27th Pan-Hellenic Conference on Progress in Computing and Informatics

A privacy policies dataset in Greek in the GDPR era

Georgia M. Kapitsaki, Maria Papoutsoglou

Department of Computer Science
University of Cyprus
Cyprus

University of Cyprus

Privacy and relevant laws

- Recent laws push providers towards compliance

- In legislation
  - EU: European General Data Protection Regulation (GDPR) (2018)
  - US: California Consumer Privacy Act of 2018 (CCPA)
  - US: California Privacy Rights Act of 2020 (CPRA)
  - Many countries have followed:
    - E.g. Brazilian General Data Protection Law (LGPD)

User rights in GDPR

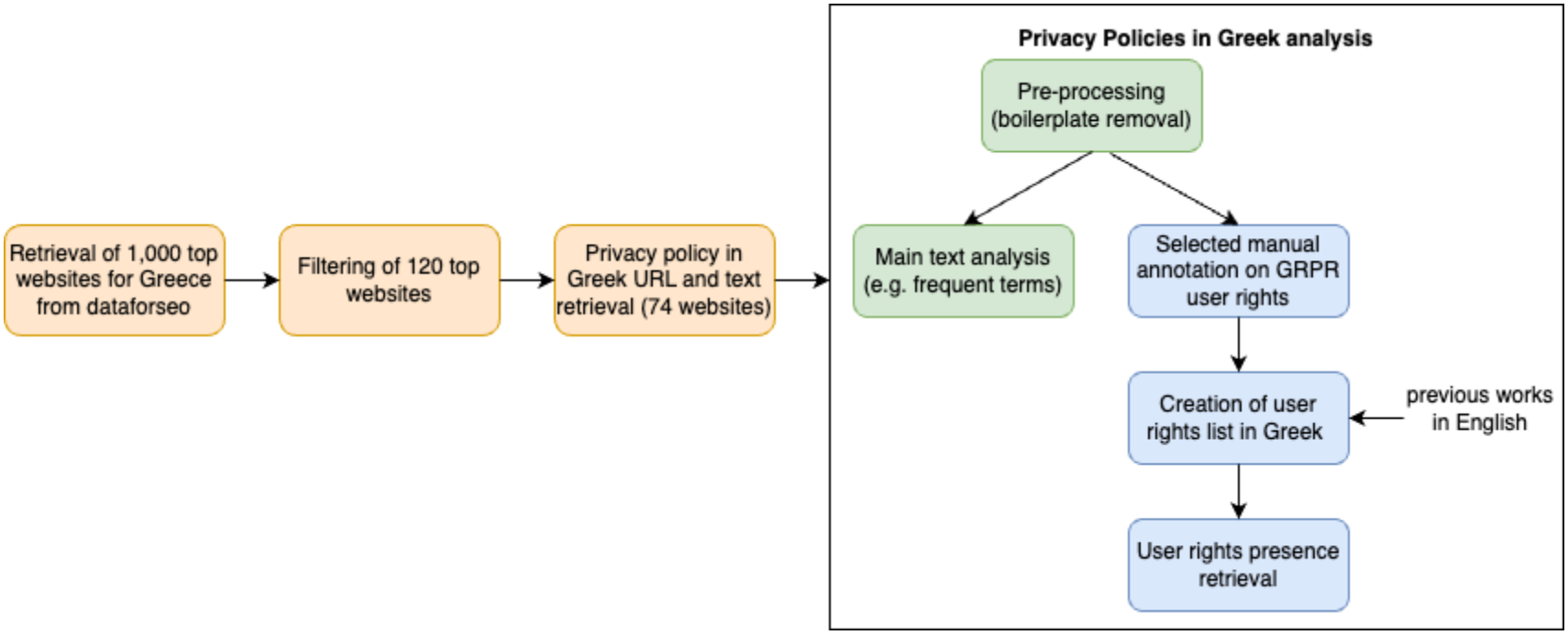
1. the right to information
2. the right of access
3. the right to rectification
4. the right to erasure (right to be forgotten)
5. the right to restriction of processing
6. the right to data portability
7. the right to object
8. the right to avoid automated decision-making

Aim and motivation

- Privacy policies need to comply with the legislation
- Datasets of privacy policies are important to understand the current state of adoption
  - Many previous works rely on policies for:
    - Text summarization
    - Making policies more user-friendly
- Datasets already available in English (e.g. OPP-115)

- **Current work aims create a dataset of privacy policies in the Greek language**
- Serves as starting point for
  - performing further analysis on privacy policies in Greek
  - understanding how recent privacy laws are handled in the Greek context

Dataset creation process



Resulted in: 74 unique privacy policies

(after removing duplicates)

Pre-processing and analysis

- Pre-processing steps:
  - Removed HTML boilerplate content (e.g. siders)
  - Removed stopwords in Greek
    - list from the quanteda package in R
  - Removed the following words (in Greek): or, when, e.g.

- Analysis tools:
  - in Java: HTMLCleaner
  - in R: quanteda, ggplot2, dplyr etc.

Names of top websites in Greece

12 different categories (categories from SpyMetrics)

Category	Rank	Website	Policy effect year
News and Media	23	iefimerida.gr	2023
	24	lifo.gr	N.A.
	26	protothema.gr	2020
	27	in.gr	2022
	29	yahoo.com	2023
	32	ethnos.gr	2019
	33	kathimerini.gr	2019
	35	newsbeast.gr	N.A.
	42	athensvoice.gr	2023
	43	newsit.gr	2023
	45	newsbomb.gr	N.A.
	50	ertnews.gr	2018
	52	insomnia.gr	N.A.
	56	thetoc.gr	N.A.
	57	stigmap.gr	N.A.
	60	inewsgr.com	N.A.
	61	11888.gr	N.A.
	63	tanea.gr	2022
	65	iefimerida.gr	2023
	66	athinorama.gr	2022
	67	tovima.gr	2022
	70	news247.gr	N.A.
	74	capital.gr	N.A.
	76	cnn.gr	N.A.
	79	citymaps.gr	N.A.
	84	aftodioikisi.gr	2023
	93	zougla.gr	N.A.
	94	star.gr	2023
	95	neolaia.gr	2022
	96	rotise.gr	N.A.
	106	naftemporiki.gr	2020
	111	alfavita.gr	2019
	113	skai.gr	2022
	117	newmoney.gr	N.A.
	120	dnews.gr	2022
Science and Education	40	sch.gr	2021
	80	scribd.com	2022
	89	auth.gr	2019
	101	uoa.gr	2018
Vehicles	21	car.gr	N.A.
Arts and Entertainment	73	spotify.com	2023
	105	e-food.gr	2023
Computers Electronics and Technology	2	facebook.com	2023
	6	instagram.com (same)	
	7	pinterest.com	2023
	8	google.com	2023
	1	youtube.com (same)	
	4	google.gr (same)	
	100	business.site (same)	
	12	tiktok.com	2023
	16	apple.com	2022
	17	wordpress.com	N.A.
	44	microsoft.com	2023
E-commerce and Shopping	5	skroutz.gr	N.A.
	11	bestprice.gr	2023
	15	vrisko.gr	N.A.
	31	public.gr	2022
	51	politeianet.gr	2018
	81	shopflix.gr	2022
	85	e-shop.gr	N.A.
	87	kotsovolos.gr	2021
	97	vendora.gr	2019
	104	plaisio.gr	N.A.
	108	e-jumbo.gr	N.A.
	110	xe.gr	2023
Games	98	ign.com	2019
Hobbies and Leisure	48	shutterstock.com	2021
	99	dreamstime.com	N.A.
Jobs and Career	86	indeed.com	2023
Law and Government	46	europa.eu	2019
Reference Materials	3	wikipedia.org	2021
	30	wiktionary.org (same)	
	14	xo.gr	2018
	112	babla.gr	N.A.
Travel and Tourism	22	tripadvisor.com.gr	2023
	39	tripadvisor.com (same)	
	28	booking.com	2023
	71	hotels.com	2023
	77	trip.com	2022
	107	gtp.gr	N.A.

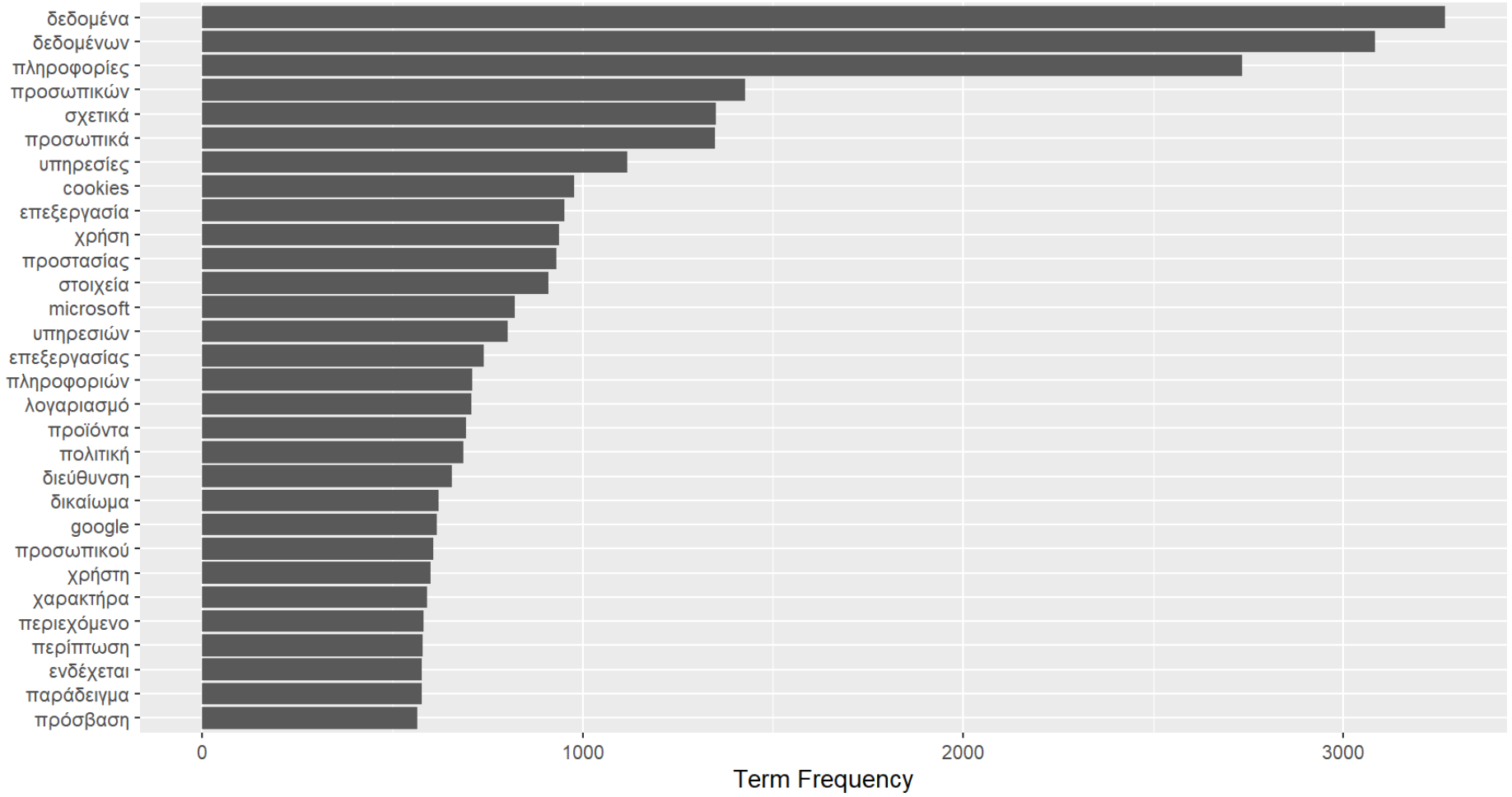
*N.A.–Not Available

Dataset structure

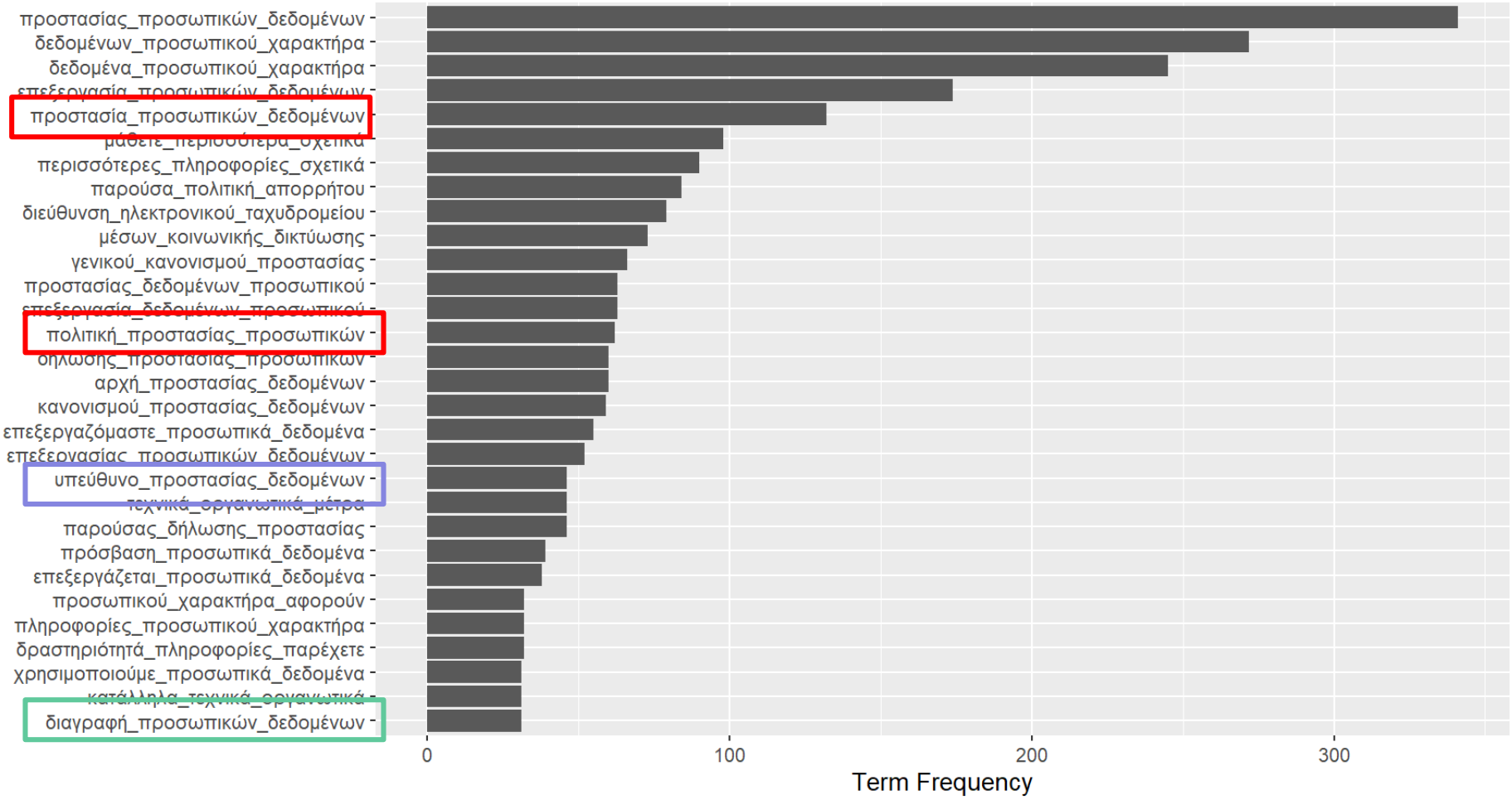
- Privacy policy length:
  - average: 5,079 words
  - shortest: 550 words (inewsgr.com)
  - longest: 49,764 words (microsoft.com)
  - 4 privacy policies have a text with less than 1,000 words
  - 5 privacy policies have a text with more than 10,000 words

Policies with update date in 2023	Policies with update date in 2022	Policies with update date in 2020-2021	Policies with earlier update date	Policies with no indication of update date
20	12	6	11	25

Most frequent 1-grams in policies



Most frequent 3-grams in policies



University of Cyprus

GDPR user rights presence

- Went though the text of 12 privacy policies in the dataset
  - created a manually annotated dataset for those
- Relying on a set of keywords
- Created a lexicon in Greek
  - compared it with terms in: Evangelia Vanezi et al.. 2021, CompLicy: Evaluating the GDPR Alignment of Privacy Policies-A Study on Web Platforms, International Conference on Research Challenges in Information Science (RCIS)
- Many rights may be combined in one sentence
- The right that is less frequent to encounter is the right to avoid a

GDPR user right	Number of policies containing the right's terms	% of policies containing the right's terms
the right to information	32	43.2%
the right of access	39	52.7%
the right to rectification	31	41.9%
the right to erasure	51	68.9%
the right to restriction of processing	38	51.4%
the right to data portability	40	54.1%
the right to object	51	68.9%
the right to avoid automated decision-making	20	27.0%

Main conclusions

- Some policies of websites with international presence are making references to privacy laws of other countries (besides GDPR)
- Introduced a privacy policies dataset of 74 privacy policies in the Greek language
- Will make the dataset publicly available
- In the future:
  - will gather more privacy policies in Greek
  - will expand the manual annotation using expert annotators
  - will improve the rights detection process

Thank you for your attention!

Questions ?

Contact @
gkapi@ucy.ac.cy

University of Cyprus