Statistical Natural Language Processing

Part X: Practical Issues

Henning Wachsmuth

https://ai.uni-hannover.de

Learning Objectives

Concepts

- Domain dependency
- Social bias
- Dual use in NLP

Methods

- Estimation of training data need
- General effectiveness tweaks
- Domain adaptation
- Social bias detection
- Social bias mitigation

Notice

- The issues covered here complement those in *Introduction to NLP*.
- To be self-contained, some overlap exists between the two parts.

Outline of the Course

I. Overview

II. Basics of Data Science

III. Basics of Natural Language Processing

IV. Representation Learning

V. NLP using Clustering

VI. NLP using Classification and Regression

VII. NLP using Sequence Labeling

VIII. NLP using Neural Networks

IX. NLP using Transformers

X. Practical Issues

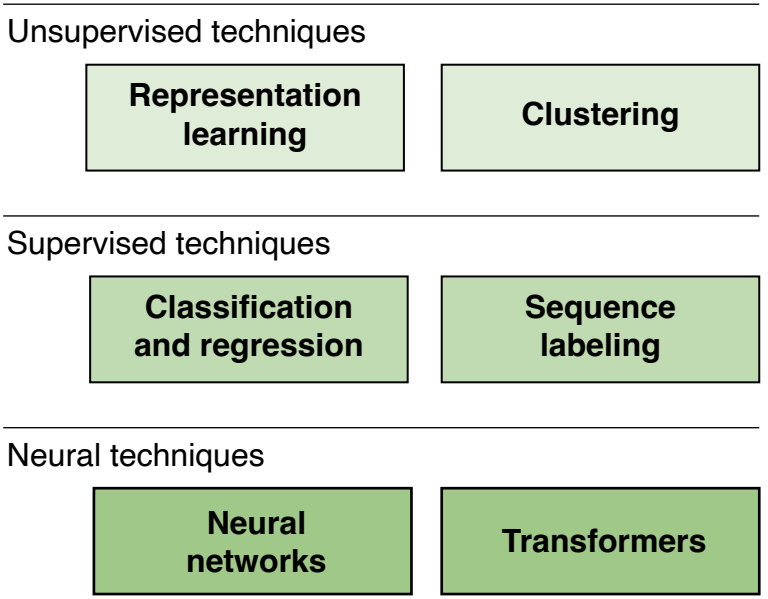
- Introduction
- Reliability Issues
- Robustness Issues
- Ethical Issues

Introduction

Practical Issues

NLP techniques (recap)

- Fully unsupervised techniques
- Fully supervised techniques
- Neural techniques, combining ideas from both



Going to the "real world"

- How to develop approaches that use these techniques in practice?
- What differences exist, when NLP is applied *in the wild*.
- What issues to take care of?

Practical issues in NLP

- Effectiveness. Limited reliability and robustness of the methods
- Efficiency. Too high run-time, energy, and/or memory consumption
- Ethics. Concerns and dangers with the behavior and use of methods

Practical Issues

Issues Covered Here

Effectiveness: Reliability issues

- Approaches may have a too low (mean) effectiveness for reliable use.
- Main reasons include error accumulation, lack of data, and the general problem complexity.

Effectiveness: Robustness issues

- Approaches may not work robustly across different types of input texts.
- Main reasons include unforeseen inputs, adversarial user behavior, and domain transfer.

Ethical issues

- Approaches may show ethically problematic behavior, including hallucinations and social bias.
- Approaches may be employed in doubtful ways.

Practical Issues

Efficiency Issues

Reasons for limited efficiency

- NLP pipelines often include several time-/energy-intensive methods.
- Large amounts of data may need to be processed, possibly repeatedly.
- Much information may be stored during processing.

Ways to improve run-time and energy efficiency

- Indexing of relevant information
- Resort to simpler and/or efficiency-optimized NLP methods
- Filtering, scheduling, and parallelization in NLP processes

  Details on approaches to efficiency are discussed in *Introduction to NLP*.

Ways to improve memory efficiency

- Scaling up is the natural solution to higher memory needs.
- Also, debugging (and minimizing) what information is stored may help.
- There may be a tradeoff between time and memory efficiency.

  Memory efficiency is often *not* the main problem.

Reliability Issues

Reliability Issues

Problem-related reasons for limited effectiveness

- Ambiguity of natural language

  "Death penalty — why not?" → Stance on death penalty?

- Missing context and world knowledge

  "I hope Biden will keep his attitude towards capital punishment." → And here?

Approach-related reasons for limited effectiveness

- Error accumulation. Errors propagate in a sequence of NLP methods.
- Lack of data. Training data may not suffice to learn an effective method.
- Need for Training. What amount is sufficient, is unclear in general.

  Details on the next slides

Application-related reasons for limited effectiveness

- Low robustness. Methods may fail on data different from training data.

  Details in the next section of this lecture part

Reliability Issues

Error Accumulation

Error accumulation

- When NLP methods are applied in sequence, errors propagate, since the output of one method serves as input to subsequent ones.
- Even when each method works well, overall effectiveness may be low.

  Ways to alleviate error propagation are discussed in *Introduction to NLP*.

Example: Review sentiment analysis

- Automatically classified vs. ground-truth local sentiment

  Subjectivity accuracy 78.1%, polarity accuracy 80.4%
- Root mean squared errors (RMSE) of sentiment score regression:

Feature type	Automatic	Ground-truth
Standard features	1.11	1.11
Local sentiment distribution	0.99	0.77
Discourse relation distribution	1.01	0.84
Sentiment flow patterns	1.07	0.86
Combination of features	**0.93**	**0.75**

Reliability Issues

Lack of Data (before LLMs)

No training data available?

- Hand-crafted rules are the only way to go.
- Careful human tuning on some validation set is important.

A small amount of training data?

- If any, use a “high-bias” learning algorithm, such as Naïve Bayes.
- Also, semi-supervised learning methods may help.

A sufficient amount of training data?

- Suitable for advanced feature-based learning algorithms, e.g., SVMs
- If interpretability is needed, decision trees should be preferred.

A huge amount of training data?

- SVMs and particularly neural networks may achieve high effectiveness.
- But algorithms such as Naïve Bayes scale much better.

  With enough data, the learning algorithm often matters less.

Reliability Issues

Lack of Data (now, with LLMs)

No training data available?

- Zero-shot learning. Train LLM on related task, add auxiliary information.
- Alternatively, an instruction-following LLM can simply be prompted.

A small amount of training data?

- Few-shot learning. Fine-tune pretrained LLM on a few selected cases.
- With instruction-following LLMs, this is often done directly in the prompt.

A sufficient amount of training data?

- Fine-tuning. Fine-tune pretrained LLM on task-specific training data.
- This also works for instruction-following LLMs.

A huge amount of training data?

- Training from scratch. Learn a new LLM on the whole training data.
- Starting with a pretrained LLM works, too.

  Training from scratch requires powerful computing facilities.

Reliability Issues

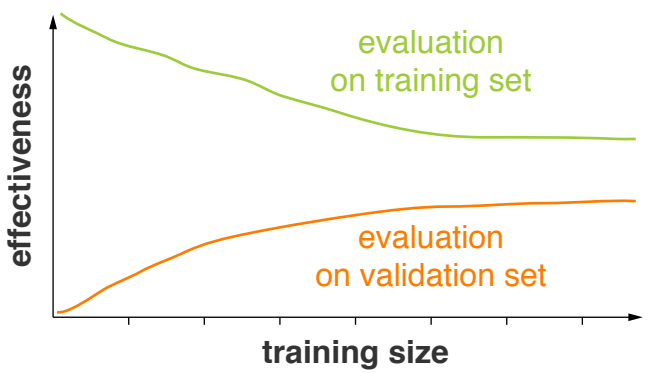
Need for Training

How much training data is sufficient?

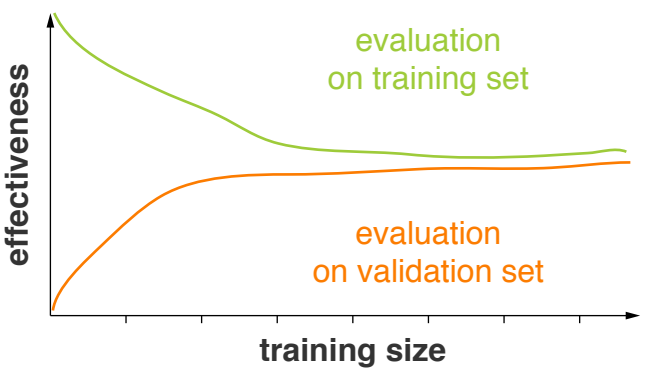
- In general, hard to say
- Depends on the complexity of the task, the heterogeneity of the data, ...

One way to find out

- Test different training sizes.
- Evaluate effectiveness on training set and on validation set.



(more data is likely to help)



(more data likely won't help)

- Validation effectiveness is unlikely to ever exceed training effectiveness.

Reliability Issues

Practical Effectiveness Tweaks

Exploiting domain knowledge

- Rule of thumb. The narrower the domain, the higher the effectiveness
- Encoding domain-specific knowledge is important in practice.
- In-domain training is often a must for high effectiveness.

Combining statistics and rules

- Real-world NLP applications mostly combine statistical learning with hand-crafted rules.
- Rules are derived from a manual review of uncertain and difficult cases.

Scaling up

- At large scale, precision can be preferred over recall, assuming that the information sought for appears multiple times.
- A smart use of redundancy increases confidence.

"In 1998, he founded Google"	"Google exists since '98"	"Google, estd. 1998"

Robustness Issues

Robustness Issues

Reasons for limited robustness

- The term *robustness* may refer to different aspects of NLP methods.
- We here mean robustness with respect to the given input.
- Main reasons. Unforeseen inputs, adversarial users, domain differences

Unforeseen inputs

- Methods are often trained on the expected type of input text only.
- Different styles, length, encodings, etc. may cause unexpected behavior

Adversarial users

- Users may test how they can trick the behavior of NLP methods.
- In times of prompting, this may also cause unwanted information leaks.

Domain differences (details below)

- NLP often needs to be applied to texts with unknown properties.
- Robustness means here that it is effective on texts across *domains*.

Robustness Issues

Domains and Domain Robustness

Domain (in NLP)

- A set of texts that share certain properties
- May refer to a topic, genre, style, language — or combinations
- Texts from a domain can be seen as being drawn from the same underlying feature distribution.

  This means that similar feature values imply similar output information.

Topics	Genres	Styles	Languages	...
Books	News articles	Formal	English	...
Movies	Forum posts	Informal	German	...
Hotels	Reviews	Opinionated	Mandarin	...
...	...	...	...	...

Reasons for limited domain robustness

- Learning domain-specific rules and features
- Learning from a biased dataset
- Learning a model with too much variance (i.e., overfitting)

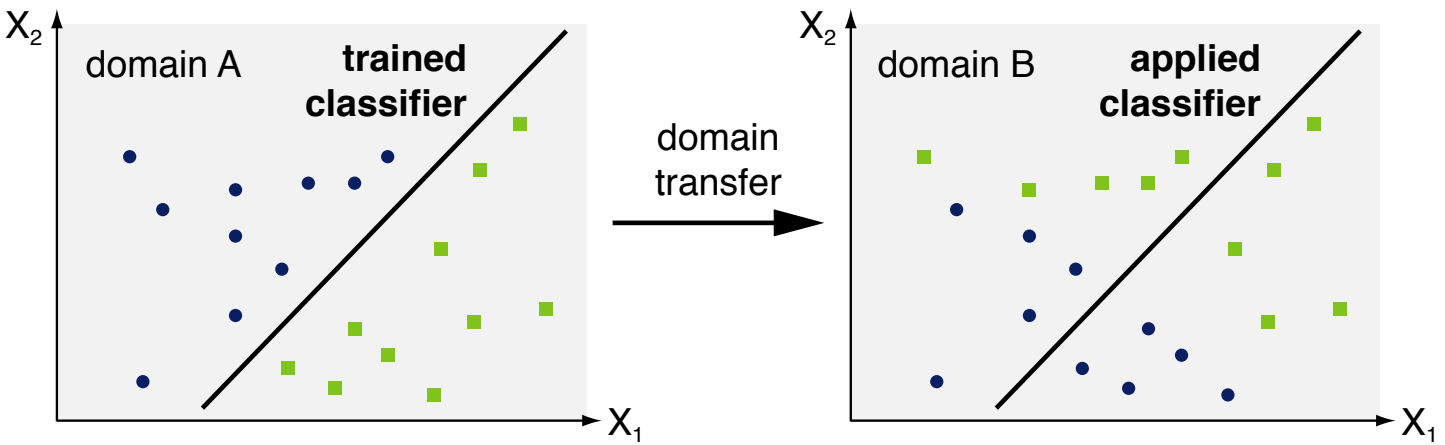
  Missing domain robustness is a fundamental problem of machine learning in general.

Robustness Issues

Domain Dependency

Domain dependency

- If a method works notably better in the domain of training texts than in others, it is said to be domain-dependent.



Differences between domains

- The same feature values result in different output information.
- Different features are discriminative regarding the target variable.

"Read the book" in book reviews vs. movie reviews vs. hotel reviews?

Robustness Issues

General Approaches to Domain Robustness

Heterogeneous training sets

- A simple way to make a method more robust is to train it on texts from multiple (notably different) domains.
- Avoids overfitting to domain-specific features
- In in-domain settings, typically worse than domain-specific methods

Domain-independent features

- For many tasks, there are features that behave similar across domains.

  unambiguous polarity indicators in sentiment analysis

  spaces in tokenization

- By focusing on such features, robustness can be improved.
- Features that model structure or style (rather than content) tend to be more domain-independent, but exceptions exist.
- The sentiment flow patterns from lecture parts V and VI follow this idea.

  See evaluation below.

Robustness Issues

Domain Adaptation

Domain adaptation

- Adjust a method trained in some source domain to some target domain.
- Requires at least a few training texts from the target domain
- Often based on *structural correspondences* of the domains

Learning of structural correspondences

- Identify features that work robustly across source and target domain.

  "horrible" is likely to be negative in every review.

- Find cooccurrences of domain-specific with domain-robust features.

  "Read the book!" occurs at the end of movie reviews with "horrible".

  "Stay away!" occurs at the end of hotel reviews with "horrible".

- Align domain-specific features based on learned correspondences.

  "Read the book!" in movie reviews ~ "Stay away!" in hotel reviews

Review Sentiment Analysis

Sentiment classification of reviews (recap)

- Classification of the nominal sentiment polarity or score of a customer review on a product, service, or work of art

Data

- 2100 English hotel reviews from TripAdvisor, scores $\in \{1, ..., 5\}$
  Below, 1–2 mapped to 0, 3 to 1, 4–5 to 2
- 5,006 English movie reviews from Rotten Tomatoes, scores $\in \{0, 1, 2\}$
  From the Cornell movie review dataset (Pang and Lee, 2005)

Tasks

- 3-class sentiment classification across domains

Approach

- Features. Same as in lecture part VI
- Algorithm. Linear SVM with one-versus-all multi-class handling
  Default cost hyperparameter value ($C = 1.0$) in cross-domain evaluation

Review Sentiment Analysis

Evaluation of Cross-Domain Classification

Evaluation

- SVMs trained on texts from one domain, tested on texts from the other
- Training in-domain vs. out-of-domain, testing on the same test set

  By comparing on the same test set, the “difficulty” of corpora is ruled out.

Effectiveness results (accuracy)

		Test on Hotel			Test on Movie		
Feature type	**Trained on:**	**Hotel**		**Movie**	**Movie**		**Hotel**
Standard features		58.9%	– 16.1	42.8%	63.8%	– 29.8	34.0%
Local sentiment distribution		69.8%	– 11.0	**58.8%**	52.7%	– 12.9	39.8%
Discourse relation distribution		65.3%	– 10.8	54.5%	52.3%	– 7.2	45.1%
Sentiment flow patterns		63.1%	– 6.0	57.1%	53.9%	– 8.6	**45.3%**
Combination of features		**71.5%**	– 22.7	48.8%	**64.0%**	– 21.7	42.3%

Observations

- Sentiment flow patterns have lowest effectiveness loss across domains.
- Full domain independence seems hard to achieve.

Ethical Issues

Ethical Issues

Ethics and NLP

- AI mimics human behavior, and its decisions often affect humans.
- The use of AI therefore generally brings up many ethical questions.
- NLP is particularly sensitive, as language is at the core of human communication and intelligence.

Selected ethical issues with the behavior of NLP

- Hallucinations. Creation of false or misleading information (more below)
- Social bias. Unfair treatment of certain social groups (more below)
- Privacy violations. Processing or mining of private information
- Non-explainability. Unclear decisions with impact on humans

Selected ethical issues with the use of NLP

- Misuse. Spread of misinformation, mass surveillance, social scoring, ...
- Dual use. Methods with a “good” purpose may be misused. (more below)
- Environmental impact. Training and using NLP requires much energy.

Ethical Issues

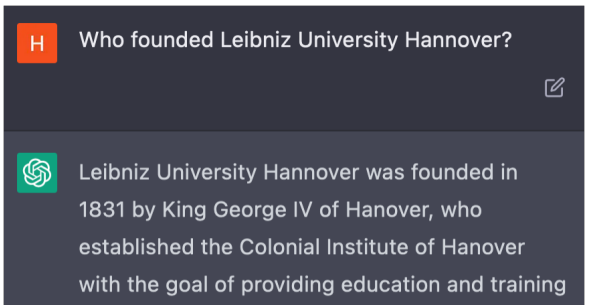
Hallucinations

Hallucination (in NLP)

- False or misleading text generated by NLP methods, such as LLMs
- The text may be nonsensical or unfaithful to available sources.
- Often originates in the basic idea of language modeling

Example: ChatGPT

- ChatGPT generates text from scratch.
- It has no mechanism to check veracity.
- Often, information may still be correct due to its enormous training data.



How to avoid hallucinations?

- Controlled text generation is the key (but far from trivial).
- Facts from knowledge bases may be integrated with free text.
- The model optimization process may include factuality criteria.

Social Bias

Social group

- A group of people defined by physical attributes (e.g., sex or skin color) or abstract concepts (e.g., culture and heritage)
- Some groups are protected by constitutional rights.

Ethnicities

Genders

Age groups

Social bias

- Stereotypes, prejudices, or discrimination against social groups
- Causes unfair treatment of the groups and their members

Social bias in NLP

- Social bias manifests in human communication and, thus, in language.
- NLP methods may process bias texts, and may reproduce the bias.

Social Bias

Detection

Social bias detection

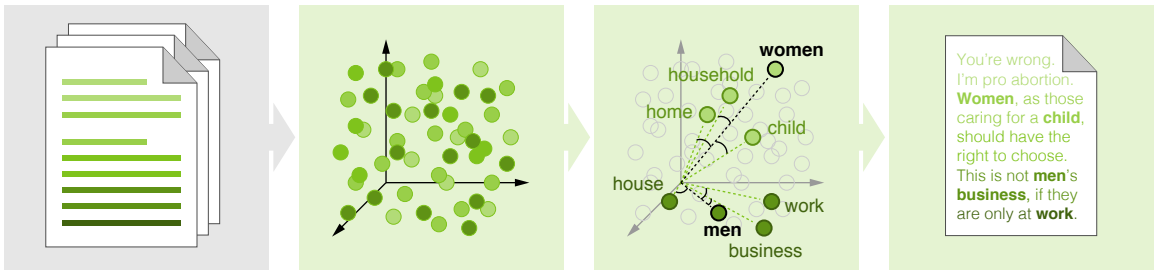
- NLP can be used to measure how biased a text corpus is.
- NLP can be used to identify biased spans in a text.

Word embedding association test (WEAT) (Caliskan et al., 2017)

- Find similarities between social groups and bias terms in embeddings.
- Map mean difference of the similarities to a bias score in [–2, 2].

Example: Genders vs. work/family terms (Spliethöver and Wachsmuth, 2020)

- Train word embedding models on different online debate corpora.
- Compute WEAT bias scores (and mine common co-occurrences).



Debate corpus | (a) Train embedding model | (b) Compute bias score | (c) Mine co-occurrences

Social Bias

Mitigation

Social bias mitigation

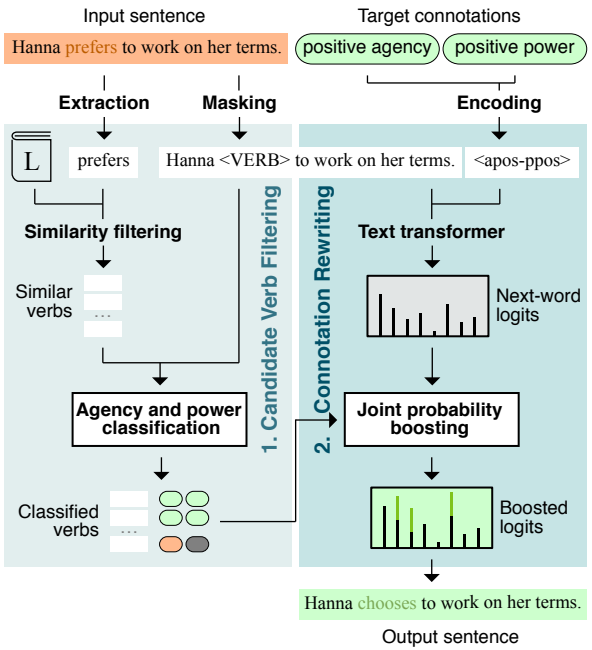
- NLP can be used to create *counterfactuals* for biased text spans.
- NLP methods can be adjusted to avoid producing bias.

Counterfactuals

- Methods trained on biased data may take on or reproduce the bias.
- Counterfactuals are instances added during training to counter the bias.

Hanna prefers to work on her terms.

→ Hanna chooses to work on her terms.



Example: Counterfactuals for gender bias

(Stahl et al., 2022)

- Set target verb connotations for counter.
- Create rewritten instance following them.

Ethical Issues

Dual Use

Dual use

- The misuse of methods for harmful (secondary) purposes besides their actual (primary) purposes
- Preventing such misuse in NLP is very hard in general.

Examples

- Text summarization can be modified to imply misleading conclusions.
- Sentiment analysis can be used to spot regime opponents.
- Bias mitigation can be misused to add bias, possibly with adjustments.

Original text (taken from right-oriented news article)

Implicit in the debate and the stalemate that left the bill to die when Congress adjourned was a recognition that the cost of immigration reform would be high, although no one knew how high. Without reform, though, the presence of what may be six million illegal aliens in this country exacts an economic and social toll.

Reframed text (with hallucinated quote)

"Illegal aliens' is a growing problem in the country," says a spokesman for the measure's sponsors. Without reform, though, the presence of what may be six million illegal aliens in this country exacts an economic and social toll.

(Chen et al., 2021)

Does a "good" purpose justify the risks?

- Better to make risks transparent than to let others exploit them secretly.

Conclusion

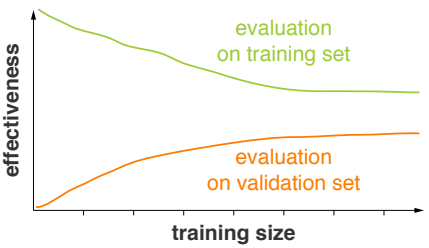
Summary

Practical issues

- Low effectiveness due to limited reliability and robustness
- Low efficiency due to expensive computations
- Ethical dangers due to imitation of human capabilites

Effectiveness issues

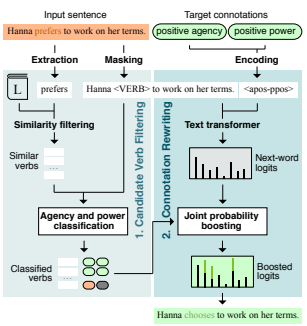
- Missing reliability often originates in lack of data.
- Recent LLM techniques reduce the need for data.
- Successful domain transfer is often challenging.



(more data is likely to help)

Ethical issues

- Avoiding hallucinations requires advanced techniques.
- Detecting and migitation social bias is often critical.
- Misuse of NLP methods is hard to avoid in general.



References

Some content and examples taken from

- **Chen et al. (2021).** Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. Controlled Neural Sentence-Level Reframing of News Articles. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2683–2693, 2021.
- **Jurafsky and Manning (2016).** Daniel Jurafsky and Christopher D. Manning (2016). Natural Language Processing. Lecture slides from the Stanford Coursera course. https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html.
- **Spliethöver and Wachsmuth (2020).** Maximilian Spliethöver and Henning Wachsmuth. Argument from Old Man?s View: Assessing Social Bias in Argumentation. In Proceedings of the 7th Workshop on Argument Mining, pages 76–87, 2020.
- **Stahl et al. (2022).** Maja Stahl, Maximilian Spliethöver, and Henning Wachsmuth. To Prefer or to Choose? Generating Agency and Power Counterfactuals Jointly for Gender Bias Mitigation. In Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 39–51, 2022.
- **Wachsmuth (2015).** Henning Wachsmuth (2015): Text Analysis Pipelines — Towards Ad-hoc Large-scale Text Mining. LNCS 9383, Springer.

References

Other references

- **Caliskan et al. (2017).** Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. Science, 356(6334):183–186, 2017.
- **Pang and Lee (2005).** Bo Pang and Lillian Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pages 115–124, 2005.