A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus

Siddhant Arora[1], Henry Hosseini[2], Christine Utz[3], Vinayshekhar Bannihatti Kumar[1], Tristan Dhellemmes[1], Abhilasha Ravichander[1], Peter Story,[1] Jasmine Mangat[1], Rex Chen[1], Martin Degeling[3], Tom Norton[4], Thomas Hupperich[2], Shomir Wilson[5], and Norman Sadeh[1]

1. School of Computer Science, Carnegie Mellon University
2. University of Munster
3. Ruhr University Bochum
4. Fordham University School of Law
5. College of Information Sciences and Technology, Penn State University

Carnegie Mellon University
School of Computer Science

WWU
MÜNSTER

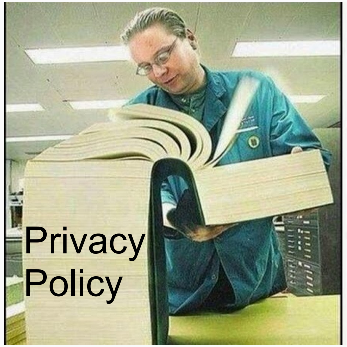
RUHR
UNIVERSITÄT
BOCHUM
RUB

FORDHAM UNIVERSITY

PennState
College of Information
Sciences and Technology

Privacy Policies are Unusable

- Privacy policies are the primary mechanism by which organizations disclose their data practices
- When was the last time you read a privacy policy for any of the websites you use?
- 2008 study by McDonald and Cranor estimated reading privacy policies would take **40 minutes** per day!



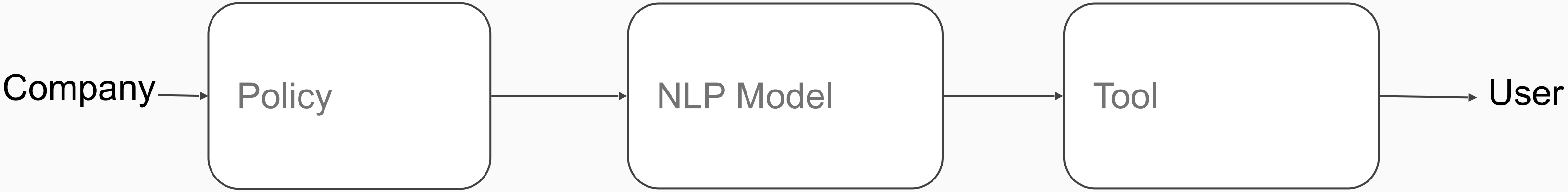
[1] McDonald, Aleecia M., and Lorrie Faith Cranor. "The cost of reading privacy policies." Isjlp 4 (2008): 543.

Privacy Regulations

- Legislative bodies have responded by imposing new requirements about the information privacy policies must disclose
- Disclosures made by the same organization are not always the same in different languages – due to location specific privacy regulations
  - Aim to systematically capture the difference to understand the impact of new privacy regulation
  - Develop a better understanding of current industry practices when it comes to accommodating regulatory requirements

Usable Privacy Policy Project

- NLP models to understand the text of privacy policies
- Tools to inform users about policies they are agree to



Company → Policy → NLP Model → Tool → User

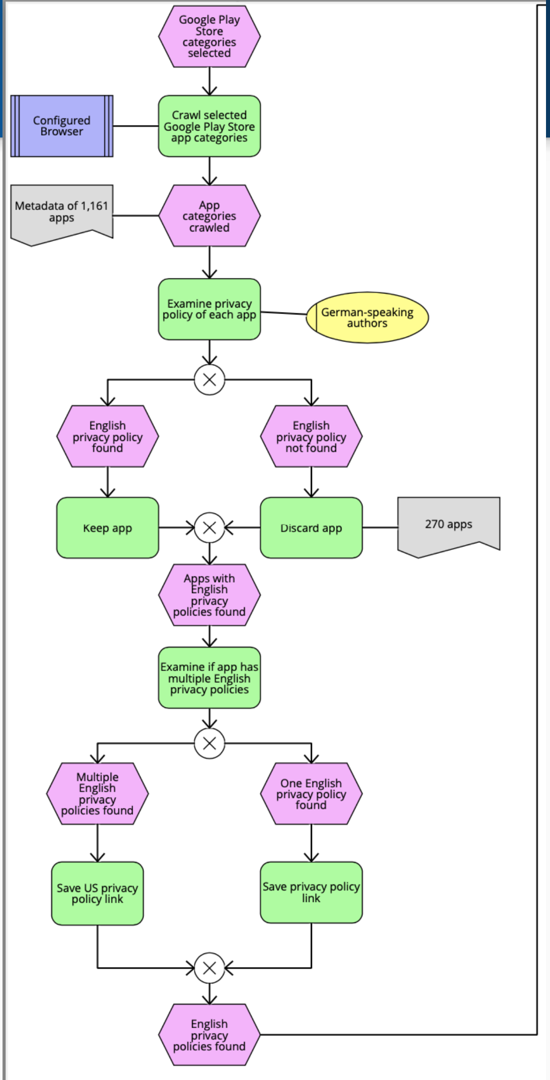
USABLEPRIVACY.ORG
the usable privacy policy project

Major Research Contributions

- New annotation scheme to capture newly introduced privacy regulations i.e. EU’s GDPR and California’s CCPA and CPRA
- First Bilingual annotated corpus of English and German privacy policies
- NLP classifiers for automatically identifying English and German privacy policy disclosures

Assembling Corpus

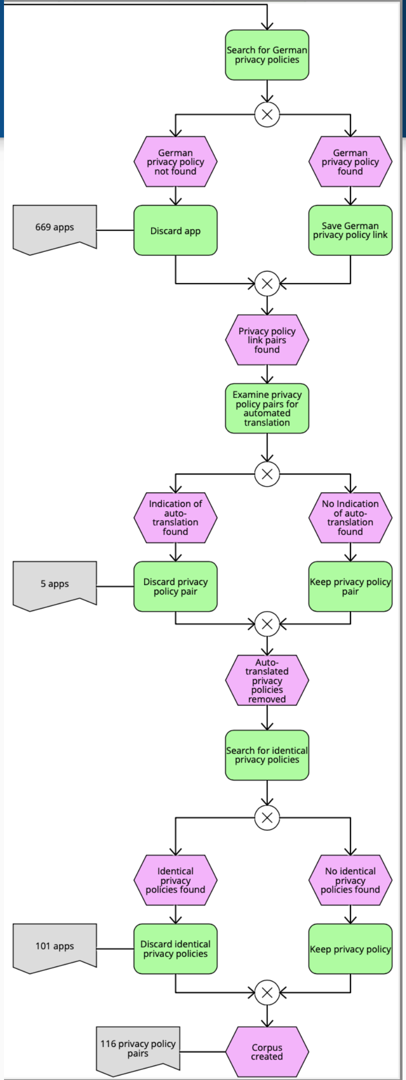
- Collected mobile app privacy policies from the Google Play Store
  - Representative subset of App Categories
- Retrieved English Privacy policy link
  - If multiple English privacy policy -> US policy downloaded



Google Play Store categories selected
Configured Browser
Crawl selected Google Play Store app categories
Metadata of 1,161 apps
App categories crawled
Examine privacy policy of each app
German-speaking authors
English privacy policy found
English privacy policy not found
Keep app
Discard app
270 apps
Apps with English privacy policies found
Examine if app has multiple English privacy policies
Multiple English privacy policies found
One English privacy policy found
Save US privacy policy link
Save privacy policy link
English privacy policies found

Assembling Corpus

- Focused on apps with policies in both German and English
- Eliminated identical policy pairs already included in the corpus
- Discard privacy policies that were automatic translations



Search for German privacy policies

German privacy policy not found

German privacy policy found

669 apps

Discard app

Save German privacy policy link

Privacy policy link pairs found

Examine privacy policy pairs for automated translation

Indication of auto-translation found

No Indication of auto-translation found

5 apps

Discard privacy policy pair

Keep privacy policy pair

Auto-translated privacy policies removed

Search for identical privacy policies

Identical privacy policies found

No identical privacy policies found

101 apps

Discard identical privacy policies

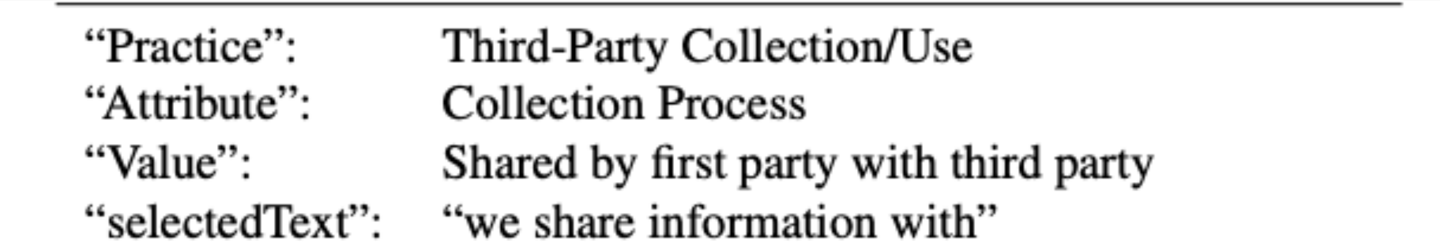
Keep privacy policy

116 privacy policy pairs

Corpus created

Annotation Scheme

- Updated OPP-115[2] scheme to capture protections introduced in EU's GDPR and California's CCPA/ CPRA
- Focus on identification of First Party Collection/Use and Third Party Collection/Use Data Practices

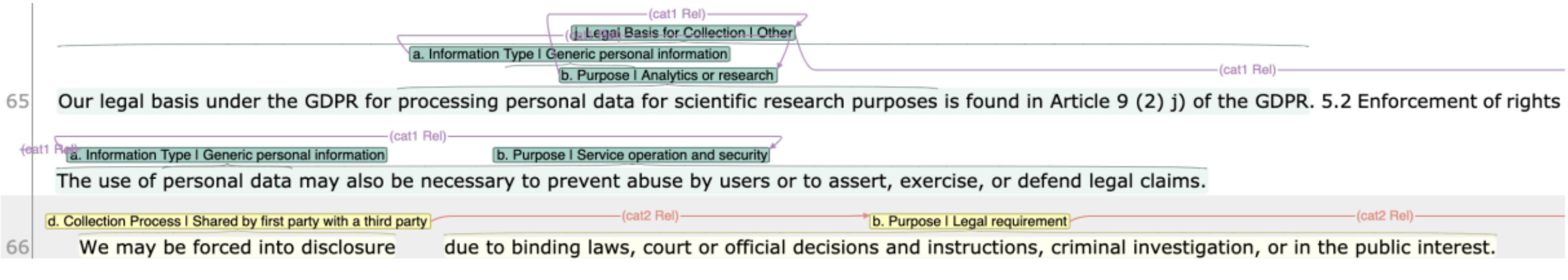


"Practice":	Third-Party Collection/Use
"Attribute":	Collection Process
"Value":	Shared by first party with third party
"selectedText":	"we share information with"

[2] Wilson et al. "The Creation and Analysis of a Website Privacy Policy Corpus." ACL 16

Annotation Process

- Configured the INCEpTION annotation platform[3]
- Recruited team of 12 and 10 law students in Germany and the US respectively as annotators
- Each privacy policy annotated with 3 annotators



65 Our legal basis under the GDPR for processing personal data for scientific research purposes is found in Article 9 (2) j) of the GDPR. 5.2 Enforcement of rights

The use of personal data may also be necessary to prevent abuse by users or to assert, exercise, or defend legal claims.

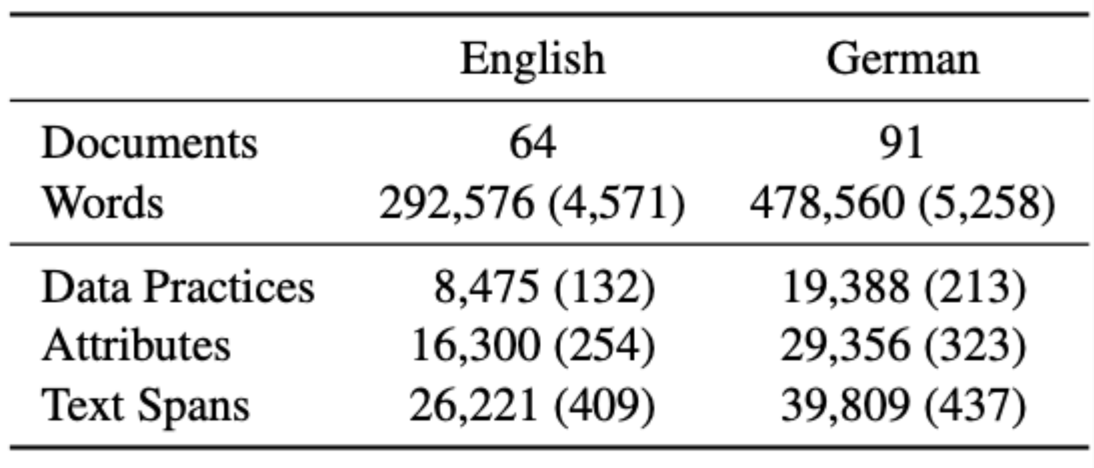
66 We may be forced into disclosure due to binding laws, court or official decisions and instructions, criminal investigation, or in the public interest.

(cat1 Rel)
j. Legal Basis for Collection | Other
a. Information Type | Generic personal information
b. Purpose | Analytics or research
(cat1 Rel)
(cat1 Rel)
(cat1 Rel)
a. Information Type | Generic personal information
b. Purpose | Service operation and security
d. Collection Process | Shared by first party with a third party
(cat2 Rel)
b. Purpose | Legal requirement
(cat2 Rel)

[3] de Castilho et al. "Linking Text and Knowledge using the INCEpTION annotation platform." eScience ’18

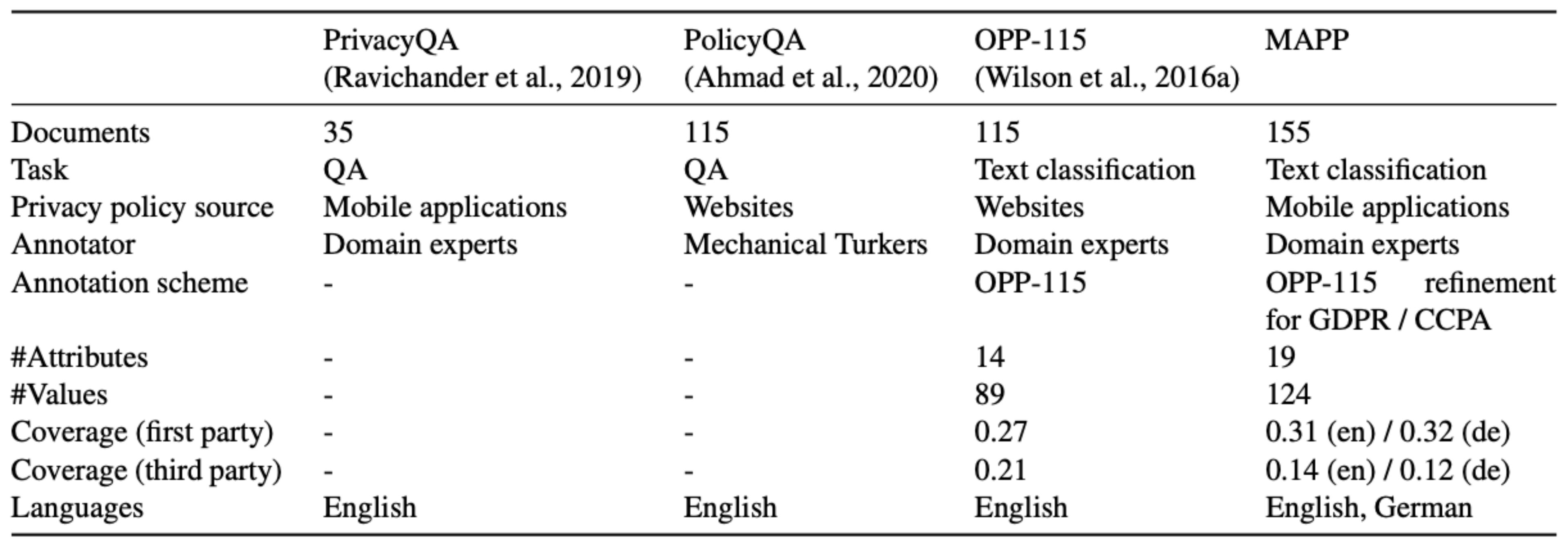
MAPP Corpus

- Consists of 91 German and 64 English fully annotated
- Semi parallel subcorpus of 59 policies - MAPP-59
  - Help to understand differences arising from different regulatory regimes
  - Linguistic differences that may change reader’s interpretation



	English	German
Documents	64	91
Words	292,576 (4,571)	478,560 (5,258)
Data Practices	8,475 (132)	19,388 (213)
Attributes	16,300 (254)	29,356 (323)
Text Spans	26,221 (409)	39,809 (437)

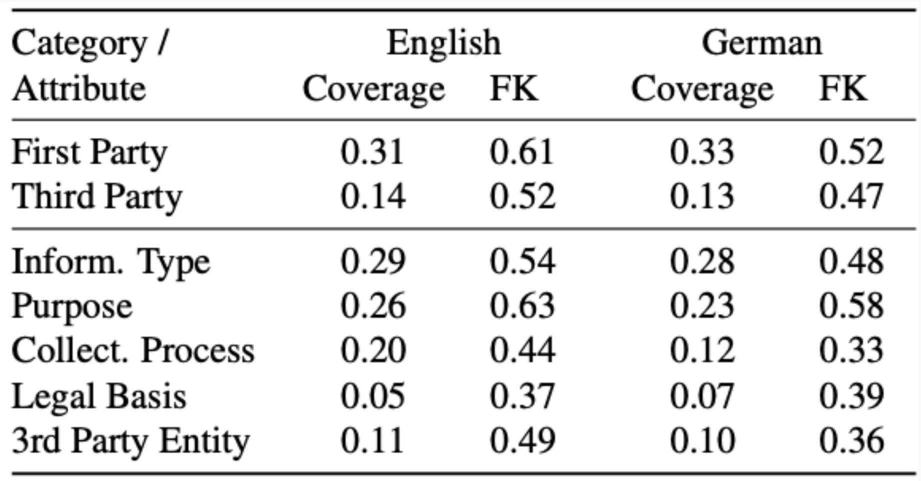
Comparison with other privacy policy corpora



	PrivacyQA (Ravichander et al., 2019)	PolicyQA (Ahmad et al., 2020)	OPP-115 (Wilson et al., 2016a)	MAPP
Documents	35	115	115	155
Task	QA	QA	Text classification	Text classification
Privacy policy source	Mobile applications	Websites	Websites	Mobile applications
Annotator	Domain experts	Mechanical Turkers	Domain experts	Domain experts
Annotation scheme	-	-	OPP-115	OPP-115 refinement for GDPR / CCPA
#Attributes	-	-	14	19
#Values	-	-	89	124
Coverage (first party)	-	-	0.27	0.31 (en) / 0.32 (de)
Coverage (third party)	-	-	0.21	0.14 (en) / 0.12 (de)
Languages	English	English	English	English, German

Inter-Annotator Agreement

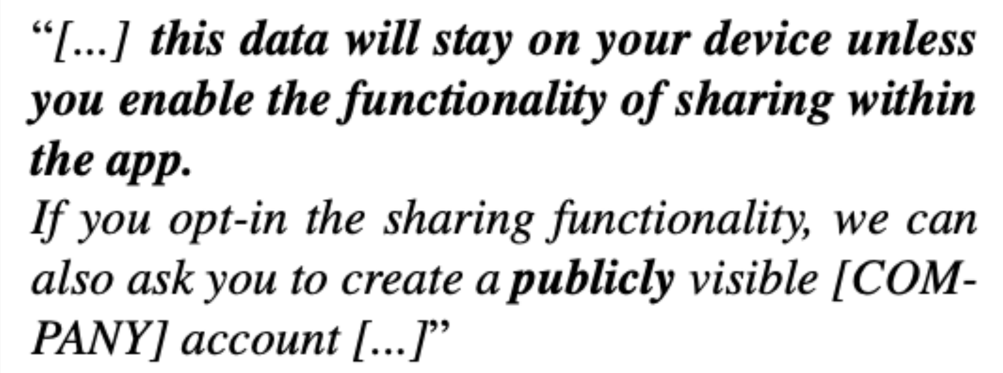
- Segmented privacy policies and calculated agreement at segment level using Fleiss Kappa
- Focus on building classifiers for attributes and values with sufficient agreement and coverage.



Category /	English		German	
Attribute	Coverage	FK	Coverage	FK
First Party	0.31	0.61	0.33	0.52
Third Party	0.14	0.52	0.13	0.47
Inform. Type	0.29	0.54	0.28	0.48
Purpose	0.26	0.63	0.23	0.58
Collect. Process	0.20	0.44	0.12	0.33
Legal Basis	0.05	0.37	0.07	0.39
3rd Party Entity	0.11	0.49	0.10	0.36

Privacy Policies are Ambiguous

- Inherent ambiguity in privacy policies
    - Even law experts[4] disagree about their interpretation
- For the example segment below
    - Annotators struggled with 2 practices being discussed in conjunction

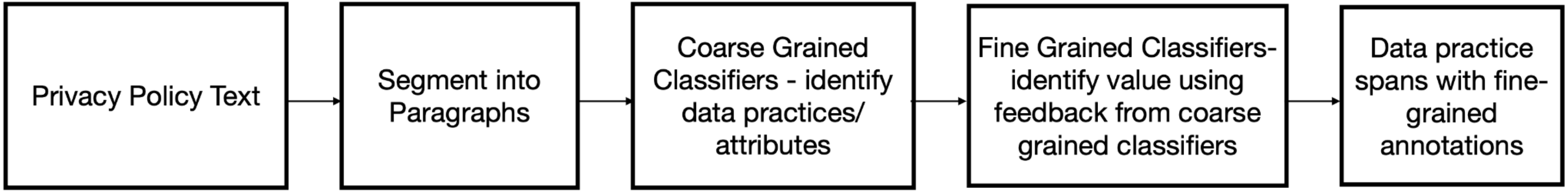


“*[...]* ***this data will stay on your device unless you enable the functionality of sharing within the app.***
If you opt-in the sharing functionality, we can also ask you to create a ***publicly*** *visible [COMPANY] account [...]*”

[4] Reidenberg et al. "Disagreeable privacy policies: Mismatches between meaning and users’ understanding." Berkeley Tech LAW Journal ’15

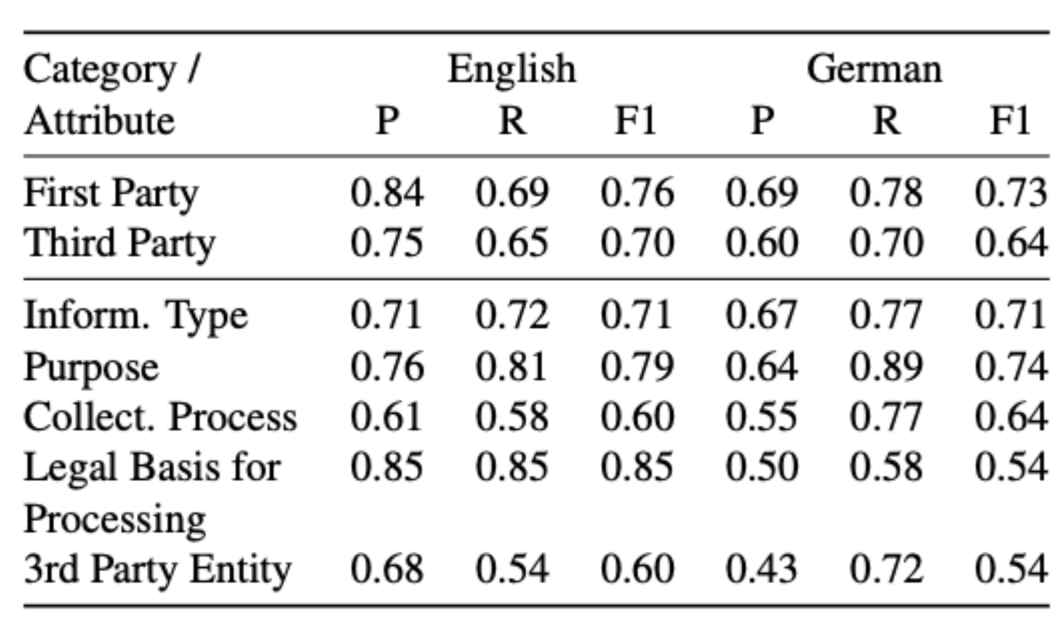
NLP Classifiers

- Train/ test split
  - 52/12 and 75/16 for training/testing in English/German
- Trained text classification models
  - Finetuned pretrained LM (BERT/M-BERT)
  - Experimented using prediction from data practice/ attribute classifiers to predict value


Privacy Policy Text
Segment into Paragraphs
Coarse Grained Classifiers - identify data practices/ attributes
Fine Grained Classifiers- identify value using feedback from coarse grained classifiers
Data practice spans with fine-grained annotations

Results

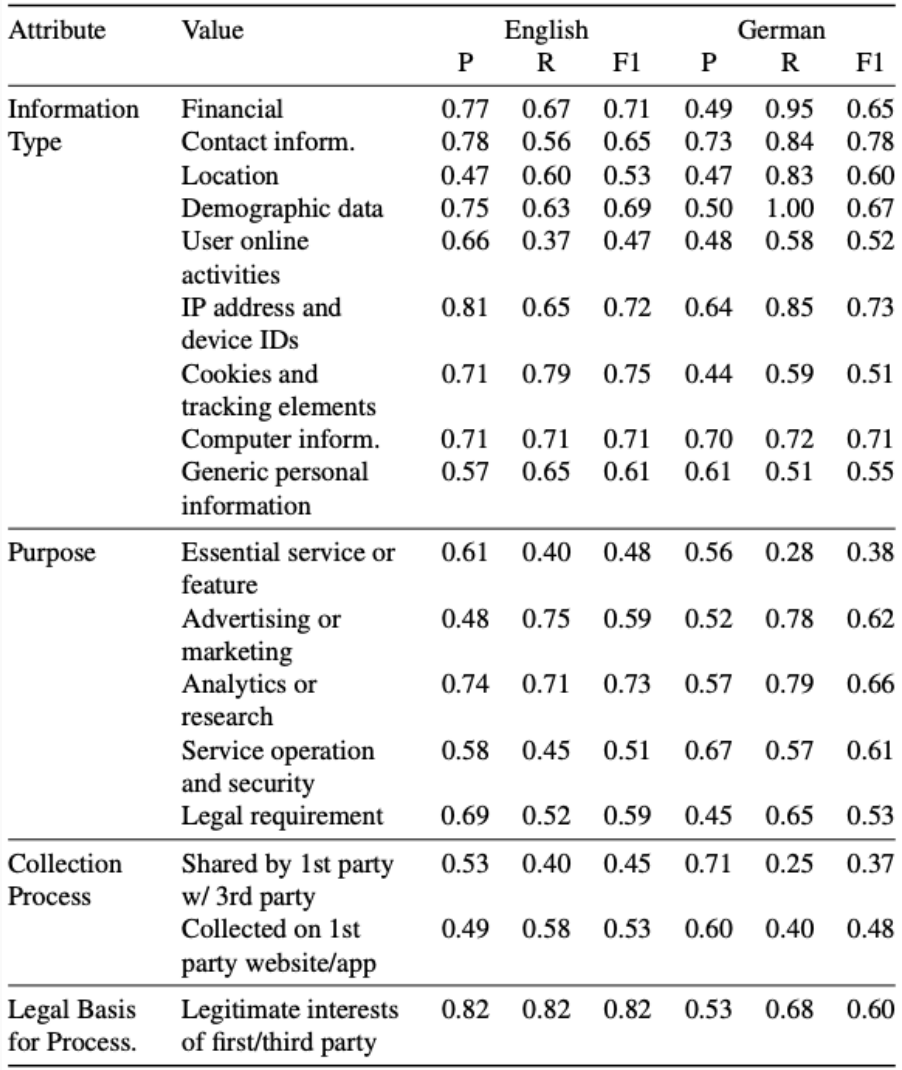
- Report F1 for positive class
- Our German classifiers are less accurate
- Classifier fairly accurate for Information Type and Purpose Attributes



Category /	English			German		
Attribute	P	R	F1	P	R	F1
First Party	0.84	0.69	0.76	0.69	0.78	0.73
Third Party	0.75	0.65	0.70	0.60	0.70	0.64
Inform. Type	0.71	0.72	0.71	0.67	0.77	0.71
Purpose	0.76	0.81	0.79	0.64	0.89	0.74
Collect. Process	0.61	0.58	0.60	0.55	0.77	0.64
Legal Basis for Processing	0.85	0.85	0.85	0.50	0.58	0.54
3rd Party Entity	0.68	0.54	0.60	0.43	0.72	0.54

Results

- Values like Financial with annotation spans containing more distinctive language yielded better performance
- Values like User Online Activities with longer annotation spans harder to identify



Attribute	Value	English			German		
		P	R	F1	P	R	F1
Information Type	Financial	0.77	0.67	0.71	0.49	0.95	0.65
	Contact inform.	0.78	0.56	0.65	0.73	0.84	0.78
	Location	0.47	0.60	0.53	0.47	0.83	0.60
	Demographic data	0.75	0.63	0.69	0.50	1.00	0.67
	User online activities	0.66	0.37	0.47	0.48	0.58	0.52
	IP address and device IDs	0.81	0.65	0.72	0.64	0.85	0.73
	Cookies and tracking elements	0.71	0.79	0.75	0.44	0.59	0.51
	Computer inform.	0.71	0.71	0.71	0.70	0.72	0.71
	Generic personal information	0.57	0.65	0.61	0.61	0.51	0.55
Purpose	Essential service or feature	0.61	0.40	0.48	0.56	0.28	0.38
	Advertising or marketing	0.48	0.75	0.59	0.52	0.78	0.62
	Analytics or research	0.74	0.71	0.73	0.57	0.79	0.66
	Service operation and security	0.58	0.45	0.51	0.67	0.57	0.61
	Legal requirement	0.69	0.52	0.59	0.45	0.65	0.53
Collection Process	Shared by 1st party w/ 3rd party	0.53	0.40	0.45	0.71	0.25	0.37
	Collected on 1st party website/app	0.49	0.58	0.53	0.60	0.40	0.48
Legal Basis for Process.	Legitimate interests of first/third party	0.82	0.82	0.82	0.53	0.68	0.60

Comparing Disclosures in English and German

- Analyzed policies for the presence of markers indicative of GDPR -> “GDPR-aware”
- 54% of apps had English privacy policies that were “GDPR-aware”
  - 43% of these apps specifically singled out EU residents
  - Remaining apps likely extend protections to non EU residents
- 36% of German apps did not acknowledge GDPR
  - Among those that are “GDPR-aware”, about 33% do not address required disclosures under GDPR

What can we learn from our classifiers

- Can we answer policy questions?
  - GDPR Article 6 prohibits collecting and processing personal data without a proper legal basis. What percentage of websites meet this requirement?
- We analysed 22,359 US and 1,864 German website privacy policies
  - 76% of German policies satisfy this requirement
  - 19% of US policies also provide this protection
- Such analysis help to understand the impact of jurisdiction specific privacy regulations

Conclusion

- ➜ We introduced MAPP, the first bilingual corpus of privacy policies
- ➜ We identified how privacy disclosures differ in policies published in English and German
- ➜ We presented initial evidence of the effectiveness of our classifiers at automatically identifying these differences
- ➜ Our study discussed how privacy regulations can account for some of these differences
- ➜ We believe that this type of analysis could ultimately help inform the development of more effective privacy regulations.

Acknowledgment

USABLE PRIVACY.ORG
the usable privacy policy project

Carnegie Mellon University
School of Computer Science

WWU
MÜNSTER

RUHR
UNIVERSITÄT
BOCHUM
RUB

FORDHAM UNIVERSITY

PennState
College of Information
Sciences and Technology

NSF