



Berner Fachhochschule  
Haute école spécialisée bernoise  
Bern University of Applied Sciences

# Data Science Projekt

## Sentiment Analyse

*Nalet Meinen, Alexander Nussbaum, Michel Utz*

# Umsetzung



```
public void init() {  
    addStep((ProcessStep) new ArffResourceInputProvider().setSource("/movie_reviews_raw.arff"));  
    addStep(new PreprocessingSimpleFilter());  
    addStep(new SentimentLexiconWeightedFilter());  
    addStep(new AttributeSelectionFilter());  
    addStep(new CrossValidationFilter());  
    addStep(new NBClassifier());  
}
```

# Umsetzung

## Eigener Ansatz

- ▶ String to Word Vector
  - ▶ WordTokenizer
  - ▶ Minimum Term-Frequency
- ▶ Sentiment Lexikon von Harvard (General Inquirer)
- ▶ Gewichtung der Features anhand der Sentimentklassen
- ▶ AttributeSelection
  - ▶ Evaluator: CfsSubsetEval
  - ▶ Search: BestFirst

# Umsetzung

## Bester Ansatz

- ▶ StringToWordVector
  - ▶ IDF-TF
    - ▶ Rainbow Stopwords-Handler
    - ▶ LovinsStemmer
    - ▶ Ngram-Tokenizer Size 1-5
- ▶ AttributeSelection
  - ▶ Evaluator: CfsSubsetEval
  - ▶ Search: BestFirst

# Resultate

Ansatz	Erfolgsquote
PP_NGramAttributeSelection	81.5%
PP_AttributeSelection	78.8%
PP_RankAttributeSelection	73.95%
PP_SentimentLexiconWeight	73.7%
PP_SentimentLexiconPercent	63.5%
PP_SentimentLexiconCount	63.4%
Diverses	50-60%

# Demo

```
-----  
  _ _ _ _ _  
 | \ / _ \ / \ / _ \ / _ \ / _ \ / _ \ / _ \ |  
 | | ) / _ \ | / _ \ \ _ \ ( _ | | | _ | .` | ( _ |  
 | _ / / \ \ / / \ \ | _ \ _ | _ | _ | _ \ \ _ |  
-----
```

Welcome to our Data Science Project.

-----

Please choose an option.

- 0 PP\_Validate\_Best\_Result | For Mr. Vogel: our best result, evaluated from arff generated by ChainConfigNGramAttrSel
- 1 PP\_SentimentLexiconCount | Preprocess data, extract features and generate ARFF
- 2 PP\_SentimentLexiconPercent | Preprocess data, extract features and generate ARFF
- 3 PP\_SentimentLexiconWeight | Preprocess data, extract features and generate ARFF
- 4 PP\_AttributeSelection | Preprocess data, extract features and generate ARFF
- 5 PP\_RankAttributeSelection | Preprocess data, extract features and generate ARFF
- 6 PP\_NGramAttributeSelection | Preprocess data, extract features and generate ARFF
- 7 PP\_Validate\_SentimentLexiconWeight | Preprocess and Validate with Sentiment Lexicon
- 8 Validate | Validate a created ARFF
- 9 Exit | Leave the program

Option:

# Demo

```
Root mean squared error          0.3128
Relative absolute error          33.3436 %
Root relative squared error      76.621 %
Total Number of Instances        200

fold 10
train classifier
evaluate classifier
Evaluation done
Error rate is: 0.17
Evaluation summary:
Correctly Classified Instances    166          83      %
Incorrectly Classified Instances  34          17      %
Kappa statistic                  0.66
Mean absolute error              0.1148
Root mean squared error          0.3139
Relative absolute error          34.4095 %
Root relative squared error      76.8866 %
Total Number of Instances        200

evaluation done, mean success rate: 0.8150000000000001
Error Rate: 0.18499999999999997
Finished Validate at 1515930532448
```

# Lessons Learned

- ▶ Stopwords und Stemming bringt nicht soviel wie es verspricht.
- ▶ Lexikon und Stemmer bringen nicht soviel zusammen.
- ▶ Bei zunehmender Features-Anzahl nimmt die Aussagekräftigkeit ab.
- ▶ Weka ist nicht performant, jedoch sehr umfangreich.