NLP4WEB PROJECT TwitterSherlock

Group 1
Fabian Otto, Clemens Biehl, Philipp Kapelle, Daniel Wehner

RESEARCH PROBLEM

- Analyze tweets from authors of different domains (e. g. news, sports, politics, …)
- Example use case:
  - Identify terrorists based on their communication/language style

- Baseline: Random classification (n authors → Accuracy 1/n)

APPROACH

- Use Twitter API to obtain the raw data because it offers a variety of tweets from different domains (or use ANC corpus in addition)
- Preprocessing of raw data using DKPro (Tokenization, POS-Tagging, Normalization, Chunking, etc.)
- Data exploration and extraction of domain-specific features based on style of writing (vocabulary, linguistic style, grammar, emoticons, hashtags, etc.)
- Train several state-of-the-art classifiers to perform the classification task (e. g. CRFSuite conditional random fields and neural networks) and tuning of parameters, feature and model selection
- Evaluate the classifier(s) using ROC-Curves, Confusion Matrix, Precision, Recall (industry-standard evaluation methods)

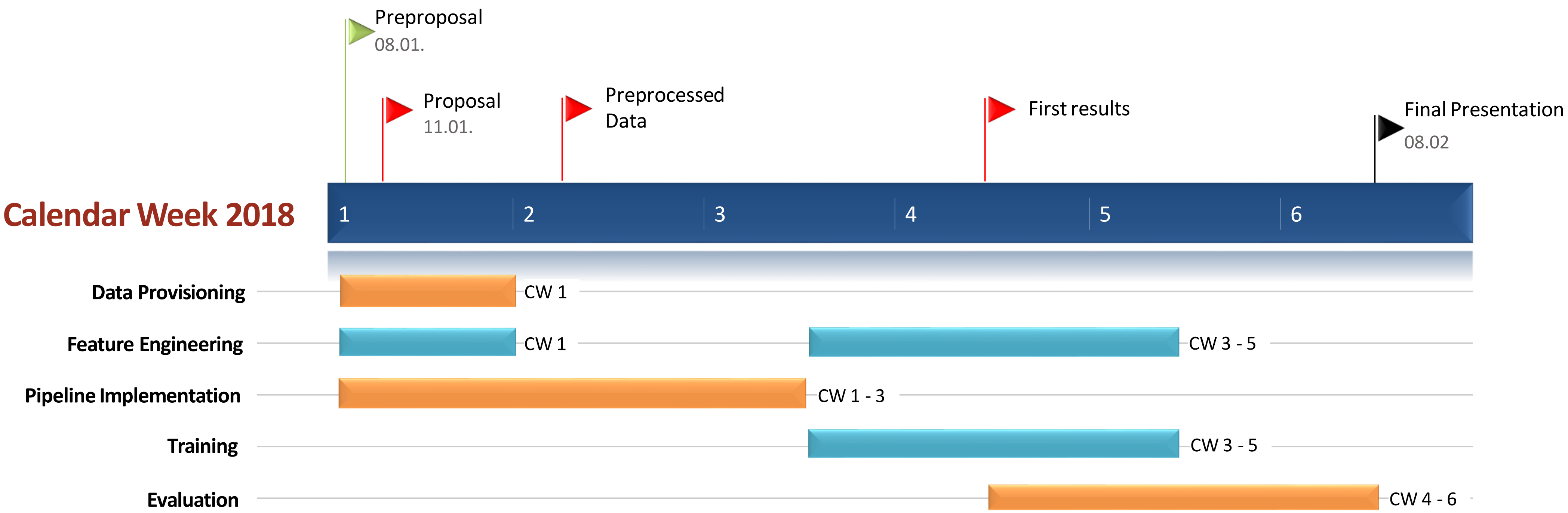
FRAMEWORKS + RESOURCES

- Frameworks, e. g.
  - DKPro (Preprocessing of raw data)/DKPro TC
  - Weka
  - CRFSuite
  - DeepLearning4J (experimental)
  - Jsoup
- Resources
  - Twitter API https://developer.twitter.com/en/docs/tweets/search/overview
  - News Authorship Identification with Deep Learning https://cs224d.stanford.edu/reports/ZhouWang.pdf
  - (ILSP Focused Crawler) http://nlp.ilsp.gr/redmine/projects/ilsp-fc/wiki/Getting_Started
  - (Crawler4J) https://github.com/yasserg/crawler4j/

ORGANIZATION

- Data Provisioning (crawler, corpus) → **Philipp Kapelle**
- Data Preprocessing (feature extraction, data format) → **Fabian Otto**
- Analysis (train and choose models) → **Clemens Biehl, Daniel Wehner**
- Evaluation → **Clemens Biehl, Daniel Wehner**

TIMELINE



Preproposal
08.01.

Proposal
11.01.

Preprocessed Data

First results

Final Presentation
08.02

Calendar Week 2018

1 2 3 4 5 6

Data Provisioning — CW 1

Feature Engineering — CW 1, CW 3 - 5

Pipeline Implementation — CW 1 - 3

Training — CW 3 - 5

Evaluation — CW 4 - 6

THANK YOU VERY MUCH

NLP4WEB

Group 1
Fabian Otto, Clemens Biehl, Philipp Kapelle, Daniel Wehner