Language-Independent Twitter Sentiment Analysis

DAI-Labor
Distributed Artificial Intelligence Laboratory

Sascha Narr
sascha.narr@dai-labor.de

Michael Hülfenhaus
michael.huelfenhaus@dai-labor.de

Prof. Dr.-Ing. Sahin Albayrak
sahin.albayrak@dai-labor.de

KDML 2012, LWA, Dortmund, Germany

Sentiment Analysis ...

- Detects emotion and opinions in text
- Polarity classification: Decides whether a text is positive or negative

... on Twitter

- 400+ million tweets per day
- Less than 40% of tweets are English
- Multilingual sentiment detection can cover a greater share of opinions

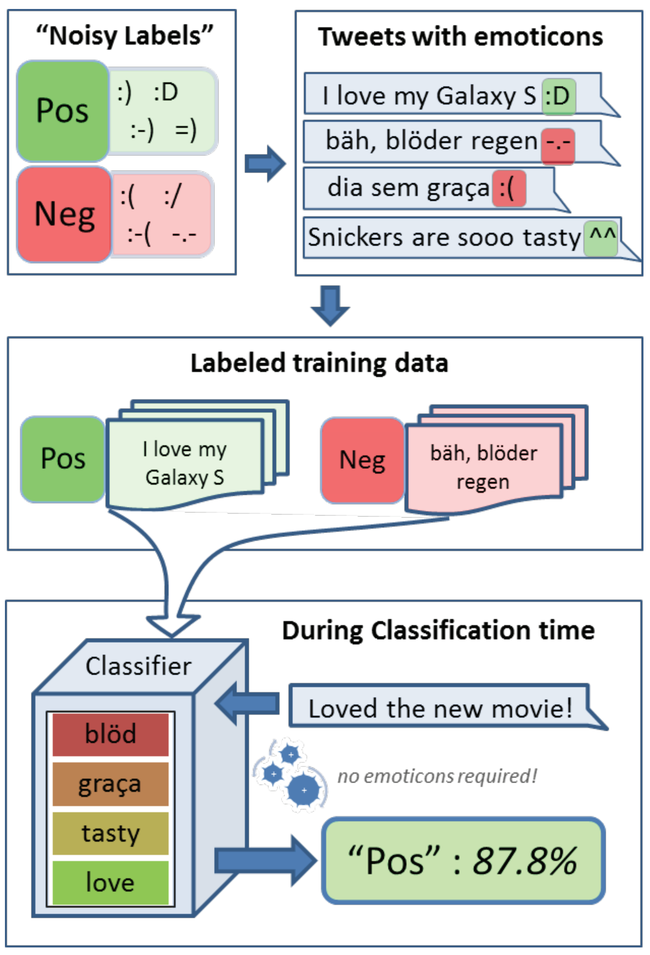
Experiments

We used millions of random tweets labeled by our emoticon heuristic as training data. With this data we trained Naive Bayes classifiers using n-gram features.

We trained one classifier each for tweets of one language, and one combined classifier for 4 languages. We tested the classifiers on the polarity classification task using our sentiment evaluation dataset.

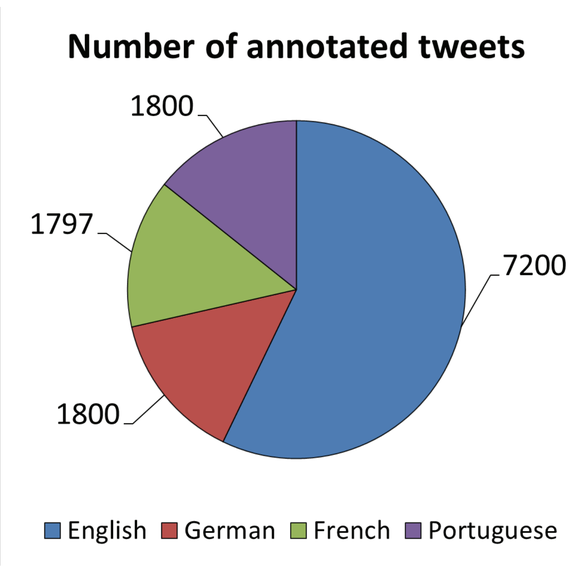
Language Independent Emoticon Heuristic

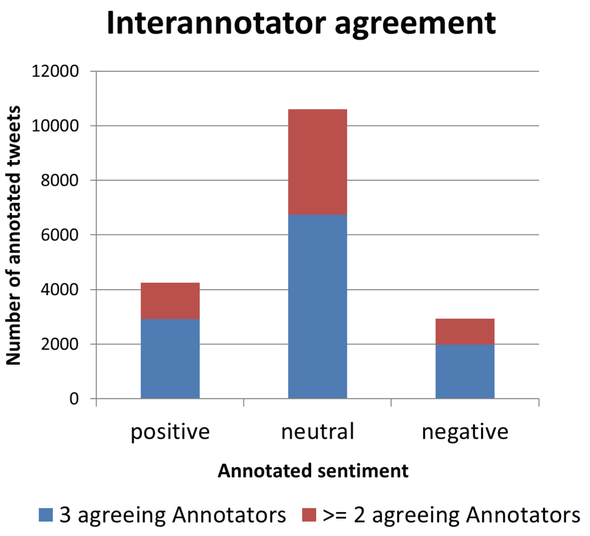
- Generates sentiment training data from tweets of any language
- Uses emoticons as “noisy label” sentiment indicators



Sentiment Eval. Dataset

- 12000+ human-annotated tweets
- Mechanical Turk - labeled **sentiment**: `positive`, `neutral`, or `negative`
- 4 languages: en, de, fr, pt
- Dataset publicly available (see bottom)

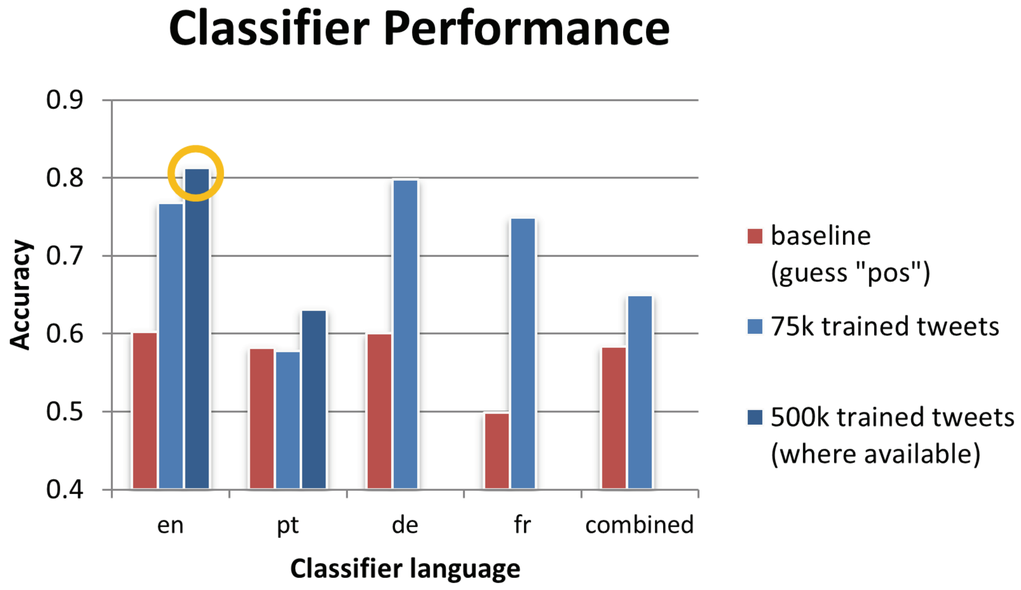




Evaluation and Results

- Classifiers were evaluated using different amounts of training tweets
- Baseline: classifier that always guesses “positive”

- Best result: **81.3% accuracy** for English, using unigrams
- Combined classifier for 4 languages is almost as accurate as the individual ones



Conclusions

- Our method is efficient without human supervision
- Good classification performance; can vary for different languages
- Publicly available evaluation dataset (sentiment-tagged tweets)

Scan to continue ...

Get the Dataset!

Read the full paper!