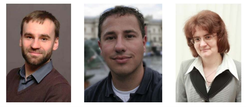
# DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data


**Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych**

Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt



**Torsten Zesch**

University of Duisburg-Essen


## Supervised Learning on Textual Data: Requirements

**Flexibility**

- single- and multi-label classification vs. regression
- feature extraction for different kinds of data: documents, document pairs, sequential data

**Replicability and Usability**

- reuse of existing components wherever possible
- detailed documentation of experiment setup
- simple reproduction of experimental results

## Why DKPro TC?

**Is DKPro TC for you?**

✓ You have **annotated**, **textual** data, and …

… you want to:

✓ train a model for **automatic learning** from your data

✓ know which **features** do a good job classifying your data

✓ make your experiments **reproducible** for you and others

… you don't want to:

¬ rerun your experiments many times to test all sorts of **different parameters** in your model

¬ **re-implement** common NLP pre-processing and feature extraction facilities

**What makes DKPro TC unique?**

- rapid prototyping due to preconfigured experiment setups
- parameter sweeping capability, i.e. testing several parameter values and parameter combinations all at once
- detailed documentation of the experimental setup
- a comprehensive, flexible, and modular architecture
- code completely open-source under ASL/GPL, i.e. you can view and modify the sources

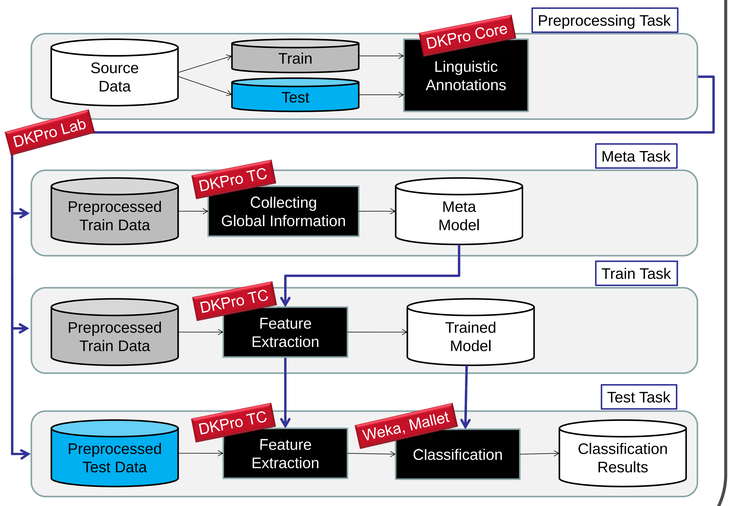

## Supported Learning Scenarios in DKPro TC

| | Single-label (Label) | Multi-label (Label, Label, Label) | Regression (0.75) |
|---|---|---|---|
| **Document Mode** | Spam Detection<br>Sentiment Detection | Text Categorization<br>Keyphrase Assignment | Text Readability |
| **Unit/Sequence Mode** | Named Entity Recognition<br>Part-of-Speech Tagging | Dialogue Act Tagging | Word Difficulty |
| **Pair Mode** | Paraphrase Identification<br>Textual Entailment | | Text Similarity |

## Architecture

**DKPro TC is built on top of several open-source NLP frameworks**

- *Apache UIMA (*Document Representation and Type Systems)
- *DKPro Lab (*Workflow Engine and Parameter Sweeping)
- *DKPro Core* (Linguistic Preprocessing for several languages)
- *Weka*, *Meka*, *Mallet (*Machine Learning)




## Example: Sentiment Detection on Tweets

**Example Experiment Configuration:** 10-fold cross-validation

```
BatchTaskCrossValidation batchTask = [
   experimentName: "Twitter Sentiment",
   preprocessingPipeline: createEngineDescription(ArkTweetTagger),
   parameterSpace: [// multi-valued parameters will be swept
      Dimension.createBundle("reader", [
         readerTrain: LabeledTweetReader,
         readerTrainParams: [LabeledTweetReader.PARAM_CORPUS_PATH, "tweets.txt"]]),
      Dimension.create("featureMode", "document"),
      Dimension.create("learningMode", "singleLabel"),
      Dimension.create("featureSet", [EmoticonRatioExtractor.name,
                                      NumberOfHashTagsExtractor.name]),
      Dimension.create("dataWriter", WekaDataWriter.name),
      Dimension.create("classificationArguments", [NaiveBayes.name], [RandomForest.name])],
      reports: [BatchCrossValidationReport], // collects results from folds
      numFolds: 10];
```

**Example Feature Extractor:** extracts the ratio of emoticons to all tokens

```
public class EmoticonRatioExtractor
   extends FeatureExtractorResource_ImplBase
   implements DocumentFeatureExtractor
{
   @Override
   public List<Feature> extract(JCas annoDb)
      throws TextClassificationException
   {
      int nrOfEmoticons = JCasUtil.select(annoDb, EMO.class).size();
      int nrOfTokens = JCasUtil.select(annoDb, Token.class).size();
      double ratio = (double) nrOfEmoticons / nrOfTokens;
      return new Feature("EmoticonRatio", ratio).asList();
   }
}
```


This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" (LOEWE) as part of the research center "Digital Humanities". The authors would like give special thanks to Richard Eckhart de Castilho, Nicolai Erbs, Lucie Flekova, Emily Jamison, Krish Perumal, and Artem Vovk for their contributions to the DKPro TC framework.


**http://dkpro-tc.googlecode.com**

VolkswagenStiftung

LOEWE Exzellente Forschung für Hessens Zukunft