** **(0.002)**	**0,97** **(0.002)**	**0.97** **(0.002)**	**0,97** **(0.002)**	**0.86** **(0.010)**	**0.89** **(0.009)**

- SimEval2007

Model	Accuracy	Precision	Recall	F1	MCC	G-Mean
Bayes	0.67	0.67	0.67	0.67	0.07	0.0
CNN	**0.94** **(0.004)**	**0.94** **(0.004)**	**0.94** **(0.004)**	**0.94** **(0.004)**	**0.47** **(0.028)**	**0.63** **(0.014)**

CNN: Results (macro)

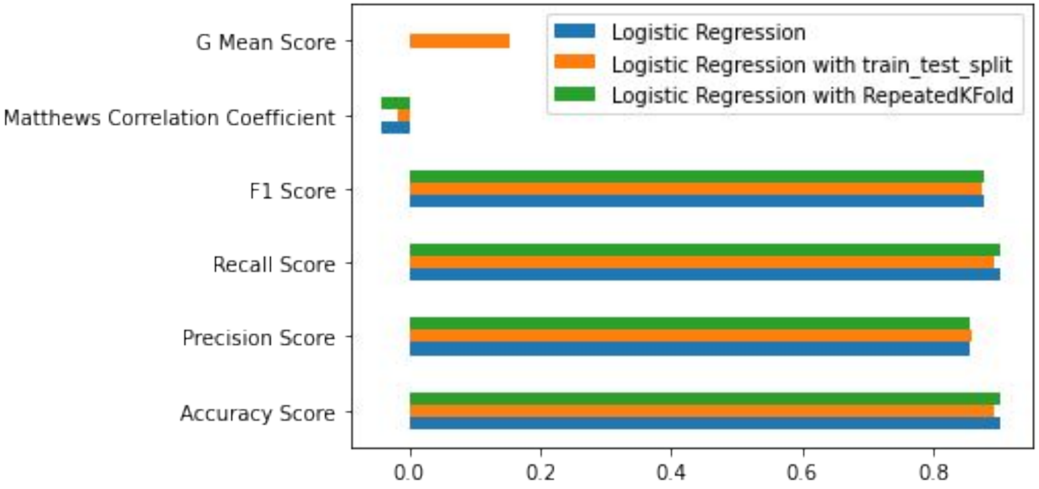
- SimEval2010

Model	Accuracy	Precision	Recall	F1	MCC	G-Mean
Bayes	0.38	0.33	0.36	0.25	0.01	0.35
CNN	**0.97 (0.002)**	**0.92 (0.007)**	**0.89 (0.008)**	**0.91 (0.007)**	**0.86 (0.010)**	**0.89 (0.009)**

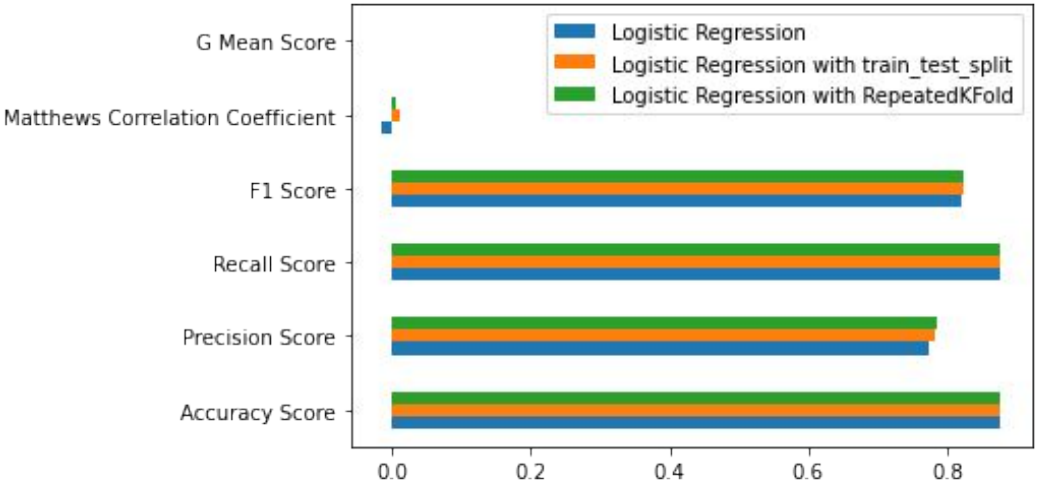
- SimEval2007

Model	Accuracy	Precision	Recall	F1	MCC	G-Mean
Bayes	0.68	0.35	0.35	0.32	0.07	0.0
CNN	**0.94 (0.004)**	**0.79 (0.033)**	**0.69 (0.004)**	**0.73 (0.009)**	**0.47 (0.028)**	**0.63 (0.014)**

Logistic Regression



simeval2007



simeval2010