

Topic-Specific Sentiment Analysis for Twitter Data of German MPs

Asmik Nalmpatian, Lisa Wimmer

Statistical Consulting

2020

Project Outline

Goal: topic-specific sentiment analysis on tweets by German MPs

→ What sentiments are expressed toward particular topics?



Two-pronged approach:

1. Comprehensive pipeline :
data extraction > data processing > sentiment analysis
 - Robust framework with rather basic methodology
 - To be implemented in R
2. Advanced sentiment analysis
 - Exploration of more complex methods
 - Most probably in Python

Key Challenges

- **Large data base of unstructured text**
 - High dimensionality
 - Run time, memory
- **German language**
 - Syntactic complexities
 - Less existing research than for English
- **Political context**
 - Specific issues/vocabulary
 - Semantic vehicles such as sarcasm, rhetorical questions
- **Twitter idiosyncrasies**
 - Short document length (140 characters)
 - Informal language (plus spelling mistakes)
 - Special features (hashtags, emojis, ...)
- **Topic extraction**
 - Upstream task where same challenges are present
- **No labels**
 - No means of evaluation with data as-is

Ideas

- **First & foremost:** labels
→ Necessary for measuring performance evaluation
- **Then:** classification in three levels of complexity
 1. **Dictionary approach**
 - Baseline model
 - n-grams, bag-of-words assumption
 - Probably low accuracy
 2. **Classic ML models**
 - Focus on feature extraction
 - Tried-and-tested classifiers (RF, SVM, ...)
 3. **BERT and friends**
 - Black-box, high-complexity approaches
 - Hope: data in, magic out
- **Eventually:** we know more about...
 - ... how far we can get with basic to medium approaches
 - ... by how much we can boost accuracy with adding complexity