SST19
CHAM, NOVEMBER 11-12, 2019

Twitter as a potential complementary source to traditional survey data: the case of Switzerland

Atelier IX: Innovationen in der Datenerhebung/
Innovations dans les méthodes
de collecte des données

PhD Maud Reveilhac
University of Lausanne
LINES
&
Prof. Stephanie Steinmetz
University of Lausanne
LINES

UNIL | Université de Lausanne

PRESENT STUDY

Background:

- **Increased use** of text based social media, especially in micro-blogs
- **Opportunities for social statistics** about opinions, attitudes & sentiments
- **Potential complement** for official statistics (especially surveys)

Why Twitter?:

- used by **many people** (~700'000 in Switzerland)
- much of its messages are **publicly available**
- receives **intense attention** from the established media (and politicians)

Corpus:

200 most recent tweets from each of the 200'000 Twitter accounts identified as Swiss

Research questions:

- A) What are the main **challenges in terms of data collection & population bias**?
- B) What are **the potentials of Twitter data** to complement official statistics?

METHODS OF DATA COLLECTION

Identification of Swiss accounts:

- Snowball procedure (followers-followees: politicians & media) + scraping by location
- Regular expression to detect “swissness” (e.g. location, hashtags, mentions, etc.)

Estimation of socio-demographics:

- Gender (using name from official statistics & face recognition with *Face++*[1])
- Age groups (idem)
- Political affiliation (using classification model SVM on the basis of political tweets with *Lightside*[2])

1 https://www.faceplusplus.com/

2 http://www.cs.cmu.edu/~cprose/LightSIDE.html

METHODS OF ANALYSIS

RQ1 "What are the main challenges in terms of data collection & population bias?"

- Collection techniques (inspired from Daas et al. 2012)
- Distribution of gender/age/political affiliation compared to official statistics

RQ2 "What are the potentials of Twitter data to complement official statistics?"

- #hashtag network (using visualisation with *Gephi*)
- Comparison of topic content with themes from official statistics
- Distribution of the topic content according to gender/age/political affiliation compared to official statistics (using LDA from *Mallet* [1])

1 http://mallet.cs.umass.edu/

DATA COLLECTION PROCEDURES

Collected via *Application Programming Interface* (API)
-> Streaming was excluded
-> REST approach was used (user identifier as point of entry to users & followers)

Depth	Total number of unique user ID's	Total number of valid* unique user ID's collected	Total number of unique Swiss ID's (% of total valid ID's)	Number of new followers ID's of the Swiss ID's
0 - politician	1	1 (100%)	1 (100%)	46'515
0 - politicians	166	166 (100%)	166 (100%)	173'733
0 - media	41	41 (100%)	41 (100%)	421'695
1 - politician	46'515	25'615 (55%)	7'440 (29%)	1'506'450
1 - politicians	173'733	87'289 (50%)	2'514 (3%)	481'007
1 - media	421'695	248'247 (60%)	53'040 (21%)	1'161'162
2 - politician	1'506'450	1'337'723 (89%)	120'210 (9%)	STOP here (exponential)
2 - politicians	481'007	412'403 (86%)	24'328 (6%)	
2 - media	1'161'162	866'787 (75%)	19'894 (3%)	
Total 3 paths			178'556	
Total scraping			13'272	
Total			**191'828**	

* valid implies that the account is still active

Sequence:
collect as many user identifiers as possible >>> identify the Swiss users >>> more messages from the Swiss users

GENDER & AGE ESTIMATION

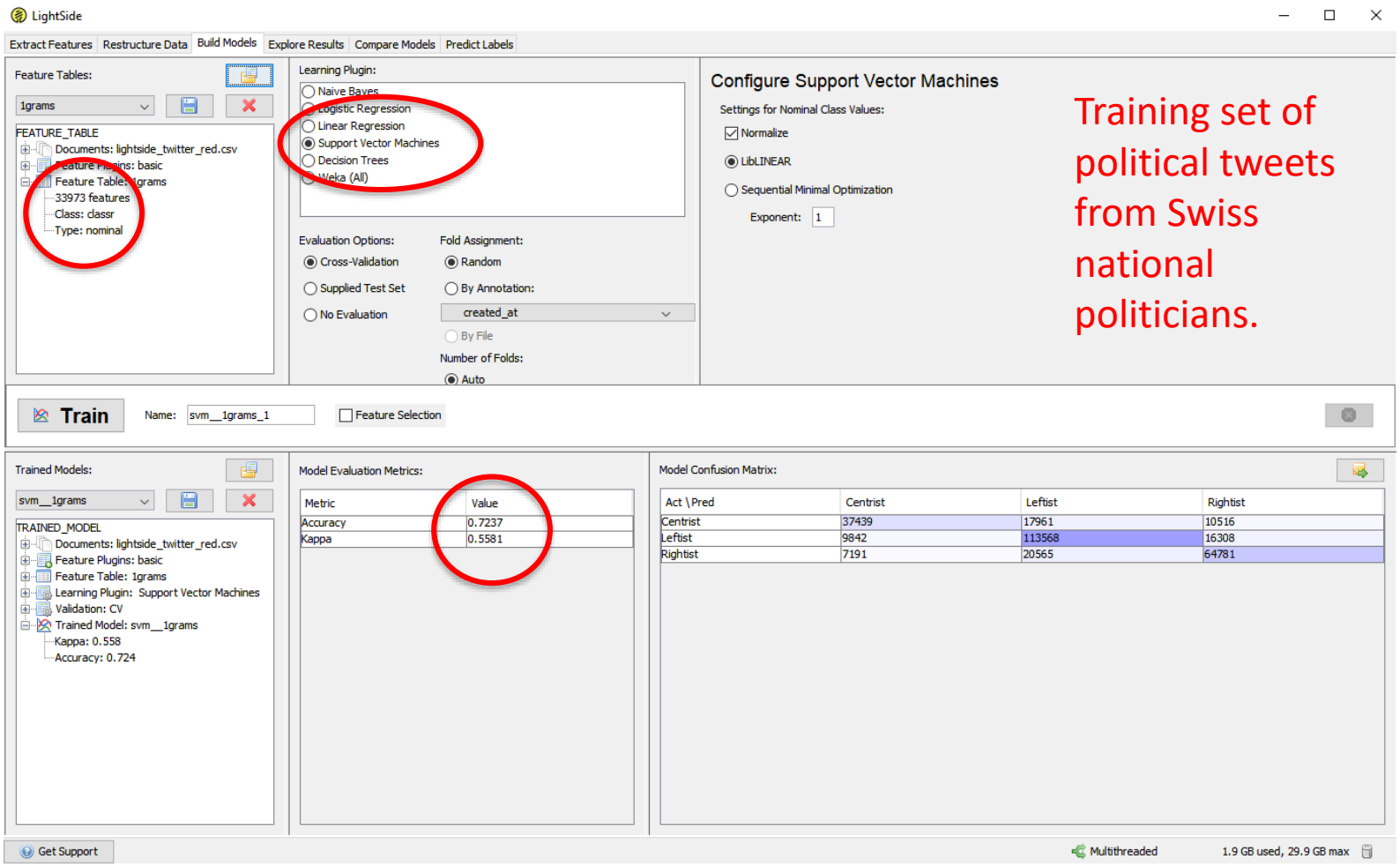
	Gender	Age
"Vornamenhitparaden" (VHP)		
Face recognition (Face++)		

		VHP		
		Woman	Man	Not found
Face++	Woman		Face++	Face++
	Man	Face++		Face++
	Not found	VHP	VHP	Not found

Overall congruence for gender:
~90%

Overall congruence for age:
~40%

POLITICAL AFFILIATION ESTIMATION



Training set of political tweets from Swiss national politicians.

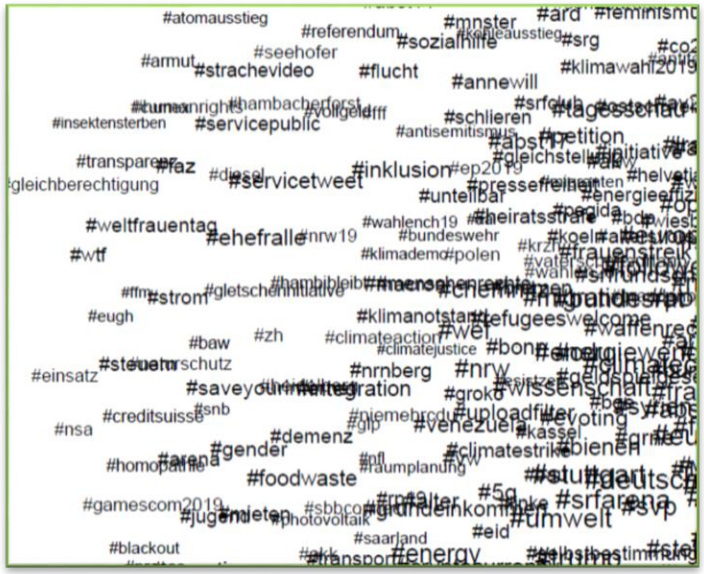
RQ1: POPULATION BIASES

		Twitter	Official stats
Gender			
	Male	52%	49.6%
	Female	48%	50.4%
Age			
	<30	33%	32%
	30-50	27%	25%
	50-70	33%	29%
	>70	7%	14%
Political affiliation			
	Leftist	76%	35%
	Centrist	12%	10%
	Rightist	12%	45%

Source: https://www.bfs.admin.ch/bfs/fr/home/statistiques/population/effectif-evolution/population.html

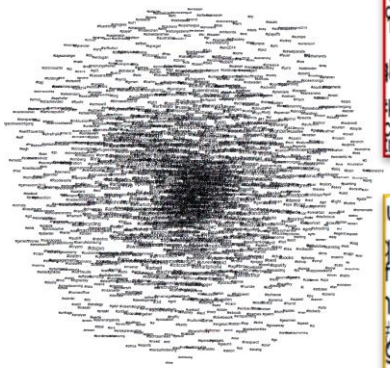
Source: https://www.pxweb.bfs.admin.ch/pxweb/fr/px-x-0102020000_103/px-x-0102020000_103/px-x-0102020000_103.px

RQ2: #HASHTAG NETWORK



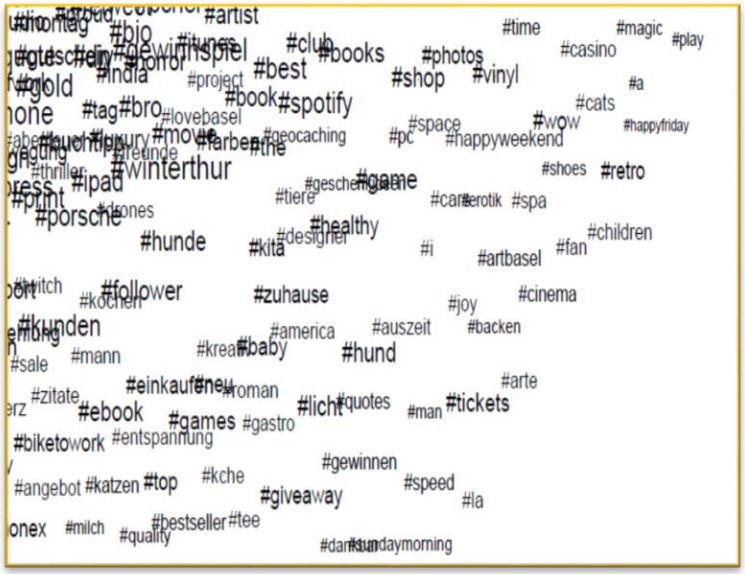
Politics

Sports

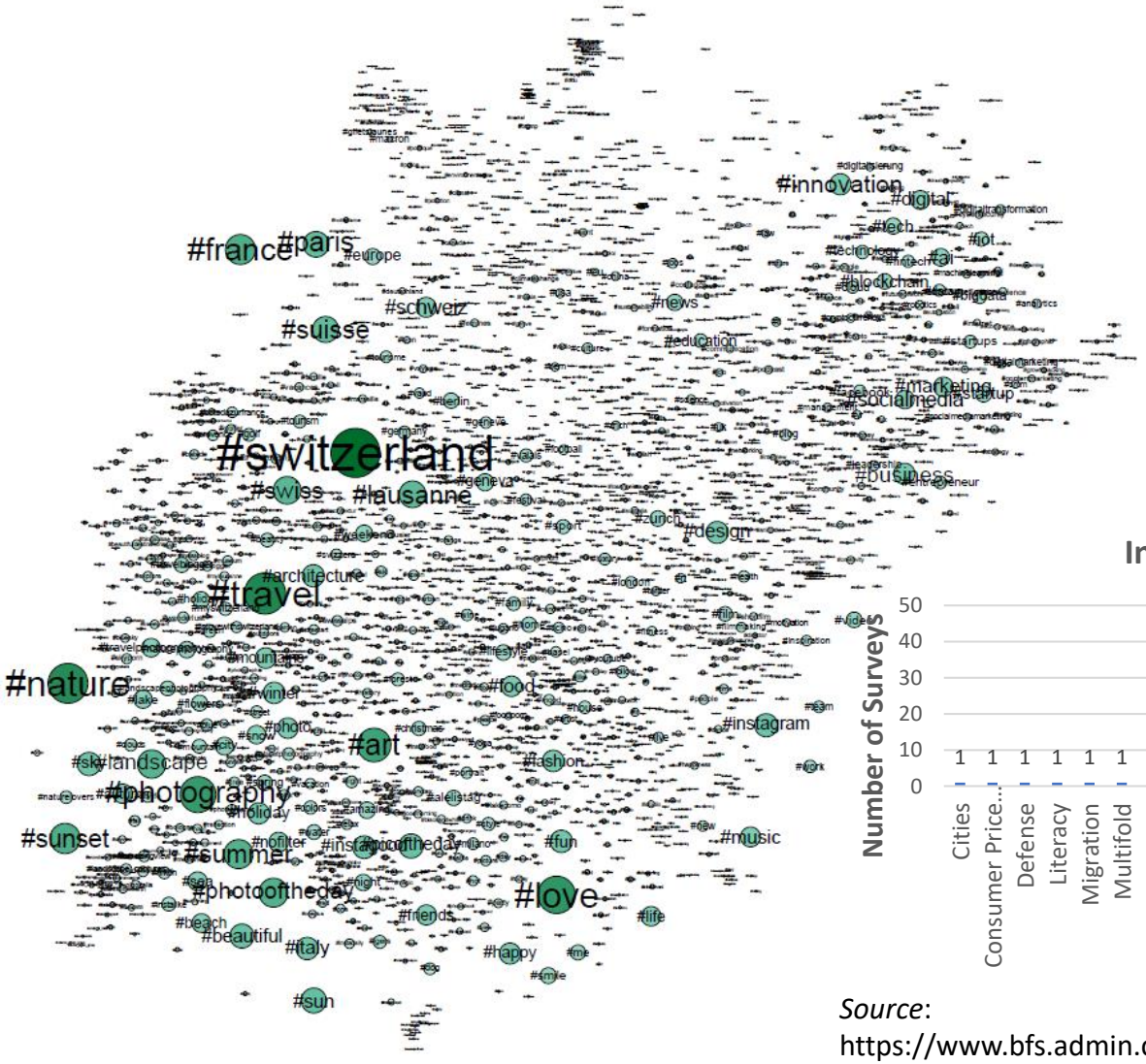


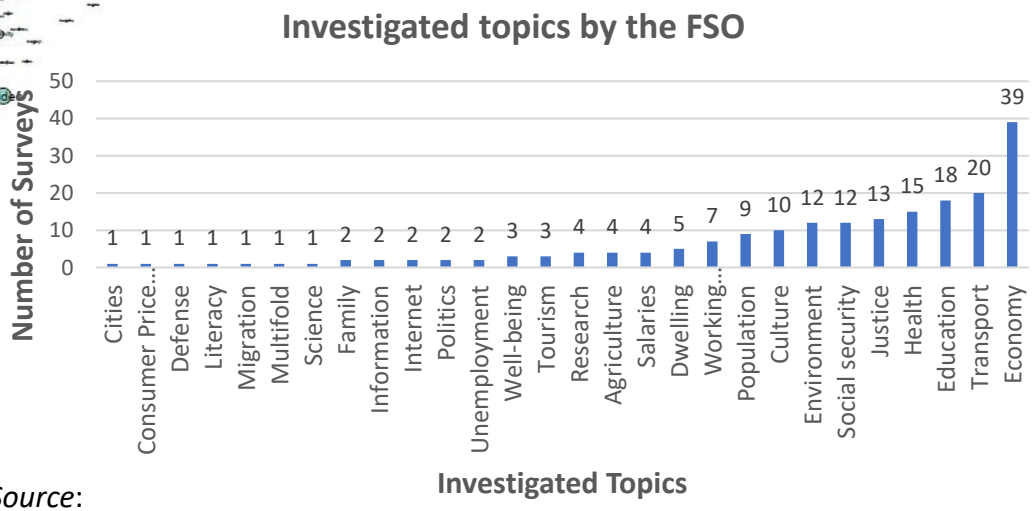
Technology

Leisure



RQ2: #HASHTAG NETWORK (TOP 10'000)





Source:
https://www.bfs.admin.ch/bfs/en/home/basics/surveys.html

UNIL | Université de Lausanne

RQ2: #HASHTAG CLASSIFICATION MODEL?

	%	N
nbr of tweets in total		**22'447'274**
nbr of tweets with one **#hashtag**	15%	
nbr of tweets with 2 or more **#hashtags**	19%	
nbr of tweets without **#hashtag**	81%	
nbr of tweets with one **@mention**	27%	
nbr of tweets with 2 or more **@mentions**	14%	
nbr of tweets without **@mention**	59%	
nbr of tweets with one **link**	33%	
nbr of tweets with 2 or more **links**	67%	
nbr of tweets without **link**	33%	

What is a tweet? short text messages (max 280 characters)

Who tweets? anyone who creates an account and gets a unique identifier

Other features? username, full name, location, biography, age, pictures, links, privacy settings, dates, etc.

What kind of relationships? sending (direct) messages, following users & being followed (not reciprocal)

What conventions? @replying, @mentioning, #hashtagging, forwarding (RT), linking (http://)

RQ2: TOPIC MODELLING

Preprocessing steps:

- Pooled (1 document = 1 user)
- Cleaned and POS-tagged[1] corpus keeping only nouns and adjectives to conduct topic modelling

1 We used *TreeTagger* (Schmid 1994)

Automated content analysis:

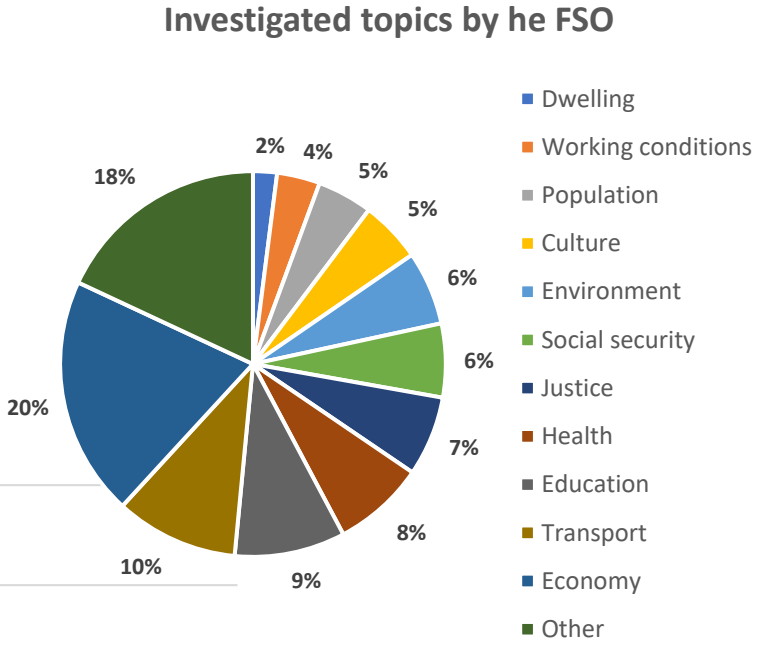
Analysis of the **manifest and latent content** of a body of textual material through classification of its **key themes** in order to ascertain its **meaning** (Blei 2012).
Topic models[2] maximize:

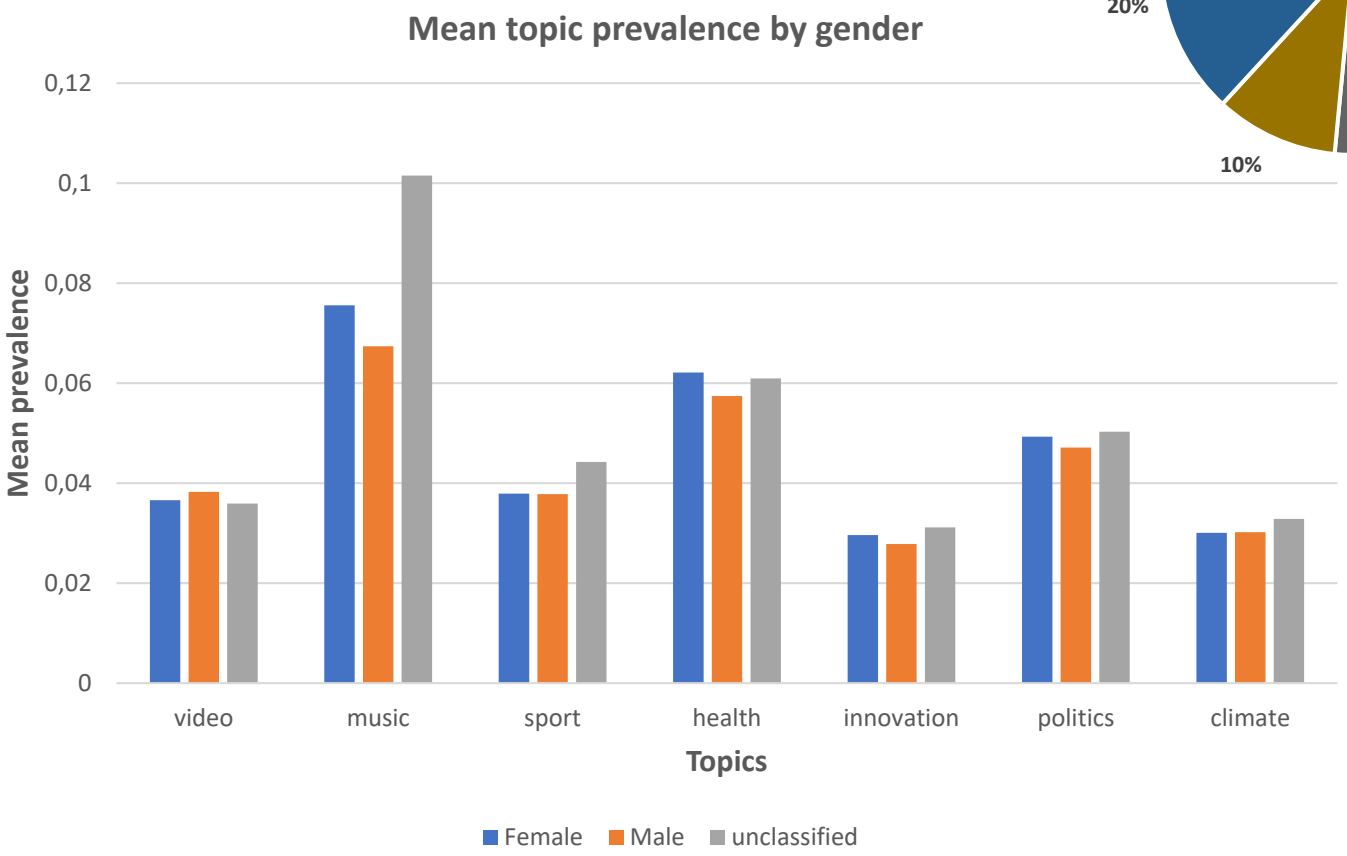
$$p(\text{topic} \mid \text{document}) \cdot p(\text{word} \mid \text{topic})$$

We used N=50 topics. About 80% of them are interpretable. We discuss a subset.

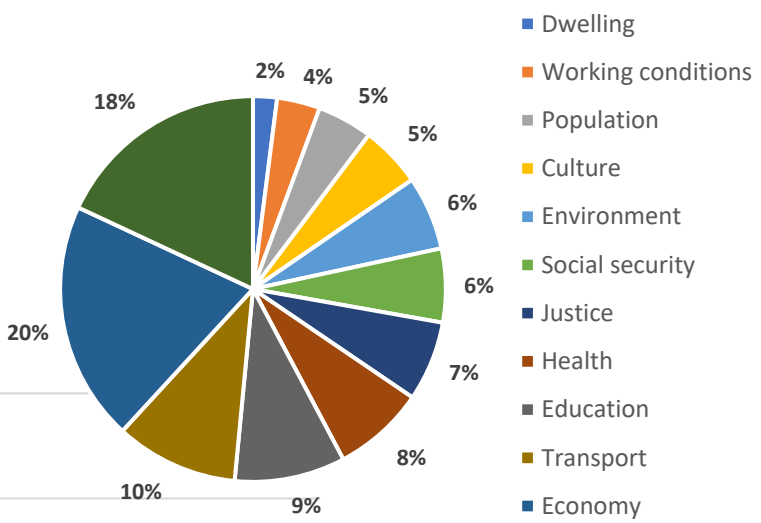
2 We used *Mallet* (http://mallet.cs.umass.edu/topics.php)

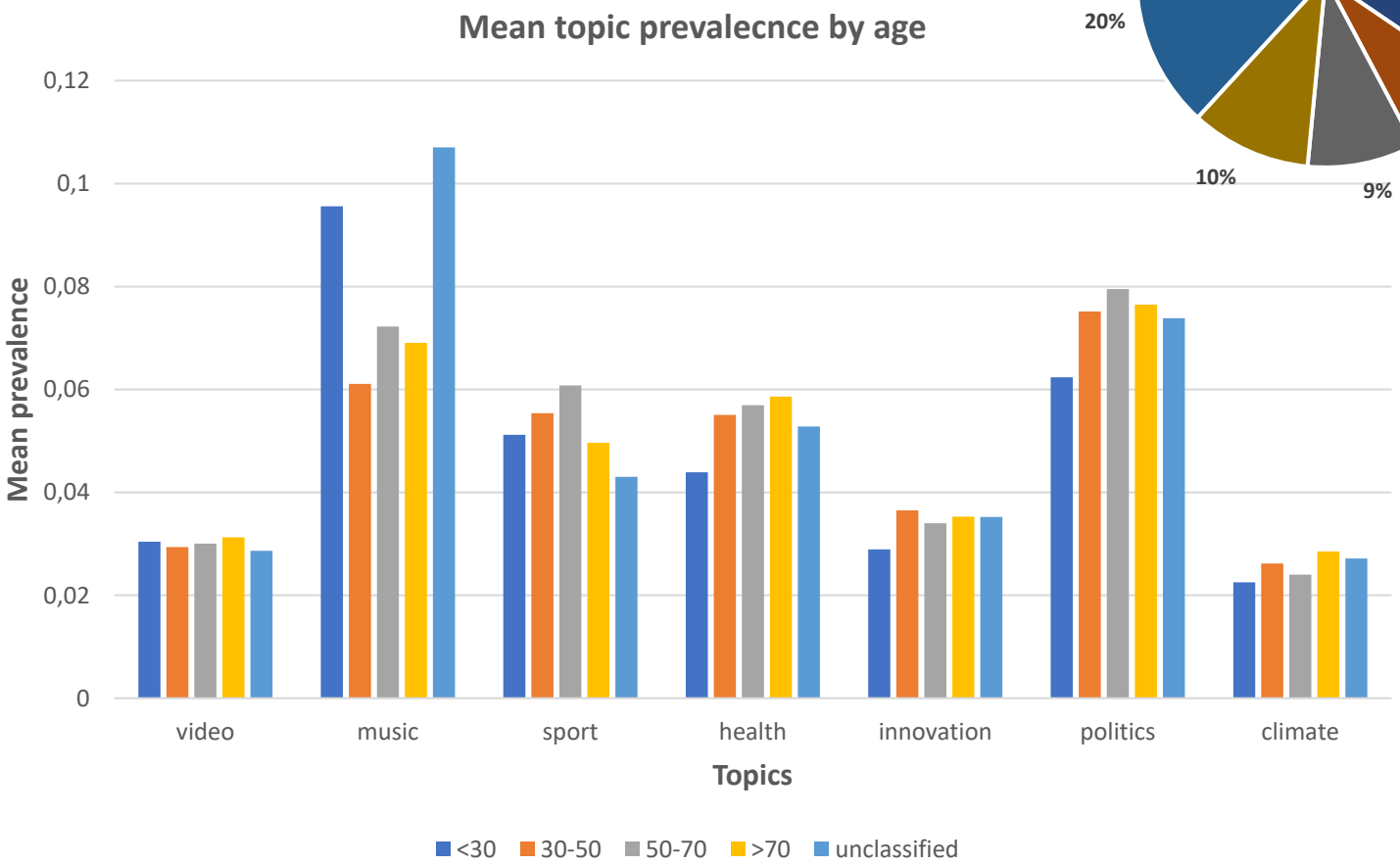
RQ2: TOPICS BY GENDER



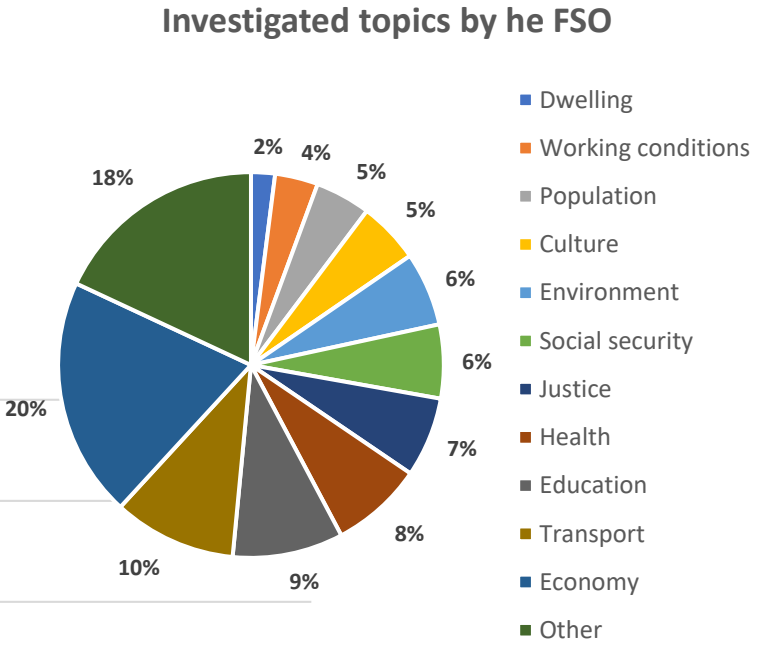


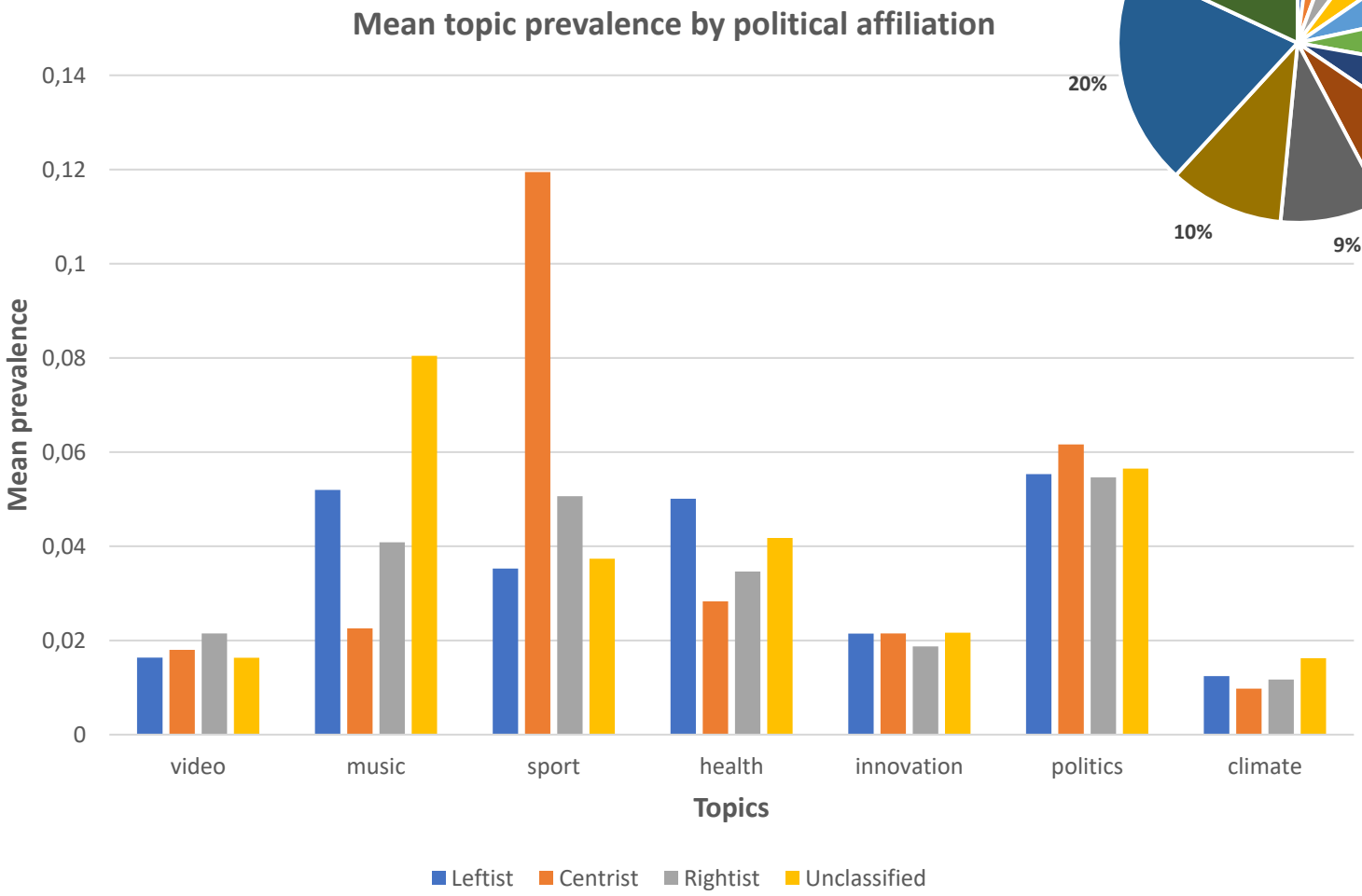
RQ2: TOPICS BY AGE





RQ2: TOPICS BY POLITICAL AFFILIATION





CONCLUSION

Data collection and representativity issues:

- relating Twitter-based findings to the opinion of the Swiss population as a whole demands using **additional steps** (e.g., non-individual agents recognition, level of activity analysis, etc.).
- this is caused by these facts (see Sen et al. 2019):
  - **not every Swiss citizen is active on Twitter**
    -> platform selection (coverage error)
  - **not all Swiss users can be identified based on the information they provide**
    -> entity selection (sampling error) + entity augmentation/reduction (processing error)
  - **not every user has the same level of activity**
    -> signal selection (measurement error)

Potentials of Twitter data to complement official statistics

- **topics of interest** for official statistics are discussed, especially those related to work and politics (spare time activities and events are also interesting options)
- **informative *versus* non-informative** messages are hard to discriminate (implies pre-selection of tweets, for e.g. based on the occurrence of topic specific words).

OUTLOOK

- refinement of data collection method
- further focus on background characteristics of Swiss Twitter users
- further improvement of automatic classification of topics (e.g. multi-language)
- mining of sentiment (especially aspect-based)
- focus of links contained in tweets

REFERENCES

- Blei, David M.. 2012. “Probabilistic Topic Models”. *Communications of the ACM*, 55(4): 77-84.
- Daas, P., Roos, M., van de Ven, M., & Neroni, J. (2012). “Twitter as a potential source for official statistics in the Netherlands”. *Paper for the 67th AAPOR 2012 conference* (pp. 17-20).
- Schmid, Helmut. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees". *Proceedings of International Conference on New Methods in Language Processing*.
- Sen, I., Floeck, F., Weller, K., Weiss, B., & Wagner, C. (2019). “A Total Error Framework for Digital Traces of Humans”. arXiv preprint arXiv:1907.08228.