FETT: From **E**mbeddings **t**o **T**ransformers

course overview and introduction

Let’s get to know each other

who we and who you

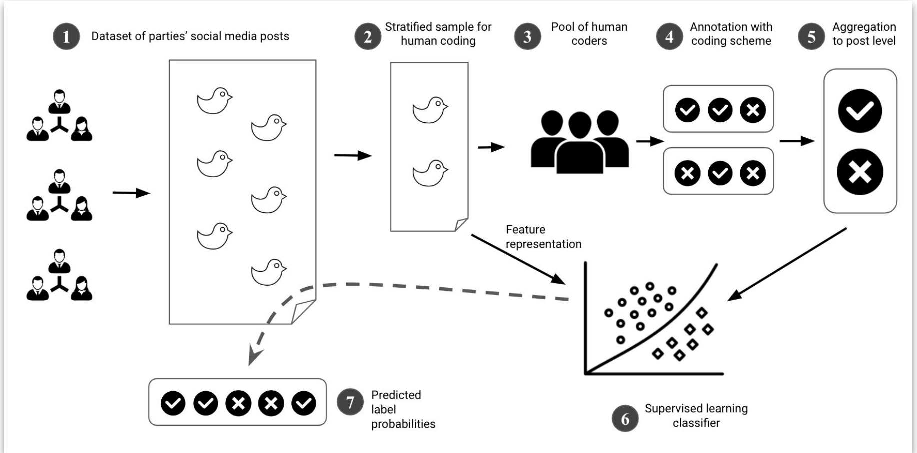
- Hauke is postdoc at U Cologne and interested in elites’ strategic use rhetoric in politics and multilingual text analyses
- Jennifer is a PhD candidate at ETH Zurich interested in human-AI interaction

Hauke

- **postdoc at the *University of Cologne* (PolSci) since 2022**

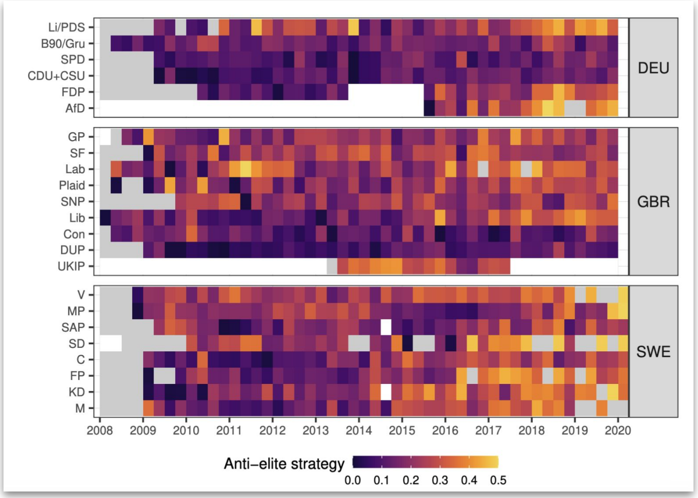
Hauke

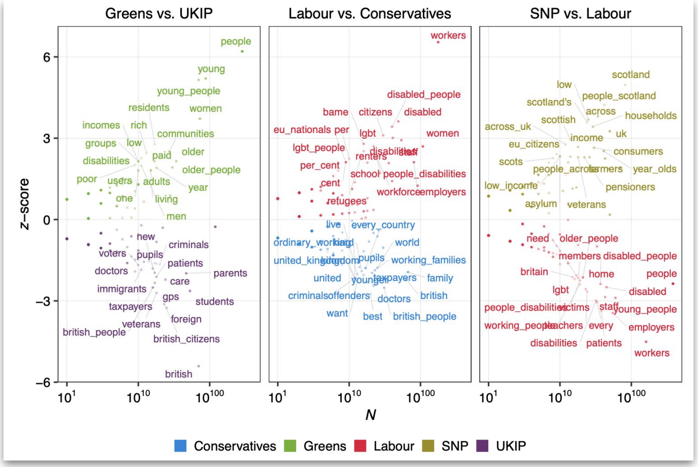
- postdoc at the *University of Cologne* (PolSci) since 2022
- **got interest in NLP when wanting to classify 500K tweets written in ~16 different languages (paper)**


1 Dataset of parties' social media posts
2 Stratified sample for human coding
3 Pool of human coders
4 Annotation with coding scheme
5 Aggregation to post level
Feature representation
6 Supervised learning classifier
7 Predicted label probabilities

Hauke

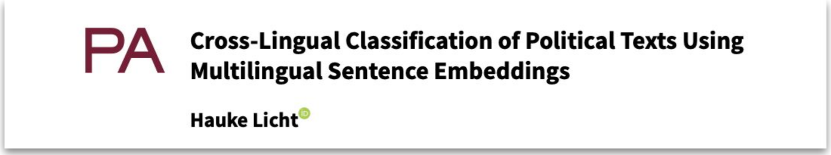
- postdoc at the *University of Cologne* (PolSci) since 2022
- got interest in NLP when wanting to classify 500K tweets written in ~16 different languages (paper)
- **interested in**
  - **elites' strategic use of political rhetoric (anti-elite rhetoric, group-based appeals)**





Hauke

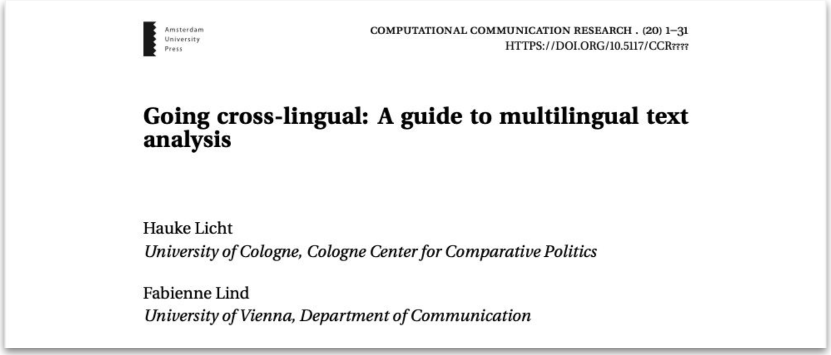
- postdoc at the *University of Cologne* (PolSci) since 2022
- got interest in NLP when wanting to classify 500K tweets written in ~16 different languages (paper)
- **interested in**
    - elites’ strategic use of political rhetoric (anti-elite rhetoric, group-based appeals)
    - **multilingual text analysis**



PA

Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings

Hauke Licht



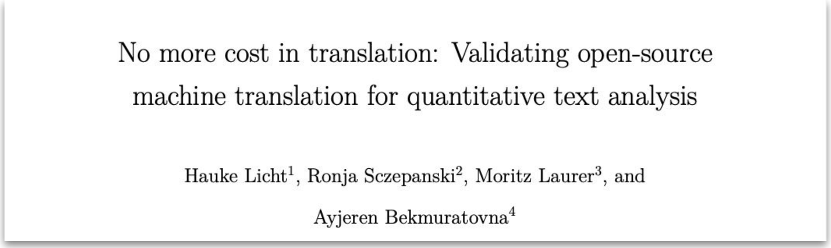
Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH . (20) 1–31
HTTPS://DOI.ORG/10.5117/CCR????

Going cross-lingual: A guide to multilingual text analysis

Hauke Licht
University of Cologne, Cologne Center for Comparative Politics

Fabienne Lind
University of Vienna, Department of Communication



No more cost in translation: Validating open-source machine translation for quantitative text analysis

Hauke Licht[1], Ronja Sczepanski[2], Moritz Laurer[3], and Ayjeren Bekmuratovna[4]

Jennifer

- PhD candidate at ETH Zurich
- BA & MA in Political Science from University of Zurich

Jennifer

- PhD candidate at ETH Zurich
- BA & MA in Political Science from University of Zurich
- **got interested in NLP / computational social science during my Masters**

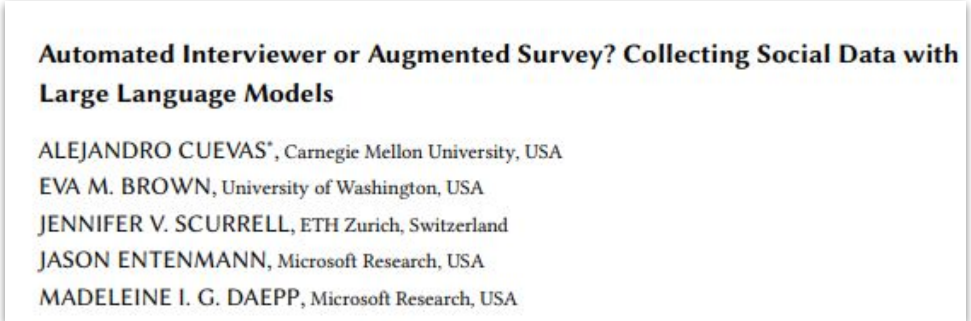
University of Zurich^UZH

Master's thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of
Master of Arts UZH in Social Sciences

Visual Party Communication:

Political Image Analysis with Deep Learning

Jennifer



Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models

ALEJANDRO CUEVAS*, Carnegie Mellon University, USA
EVA M. BROWN, University of Washington, USA
JENNIFER V. SCURRELL, ETH Zurich, Switzerland
JASON ENTENMANN, Microsoft Research, USA
MADELEINE I. G. DAEPP, Microsoft Research, USA

- PhD candidate at ETH Zurich
- BA & MA in Political Science from University of Zurich
- got interested in NLP / computational social science during my Masters
- **dissertation: impact of bots on political opinion**
- **research interests: human-AI interaction, LLM prompt engineering & red teaming, responsible AI**

Jennifer

- PhD candidate at ETH Zurich
- BA & MA in Political Science from University of Zurich
- got interested in NLP / computational social science during my Masters
- dissertation: impact of bots on political opinion
- research interests: human-AI interaction, LLM prompt engineering & red teaming, responsible AI
- **just came back from Microsoft**

Why FETT?

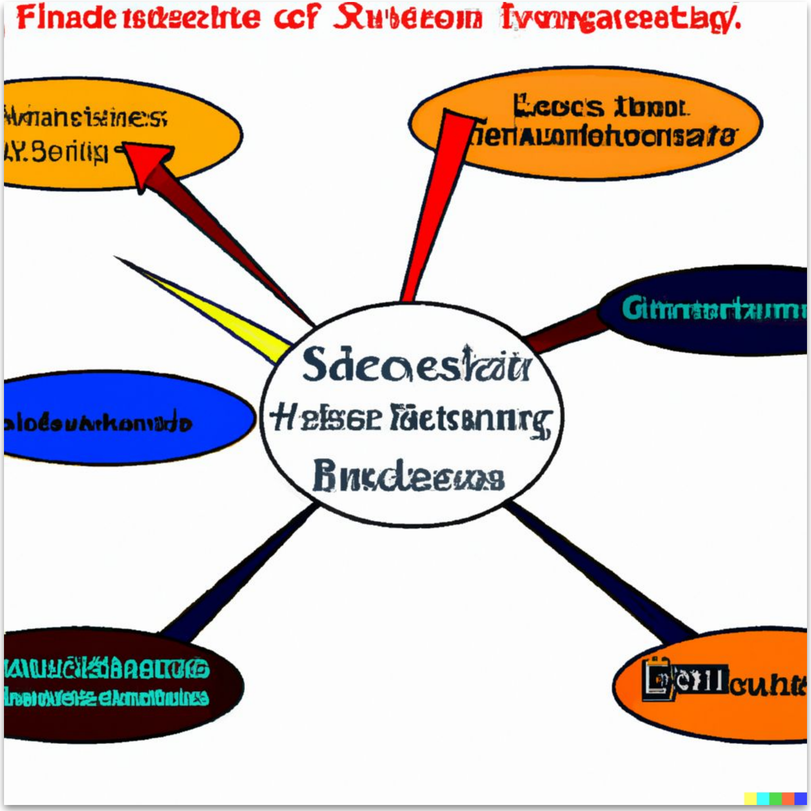
What you'll learn

- first classic word embeddings
- then Transformer models
- closing with outlook on LLMs

Computational Social Science

Goals and opportunities

- **common goals** with traditional social sciences: study social, political, and cultural phenomena
  - describe ⇒ measurement
  - explain ⇒ (causal) inference
  - predict ⇒
- but **big data** generally requires new methods and approaches



Computational Social Science

Computing with text data

text provides good raw material to learn about social and political behaviors

- abundance of text ⇒ manual, qualitative reading impossible
- *raison d’etre* of computational text analysis

But we need *numeric* representations!!!

- to compare text units (change over time, differences between “authors”)
- to perform inductive or deductive “downstream” tasks (e.g., clustering or classification)

btw: independent of whether the unit for analysis is the corpus, document, or word

Representing text with numbers

Counting words

bag-of-words representations have clear limitations

- no info about words' relations
- no contextualization of word meaning
- high-*d* ⇒ costly computation
- sparsity limits generalization

Embedding words

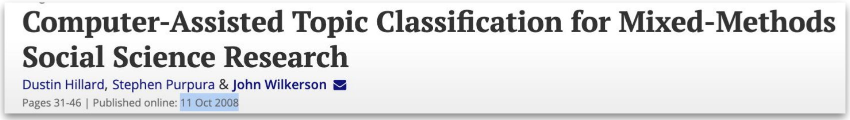
(neural) **text embedding methods** address these limitations

- word embeddings capture similarities in words' meaning and function
- Transformers' attention mechanism enables contextualized word representation
- transfer learning makes analyses and computation more efficient

Computational literacy

Methods diffusion changes CSS

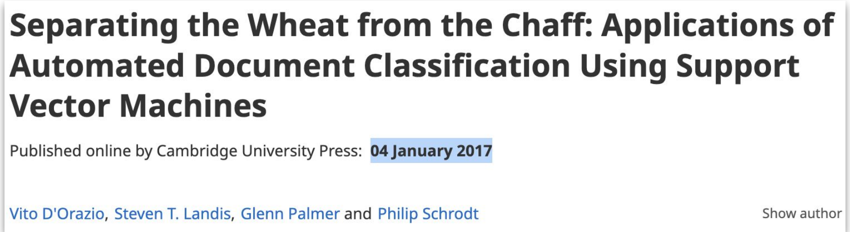
- increasing adoption of innovations from CS and NLP in applied CSS research
- known and understanding these methods
  - ✓ (potentially) better leverage and new angles in your research
  - ✓ critical evaluation of research
  - ✓ comparative advantage in job market
  - ✓ facing upcoming transformations with greater resilience



Computer-Assisted Topic Classification for Mixed-Methods Social Science Research

Dustin Hillard, Stephen Purpura & John Wilkerson

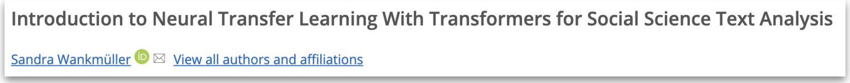
Pages 31-46 | Published online: 11 Oct 2008



Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines

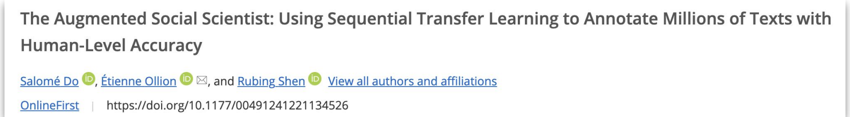
Published online by Cambridge University Press: 04 January 2017

Vito D'Orazio, Steven T. Landis, Glenn Palmer and Philip Schrodt



Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis

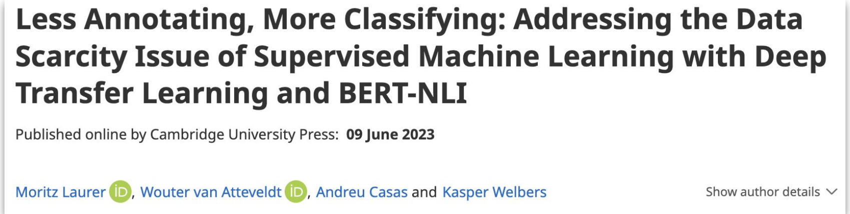
Sandra Wankmüller



The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy

Salomé Do, Étienne Ollion, and Rubing Shen

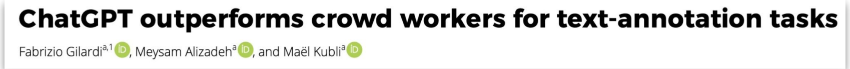
OnlineFirst | https://doi.org/10.1177/00491241221134526



Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI

Published online by Cambridge University Press: 09 June 2023

Moritz Laurer, Wouter van Atteveldt, Andreu Casas and Kasper Welbers



ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi[a,1], Meysam Alizadeh[a], and Maël Kubli[a]

Day-by-day Schedule

What you’ll learn

- first classic word embeddings
- then Transformer models
- closing with outlook on LLMs

Word embedding methods and analyses

- Day 1
    - motivation and intuition
    - computing with word embeddings (similarity, nearest neighbors, analogies)
- Day 2
    - computing social *scientifically-relevant quantities* (implementation of Caliskan *et al.* 2017, Kozlowski *et al.* 2019, and Gennaro & Ash 2022)
    - detailed explanation of `word2vec`
    - training from scratch and fine-tuning embedding models
- Day 3 *(morning)*
    - limitations of (static) word embeddings

Transformer models and applications

- Day 3 *(afternoon)*
  - contextualized word embeddings
  - conceptual intro to transformers (yep, some dry theory :))
- Day 4
  - transformers in the social sciences
  - about training and tuning
  - masked language models (like BERT)
  - exercises with Hugging Face transformers
- Day 5
  - BERTopic
  - input: Large Language Models
  - ethics
  - course recap / Q & A / 1-on-1 meetings

Python

Setup and trouble-shooting

- use the setup instructions on GitHub
- we’ll do trouble-shooting in the afternoon if needed

Python

we will use Python

- lots of time lag in implementation of state-of-the-art NLP methods in R
- most advanced stuff runs in Python in the background anyways

all code and materials are at on Github

https://github.com/haukelicht/advanced_text_analysis

Python setup

We have setup instruction on Github (see our mails from last week)

BUT DON'T PANIC!! But we'll only start coding in the afternoon

If you had setup issues, let us know before the lunch break – we'll help you trouble-shooting