

MEASURING OPINIONATED COMMUNICATION

Dictionaries, Machine Learning, LLMs: A framework for deciding which approach to use for computational text analysis

Table 1: Computational approaches to text analysis

Method	examples
dictionary-based methods	LIWC, Vader
supervised machine learning	logistic regression, Naïve Bayes, random forest
unsupervised machine learning	PCA, LDA
deep learning	convolutional neural networks (CNN), long short-term memory (LSTM)
large language model (LLM)	BERT, BERT-NLI
generative large language model (gLLM)	ChatGPT, Llama

Table 2: Showstoppers

Potential show stopper	dict.	SML	UML	DL	LLM	gLLM	
						local	API
no (large) annotated dataset available	✓	✗	✓	✗	✓	✓	✓
no language-specific tool available	✗	✓	✓	✓	✗	✗	✗
categories are known a-priori	✓	✓	✗	✓	✓	✓	✓
categories are not known a-priori	✗	✗	✓	✗	✓	✓	✓
no access to high-performance computing	✓	✓	✓	✓	(✗)	(✗)	✓
no programming skills	✓	✗	✗	✗	✗	✗	(✓)
no statistics/data science skills	✓	✗	✗	✗	✗	✓	✓
explainability of method required	✓	✓	✓	✗	✗	✗	✗
confidential data	✓	✓	✓	✓	✓	✓	✗
no money available	✓	✓	✓	✓	✓	✓	✗

Note. ✓ = You may still consider this method. ✗ = Do not consider this method further. Brackets indicate exceptions discussed in the main text.

Big question:
Which
methods to
choose for
measuring
OPINIONS in
tekst?

- There is no one-size-fits-all solution
- Challenge: Avoiding outdated approaches, but also recognizing drawbacks of new approaches

(g)LLMs for everything?
→ Not necessarily (but they do offer a lot of possibilities)

Performance is important, but don't forget financial and/or environmental costs, reproducibility, explainability, ...

Table 3: A possible evaluation matrix

Method	Env.C.	Fin.C.	Perform.	LowR-Ls	Det	...
dictionaries	++	++	-	-	++	...
supervised ML	++	++	+	+	+	...
unsupervised ML	++	++	+	++	-/+	...
deep learning	+	++	+	+	-/+	...
LLM	-	-	++	-	-	...
generative LLM	--	--	++ (?)	--	--	...

Note. Env = environmental cost; Fin.C = financial costs; Perf. = (typical) performance; LowR-Ls = likely to be available for small (low-resource) languages; Det = deterministic output