TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Automatisierte Analyse des gesellschaftlichen Impacts von Forschung mit Sprachmodellen

10. Oktober 2024 3. Text+ Plenary 2024 Dr. Maria Becker

Leibniz Gemeinschaft

FORSCHUNGSFRAGE

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Wie können wir den Impact wissenschaftlicher Forschung innerhalb und insbesondere außerhalb der Wissenschaft vorhersagen, klassifizieren und messen?

→ Erfassung wissenschaftlichen Impacts beruht zumeist auf der Analyse wissenschaftlicher Veröffentlichungen und ihrer Verbreitung (v.a. Methoden der Bibliometrie und Informetrie).

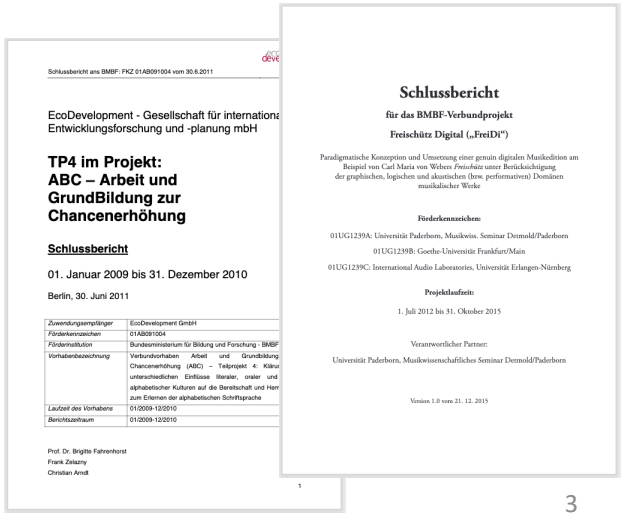
→ Wir interessieren uns für die Auswirkungen wissenschaftlicher Tätigkeit insbesondere über den akademischen Bereich hinaus, z.B. auf die Wirtschaft, Kultur, Politik, Recht, oder Umwelt.

DATENGRUNDLAGE

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Ziel: Empirische Nutzung von Projektberichten zur Vorhersage, Messung und Kategorisierung wissenschaftlichen Impacts

- Datengrundlage: Abschlussberichte geförderter Projekte in den Bereichen
  - Mobilität – 906 Berichte (aus 239 Projekten)
  - Künstliche Intelligenz – 497 Berichte (aus 116 Projekten)
  - Linguistik – 89 Berichte (aus 40 Projekten)
  - Musikwissenschaft – 59 Berichte (aus 27 Projekten)


Schlussbericht ans BMBF: FKZ 01AB091004 vom 30.6.2011

EcoDevelopment - Gesellschaft für internationa
Entwicklungsforschung und -planung mbH

TP4 im Projekt:
ABC – Arbeit und GrundBildung zur Chancenerhöhung

Schlussbericht

01. Januar 2009 bis 31. Dezember 2010

Berlin, 30. Juni 2011

Zuwendungsempfänger	EcoDevelopment GmbH
Förderkennzeichen	01AB091004
Förderinstitution	Bundesministerium für Bildung und Forschung - BMBF
Vorhabenbezeichnung	Verbundvorhaben Arbeit und Grundbildung Chancenerhöhung (ABC) – Teilprojekt 4: Klärung unterschiedlichen Einflüsse literaler, oraler und alphabetischer Kulturen auf die Bereitschaft und Herausbildung zum Erlernen der alphabetischen Schriftsprache
Laufzeit des Vorhabens	01/2009-12/2010
Berichtszeitraum	01/2009-12/2010

Prof. Dr. Brigitte Fahrenhorst
Frank Zelazny
Christian Arndt

1

Schlussbericht

für das BMBF-Verbundprojekt

Freischütz Digital („FreiDi")

Paradigmatische Konzeption und Umsetzung einer genuin digitalen Musikedition am Beispiel von Carl Maria von Webers *Freischütz* unter Berücksichtigung der graphischen, logischen und akustischen (bzw. performativen) Domänen musikalischer Werke

Förderkennzeichen:

01UG1239A: Universität Paderborn, Musikwiss. Seminar Detmold/Paderborn

01UG1239B: Goethe-Universität Frankfurt/Main

01UG1239C: International Audio Laboratories, Universität Erlangen-Nürnberg

Projektlaufzeit:

1. Juli 2012 bis 31. Oktober 2015

Verantwortlicher Partner:

Universität Paderborn, Musikwissenschaftliches Seminar Detmold/Paderborn

Version 1.0 vom 21. 12. 2015

WIE KANN MAN IMPACT MESSEN? DREI ANSÄTZE

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Machine Learning	Annotation der Berichte mit Impactkategorien als Trainingsdaten + überwachte Lernverfahren
Korpuslinguistische Analysen	Korpuslinguistische Detektion und Analyse von Pressetexten über die Forschungsprojekte
Umfragen	Online-Umfragen zu den Forschungsprojekten

Impacterfassung durch Annotationen und maschinelle Lernverfahren

ZIELE DES SUPERVISED MACHINE LEARNING

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Ziel: Entwicklung eines Modells, das in Projektberichten automatisch Sätze/ Abschnitte detektiert, die den (potenziellen) Impact des Projekts ausdrücken und diesen Impact klassifiziert (z.B. gesellschaftlicher Impact, ökonomischer Impact…)
- Grundlage des Models: Manuell mit Impactkategorien annotierte Projektberichte als Trainingsdaten (supervised Machine Learning)
- Erlernen von Mustern in Texten, die (potenziellen) Impact ausdrücken → Erkennen und Klassifizieren von neuen Daten basierend auf erlernten Mustern
- Modell soll auf Texte aus verschiedenen wissenschaftlichen Bereichen anwendbar sein

BEISPIELE FÜR IMPACT-INDIZIERENDE SÄTZE

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- *Ein weiteres Ergebnis des Projekts war die Ableitung von Empfehlungen für Sicherheitsstandards für Elektrofahrzeuge auf Basis der Ergebnisse von Nutzerstudien.* (Domäne (Elektro-)Mobilität)
- *Die aus diesem Projekt gewonnenen Erkenntnisse sollen zu einer besseren Betreuung hilfsbedürftiger älterer Menschen beitragen.* (Domäne Künstliche Intelligenz)
- *Unsere Studien zeigen den Einfluss von Migration auf die Entwicklung individueller und sozialer Werte als zentrale Aspekte jugendlicher Identität.* (Domäne Linguistik)
- *Mit der Software „Annotation Tool“ wurde ein ebenso effektives wie flexibles Werkzeug zur detaillierten manuellen Musikbeschreibung geschaffen.* (Domäne Musikwissenschaften)

SCHRITT 1: IDENTIFIKATION IMPACTRELEVANTER TEXTPASSAGEN

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Problem: Forschungsberichte sind i.d.R sehr umfangreich (bis zu 150 Seiten) → Zeitaufwändige und kostspielige Annotationen
- Lösung: Modell zur automatischen Identifizierung impactrelevanter Abschnitte → können dann für manuelle Annotationen verwendet werden
- Mixed-Methods-Ansatz: Kombination überwachter maschineller Lernverfahren mit Heuristiken
- Ergebnisse: Bis zu 0,81 Accuracy (Akkuratheit)
- Methode ist übertragbar auf verschiedenen Forschungsbereiche und Sprachen

Becker et al. 2024: Detecting impact relevant sections in scientific research. LREC-COLING 2024

SCHRITT 2: MANUELLE ANNOTATION DER EXTRAHIERTEN PASSAGEN MIT IMPACTKATEGORIEN

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

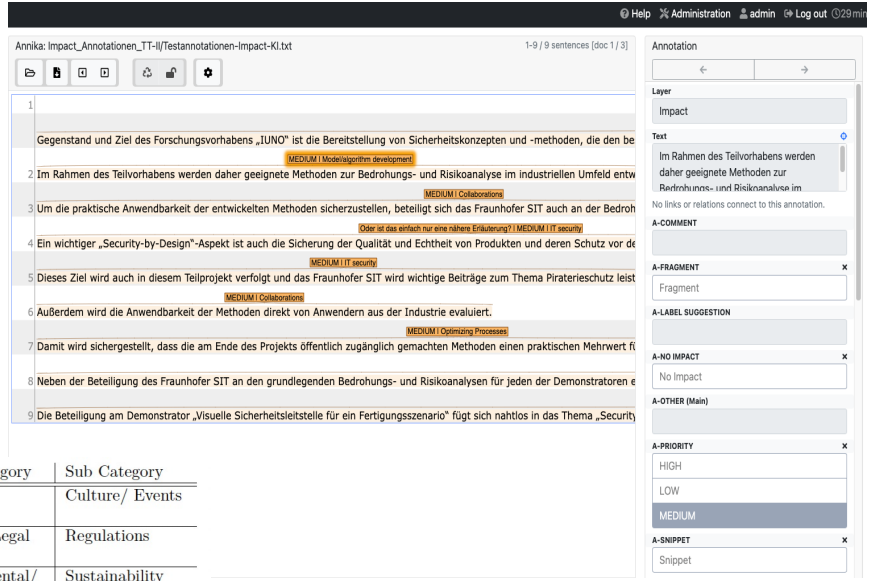
Annotationsschema mit sieben Hauptkategorien plus Subkategorien

i. Gesellschaftlicher Impact
ii. Politischer und rechtlicher Impact
iii. Ethischer Impact
iv. Wirtschaftlicher Impact
v. Ökologischer Impact
vi. Technischer Impact
vii. Akademischer Impact

MAIN CATEGORY	Description and Subcategories	Examples
Societal Impact	occurs when a project influences societal groups or institutions like schools, local authorities, foundations, or clubs, refugees/migration, religious persecution, etc.	
	Education outside academia, e.g. new/improved learning and teaching methods for schools, learning efficiency, learning practical skills	[MUSIC] The temporal structure of the further training of the highschool teachers was assessed as positive, despite the difficulties mentioned with the compatibility with family and job, which speaks for the long further training time frame of MuBiKi.
	Culture, e.g. the organization of concerts, theaters, novels…	[MUSIC] At the end of the workshop, the participant gave a concert at the Altstadtfestival, where children could sign up for future classes.
	Physical health, e.g. fewer respiratory diseases, vaccination campaigns in emergency situations etc.	[AI] With the help of the trend recognition developed in the project on the basis of recorded processes and vital parameters, it should be possible in future to recognise possible dangers in the development of the course of the disease at an early stage and to initiate suitable measures in the care of the respective patient concerned.
	Life quality/Mental health, e.g. fewer depressions, work-life balance, smart home systems, personal growth etc.	[AI] the knowledge gained from this project was to contribute to better care for elderly people in need of help.
	Safety, e.g., road safety, users' safety, general safety, fewer accidents… *(but not IT security/data privacy; see below)*	[MOB] Another outcome of the project was the derivation of recommendations for safety standards for electric vehicles based on the results of user studies.
	Establishing/improving collaborations or networks (outside academia)	[MUSIC] One of the most interesting and important results is certainly the cooperation with teachers from highschools, which stands out against work-sharing forms of cooperation and enables an equal cooperation against the background of an expansion of professional scopes of action.
	Other	

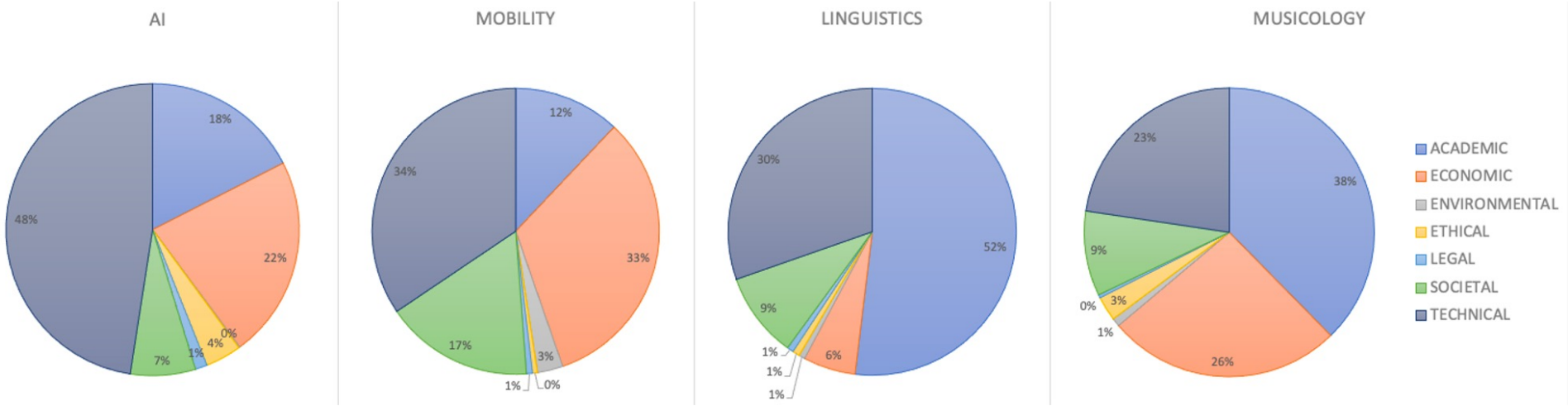
ANNOTATIONSPROZESS

- Parallele Annotation mit INCEpTION
- Annotation auf der Satzebene
- Zusätzliche Annotation der Impactintensität (*high – medium – low*)
- 60.000 annotierte Sätze
- Hohes Inter-Annotator Agreement: 74.48 (Kappa)



Sentence	Domain	Main Category	Sub Category
To conclude the workshop, the participant gave a concert at the Old Town Festival, where children could sign up for future courses.	Musicology	Societal	Culture/ Events
The insights gained from the analyses were presented to the Federal Ministry of Education and Research.	AI	Political/Legal	Regulations
The developed battery lasts an average of 5 years longer than conventional batteries, contributing to reduced electronic waste in this way.	Mobility	Environmental/ Ecological	Sustainability
We demonstrate the impact of migration on the development of individual and social values as central aspects of youth identity.	Linguistics	Ethical	Justice
We can offer an effective and robust algorithm for extracting profession-specific information from English and German texts.	Linguistics	Technical	Model Development
Optimization of the module led to greater satisfaction among test users.	AI	Economic	Employee Satisfaction
One of the most interesting and important results is undoubtedly the collaboration with teachers from high schools, which differs from division-of-labor forms of cooperation and enables equal participation.	Musicology	Other (Main)	Collaborations
We propose a new, semi-automated research method for a contrastive comparison of German and English grammars.	Linguistics	Academic	Research Methods

DATENAUSWERTUNG: IMPACTKATEGORIEN



Verteilung der Impactkategorien auf die Domänen (in Prozent)

DATENAUSWERTUNG: SCHLÜSSELWÖRTER

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

academic		economic		environmental		ethic		legal		societal		technical	
projekt	583	möglich	217	nachhaltig	52	ziel	29	rechtlich	15	ziel	142	entwickeln	475
ergebnis	455	neu	212	neu	16	thema	24	ziel	13	projekt	108	projekt	319
weit	326	unternehmen	169	ziel	16	entwicklung	18	rahmenbedingung	9	möglich	89	entwicklung	291
rahmen	326	ziel	161	optisch	12	sozial	18	betrieb	9	schüler	87	ergebnis	271
können	325	projekt	155	konzept	11	möglich	16	möglich	8	können	79	rahmen	255
wissenschaft	287	wirtschaft	153	möglich	10	hochschule	16	technisch	8	entwicklung	78	neu	234
bereich	243	entwicklung	133	technisch	10	fh	16	rahmen	8	ergebnis	74	ziel	231
entwicklung	231	ergebnis	129	hersteller	9	umgang	14	entwicklung	8	rahmen	68	können	229
ziel	208	weit	117	rahmen	9	aspekt	13	neu	8	neu	66	weit	227
digital	207	können	116	betrieb	9	technisch	13	projekt	7	teilnehmen	64	möglich	207
entwickeln	206	entwickeln	113	nutzung	9	stark	12	daten	7	weit	63	daten	203
neu	199	rahmen	107	arbeit	9	massnahme	12	organisator	7	angebot	59	verfahren	181
möglich	196	nutzen	106	entwicklung	8	projekt	12	ergebnis	6	entwickeln	57	erstellen	172
hochschule	172	prozess	91	ergebnis	8	daten	12	verein	6	nutzen	53	basis	159
kooperation	170	region	88	vorhaben	7	apps	12	nachhaltig	6	öffentlich	52	technisch	151
universität	169	bereich	87	mobilität	7	studieren	12	grundlage	6	elektromobilit	51	gross	139
arbeit	162	geschäftsmodel	82	langfristig	7	können	11	jurist	6	bereich	46	implementieren	131
durchführen	161	gross	81	beitrag	7	weit	11	können	5	durchführen	44	analyse	131
verschieden	157	international	80	deutlich	7	datenschutz	11	ermöglichen	5	kind	40	automatisch	130
gemeinsam	157	kunde	79	reduzieren	6	nutzer	11	prüfen	5	zukunft	38	digital	126

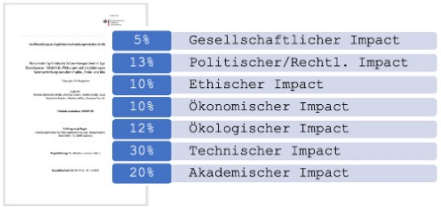
Häufigste Wörter pro Impactkategorie (abs. Zahlen). Gelbe Markierungen: Allgemeinen Impact indizierende Wörter. Grüne Markierungen: Wörter, die Hinweise auf eine bestimmte Impactkategorie geben

MACHINE LEARNING ZUR IMPACTERFASSUNG

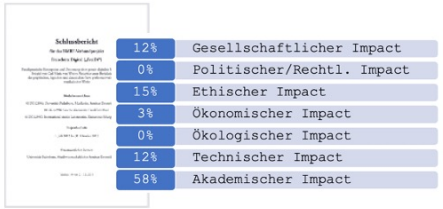
TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Input: Projektberichte

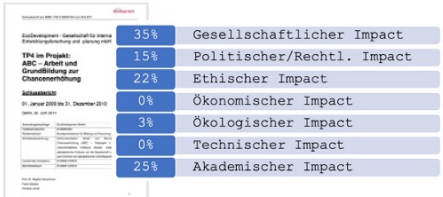
Output: Vorhersage des Impactpotenzials der Projekte



5%	Gesellschaftlicher Impact
13%	Politischer/Rechtl. Impact
10%	Ethischer Impact
10%	Ökonomischer Impact
12%	Ökologischer Impact
30%	Technischer Impact
20%	Akademischer Impact



12%	Gesellschaftlicher Impact
0%	Politischer/Rechtl. Impact
15%	Ethischer Impact
3%	Ökonomischer Impact
0%	Ökologischer Impact
12%	Technischer Impact
58%	Akademischer Impact



35%	Gesellschaftlicher Impact
15%	Politischer/Rechtl. Impact
22%	Ethischer Impact
0%	Ökonomischer Impact
3%	Ökologischer Impact
0%	Technischer Impact
25%	Akademischer Impact

MASCHINELLE LERNVERFAHREN

Experimente mit verschiedenen Large Language Modellen (Llama, BERT, GPT)

(I) Zero Shot Learning (Prompting)

(1) You will be provided with a text that possibly outlines the impact of a research project.
(2) We define impact as an effect of scientific activities within academia or beyond the academic field, e.g. on the scientific field, economy, society, culture, politics, law, technology or the environment. In a report, impact may be represented by describing methods and routines implemented for a project, or by the impact that authors anticipated when writing their final reports. This means that estimated impact is also considered as impact. Impact could also be the maintenance or the avoidance of change.
(3) Your task is to assign one impact category to the text from the following options:
(4) NO-IMPACT: the sentence does not express any impact.
TECHNICAL IMPACT: refers to technologies that are used outside of the original project, e.g. software prototype development, Improving IT security, or data release
ECONOMIC IMPACT: refers to the use of research results for economic developments, e.g. development of business models, service quality, or economic strategies
ACADEMIC IMPACT: refers to impact within academia - within or beyond the own field/institution, e.g.improved learning and teaching, new research methods, or publications
SOCIETAL IMPACT: occurs when a project influences societal groups or institutions like schools, local authorities, foundations, or clubs, refugees/migration, religious persecution, etc.
POLITICAL-LEGISLATIVE IMPACT: refers to using the project results in political or legislative contexts, e.g. contributions to laws or political regulations
ETHICAL IMPACT: refers to ethical impact, e.g. equality, awareness, or charity
ENVIRONMENTAL IMPACT: refers to changes of ecological or environmental aspects, e.g. climate protection, protection of species, or sustainability of products
OTHER: any sentences that express impact, which are not captured by the other categories.
(5) Simply return one category without any explanation.
(6) Sentence: <Sentence>. Answer:

MASCHINELLE LERNVERFAHREN

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Experimente mit verschiedenen Large Language Modellen (Llama, GPT, BERT)

(II) Few Shot Learning (Prompting + gute Beispielinstanzen):

(1) You will be provided with a text that possibly outlines the impact of a research project.
(2) We define impact as an effect of scientific activities within academia or beyond the academic field, e.g. on the scientific field, economy, society, culture, politics, law, technology or the environment. In a report, impact may be represented by describing methods and routines implemented for a project, or by the impact that authors anticipated when writing their final reports. This means that estimated impact is also considered as impact. Impact could also be the maintenance or the avoidance of change.
(3) Your task is to assign one impact category to the text from the following options:
(4) NO-IMPACT: the sentence does not express any impact.
TECHNICAL IMPACT: refers to technologies that are used outside of the original project, e.g. software prototype development, Improving IT security, or data release
ECONOMIC IMPACT: refers to the use of research results for economic developments, e.g. development of business models, service quality, or economic strategies
ACADEMIC IMPACT: refers to impact within academia - within or beyond the own field/institution, e.g.improved learning and teaching, new research methods, or publications
SOCIETAL IMPACT: occurs when a project influences societal groups or institutions like schools, local authorities, foundations, or clubs, refugees/migration, religious persecution, etc.
POLITICAL-LEGISLATIVE IMPACT: refers to using the project results in political or legislative contexts, e.g. contributions to laws or political regulations
ETHICAL IMPACT: refers to ethical impact, e.g. equality, awareness, or charity
ENVIRONMENTAL IMPACT: refers to changes of ecological or environmental aspects, e.g. climate protection, protection of species, or sustainability of products
OTHER: any sentences that express impact, which are not captured by the other categories.
(5) Simply return one category without any explanation.
(6) Sentence: <Sentence>. Answer:

(7) For example:
Sentence: <NO-IMPACT sentence> Answer: No-IMPACT
Sentence: <TECHNICAL sentence> Answer: TECHNICAL
...
Sentence: <OTHER sentence> Answer: OTHER

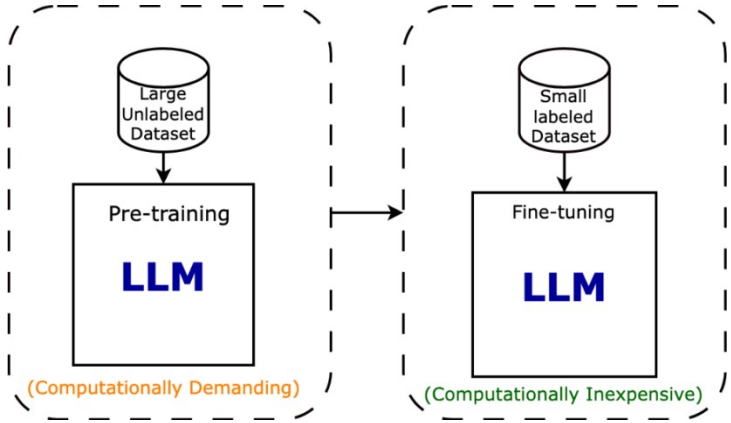
TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

MASCHINELLE LERNVERFAHREN

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Experimente mit verschiedenen Large Language Modellen (Llama, GPT, BERT)

(III) Finetuning (Nutzung unserer annotierten Daten als Trainingsinstanzen)



ERGEBNISSE (F1 SCORES)

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

		Domain				Intensity		
Method	Model	Music	Linguistics	Mobility	AI	High	Medium	Low
Zero-shot	Llama2	0.46	**0.58**	0.36	0.49	0.45	**0.53**	0.43
	ChatGPT	0.54	**0.62**	0.39	0.49	0.49	**0.59**	0.45
Few-shot	Llama2	0.54	**0.58**	0.44	0.51	0.45	**0.57**	0.50
	ChatGPT	0.55	**0.62**	0.41	0.53	0.51	**0.60**	0.46
Fine-tuning	Llama2	0.62	**0.65**	0.53	0.63	0.61	**0.65**	0.55
	BERT	0.66	**0.69**	0.61	0.61	0.61	**0.68**	0.61

BEOBACHTUNGEN

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Finetuning funktioniert signifikant besser als Zero- oder Few-Shot Learning
  - Neue Klassifikationsschemata sind eine Herausforderung für vortrainierte (aber nicht finegetunte) LLMs
  - Trainingsdaten sind wichtig → Der Annotationsaufwand hat sich also gelohnt!

		Domain				Intensity		
Method	Model	Music	Linguistics	Mobility	AI	High	Medium	Low
Zero-shot	Llama2	0.46	**0.58**	0.36	0.49	0.45	**0.53**	0.43
	ChatGPT	0.54	**0.62**	0.39	0.49	0.49	**0.59**	0.45
Few-shot	Llama2	0.54	**0.58**	0.44	0.51	0.45	**0.57**	0.50
	ChatGPT	0.55	**0.62**	0.41	0.53	0.51	**0.60**	0.46
Fine-tuning	Llama2	0.62	**0.65**	0.53	0.63	0.61	**0.65**	0.55
	BERT	0.66	**0.69**	0.61	0.61	0.61	**0.68**	0.61

BEOBACHTUNGEN

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Finetuning funktioniert signifikant besser als Zero- oder Few-Shot Learning
  - Neue Klassifikationsschemata sind eine Herausforderung für vortrainierte (aber nicht finegetunte) LLMs
  - Trainingsdaten sind wichtig → Der Annotationsaufwand hat sich also gelohnt!
- Bei einer kleinen Menge von Trainingsdaten kann das Finetuning kleinerer Modelle zu besseren Ergebnissen als das Finetuning größerer Modelle führen

		Domain				Intensity		
Method	Model	Music	Linguistics	Mobility	AI	High	Medium	Low
Zero-shot	Llama2	0.46	**0.58**	0.36	0.49	0.45	**0.53**	0.43
	ChatGPT	0.54	**0.62**	0.39	0.49	0.49	**0.59**	0.45
Few-shot	Llama2	0.54	**0.58**	0.44	0.51	0.45	**0.57**	0.50
	ChatGPT	0.55	**0.62**	0.41	0.53	0.51	**0.60**	0.46
Fine-tuning	Llama2	0.62	**0.65**	0.53	0.63	0.61	**0.65**	0.55
	BERT	0.66	**0.69**	0.61	0.61	0.61	**0.68**	0.61

BEOBACHTUNGEN

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

- Finetuning funktioniert signifikant besser als Zero- oder Few-Shot Learning
  - Neue Klassifikationsschemata sind eine Herausforderung für vortrainierte (aber nicht finegetunte) LLMs
  - Trainingsdaten sind wichtig → Der Annotationsaufwand hat sich also gelohnt!
- Bei einer kleinen Menge von Trainingsdaten kann das Finetuning kleinerer Modelle zu besseren Ergebnissen als das Finetuning größerer Modelle führen
- Ohne Finetuning können Open-Source-LLMs teilweise eine ähnliche Leistung wie kommerzielle Modelle erzielen

		Domain				Intensity		
Method	Model	Music	Linguistics	Mobility	AI	High	Medium	Low
Zero-shot	Llama2	0.46	**0.58**	0.36	0.49	0.45	**0.53**	0.43
	ChatGPT	0.54	**0.62**	0.39	0.49	0.49	**0.59**	0.45
Few-shot	Llama2	0.54	**0.58**	0.44	0.51	0.45	**0.57**	0.50
	ChatGPT	0.55	**0.62**	0.41	0.53	0.51	**0.60**	0.46
Fine-tuning	Llama2	0.62	**0.65**	0.53	0.63	0.61	**0.65**	0.55
	BERT	0.66	**0.69**	0.61	0.61	0.61	**0.68**	0.61

DISKUSSION

- Können LLMs zur Prognose und Klassifikation von Impact eingesetzt werden? Unsere Ergebnisse zeigen: Ja!
- Wie aussagekräftig sind Impactprognosen auf der Basis von Projektberichten?
- Was muss für einen verantwortungsvollen Umgang mit LLMs in der Impactforschung berücksichtigt werden? Welche Risiken gibt es?
- Wie kann der verantwortungsvolle Einsatz von LLMs zur retroaktiven Impacterfassung statt der Generierung und Optimierung von Projektanträgen kommuniziert und gelehrt werden?

TEXTTRANSFER IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Danke! Fragen?

Automatisierte Analyse des gesellschaftlichen Impacts von Forschung mit Sprachmodellen

10. Oktober 2024 3. Text+ Plenary 2024 Dr. Maria Becker

Leibniz Gemeinschaft