Martin Körner, Univ. Koblenz-Landau, mkoerner@uni-koblenz.de
Tatiana Sennikova, GESIS* & Univ. Koblenz-Landau
Florian Windhäuser, Univ. Koblenz-Landau, fwindhaeuser@uni-koblenz.de
Claudia Wagner, GESIS* & Univ. Koblenz-Landau
Fabian Flöck, GESIS*
*firstname.lastname@gesis.org

gesis
Leibniz Institute for the Social Sciences

Wikiwhere: An Interactive Tool for Studying the Geographical Provenance of Wikipedia References

Motivation

- Geographical origins of external web sources referenced in Wikipedia articles can be obscure
- Relevance: Lang. eds. can be biased towards preferring sources from certain geopolitical areas
  (E.g. the German version of an article shows an imbalance in favor of references from Russian vs. Ukrainian media, as opposed to the English version)
- Comparing different language versions on the same topic could provide insights, but is hard to do by hand
- Using only IPs to geolocate web sources seems to be too simplistic given the variety of hosting solutions

Approach

Feature Extraction
- IP-Location
- Website language
- Top Level Domain location

Ground Truth Generation
- Geo-Coordinates of Websites extracted from dbpedia.org

Machine Learning
- Best performance with Support Vector Machines (SVMs)

Data Visualization:
- Heatmap showing the predicted locations
- Bar charts and table for absolute numbers

Analysis

- IP-Location: Python geoip2 with GeoLite2 data
- Website language: Python html2text for text extraction and Python langdetect for language detection
- Top level domain: Info on country TLDs from IANA and CIA factbook
- Website locations (ground truth): SPARQL queries on dbpedia.org, manually assessed on partial set
- Total of 233,932 analyzed URLs from 9 DBpedia language endpoints
- Comparison of different machine learning models: logistic regression, random forests, and support vector machines (SVMs)

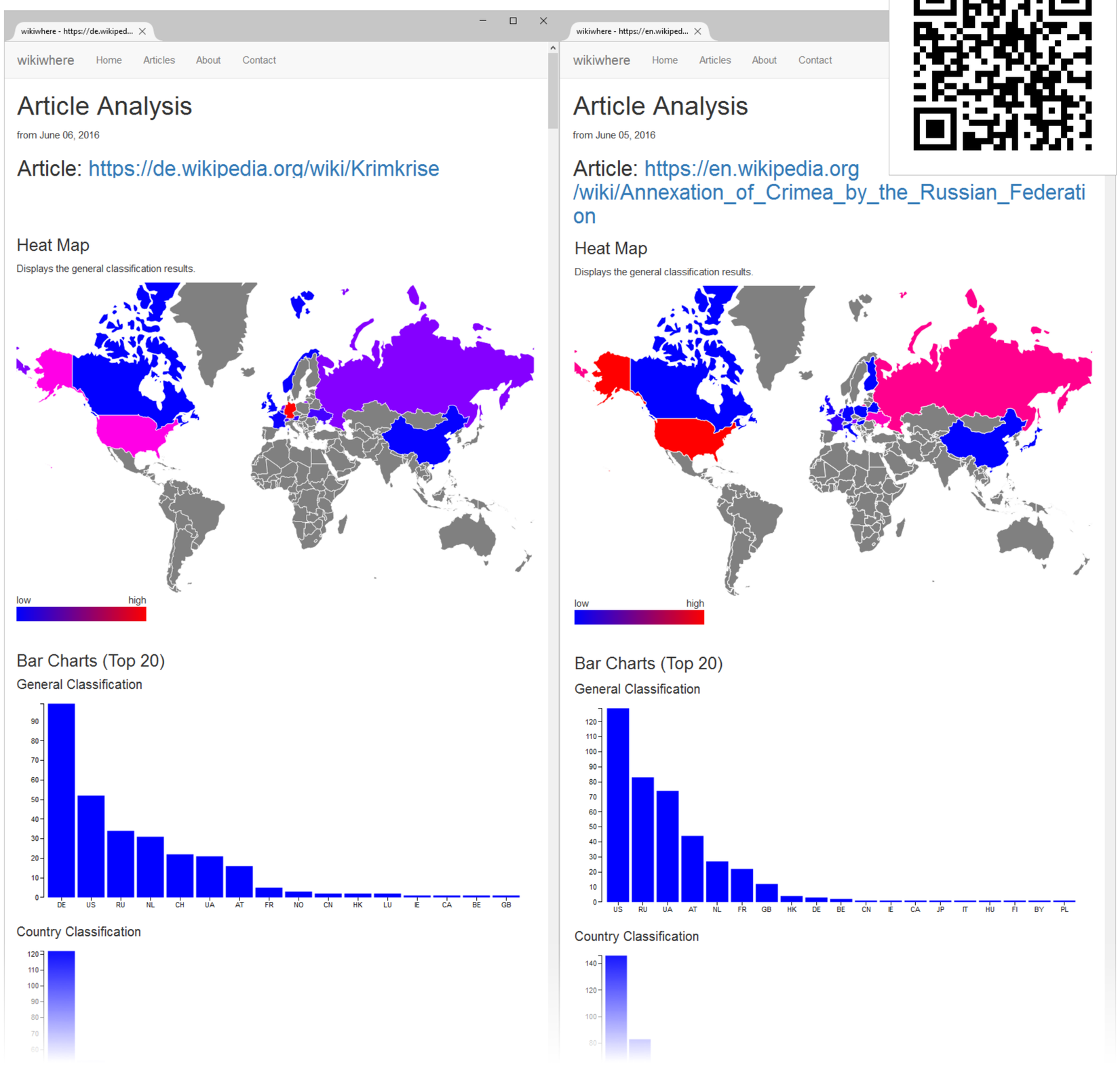
Evaluation

	Accuracy									
Method	**General**	**EN**	**FR**	**DE**	**ES**	**UK**	**IT**	**NL**	**SV**	**CS**
All data - Model	**81%**	**81%**	**91%**	**90%**	**75%**	**96%**	**91%**	**96%**	**92%**	**98%**
All data - IP only (Baseline)	61%	30%	62%	77%	29%	86%	73%	86%	81%	80%
Difficult cases - Model	**77%**	**78%**	**86%**	**80%**	71%	**89%**	**85%**	**91%**	**85%**	**93%**
Difficult cases - IP only (Baseline)	30%	57%	64%	25%	**81%**	66%	80%	74%	79%	53%

- Not perfect, but **notably** better than only using IP

Web Application

- Accessible at: **http://wikiwhere.west.uni-koblenz.de**
- User can insert URLs to any Wikipedia article
- Statistics are shown as a world heatmap, with bar charts, and in a table with individual results
- Responsive design to allow side-by-side comparison of different articles



M. Körner, T. Sennikova, F. Windhäuser, C. Wagner, F. Flöck. ***Wikiwhere: An Interactive Tool for Studying the Geographical Provenance of Wikipedia References***, *3rd Computational Social Science Winter Symposium 2016*