Joint Estimation of Sentiment and Topics in Textual Data

COMPTEXT 2022 Pre-Conference Workshop

Christian Pipal

University of Amsterdam

`github.com/cpipal/comptext-workshop2022`


HELLO
ITS ME
memegenerator.net


I DON'T KNOW WHO YOU ARE...
WHO ARE YOU?

NB!

We will need to compile source code in the 2nd half of the workshop. Please make sure that your laptop is set up to do so!

- Windows: Download and install RTools
- Mac: Download and install Xcode
- Instructions: https://clanfear.github.io/CSSS508/docs/compiling.html

Plan for today

2 blocks, each block consists of:

- Short introduction
- Coding examples
- Coding challenge

11:00 - 11:45: Intro / recap dictionaries and topic models

11:45 – 12:00: Break

12:00 – 11:45 Joint sentiment-topic modelling with *sentitopics*

A bag of words

When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will

docs	made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3

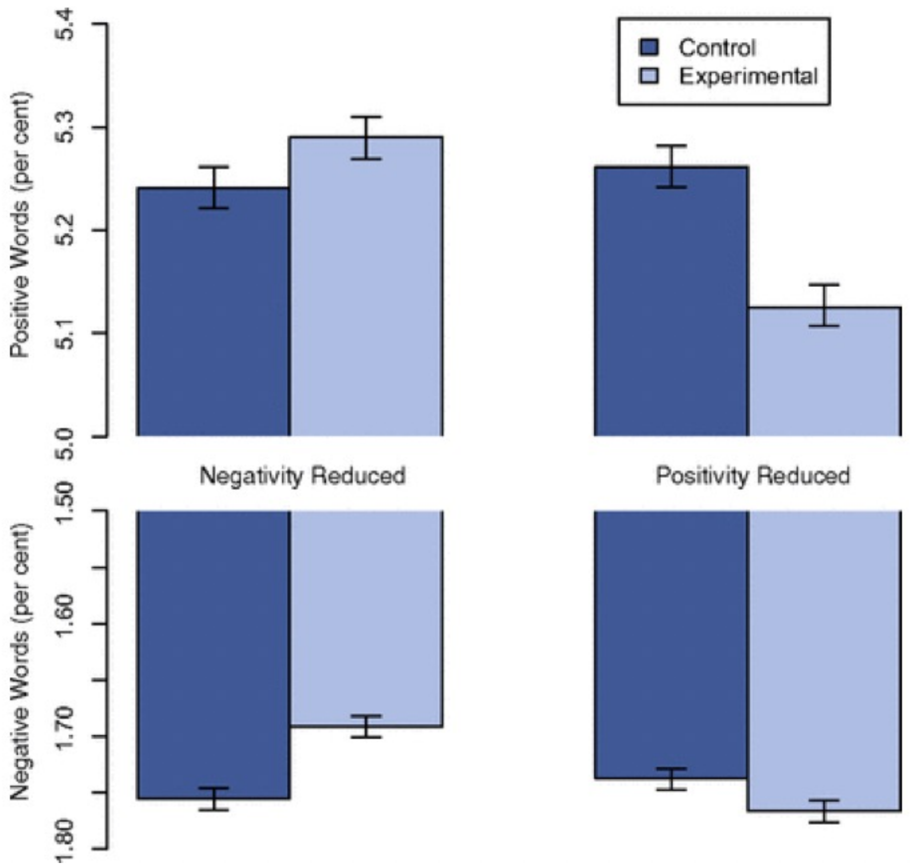
Dictionary Methods

Classifying documents when categories are known:

- List of words that correspond to each category:
    - Positive or negative, e.g. for sentiment
    - Sad, happy, angry… for discrete emotions
    - Insight, causation, discrepancy… for cognitive processes
    - Sexism, homophobia, racism… for hate speech
    - And many others: see LIWC, VADER, SentiStrength
- Count number of times they appear in each text
- Normalize by document length (optional)

- **Validate, validate, validate**

Example: Emotional Contagion on Facebook



Kramer et al. 2014

Example: Anger in US Presidency candidate speeches

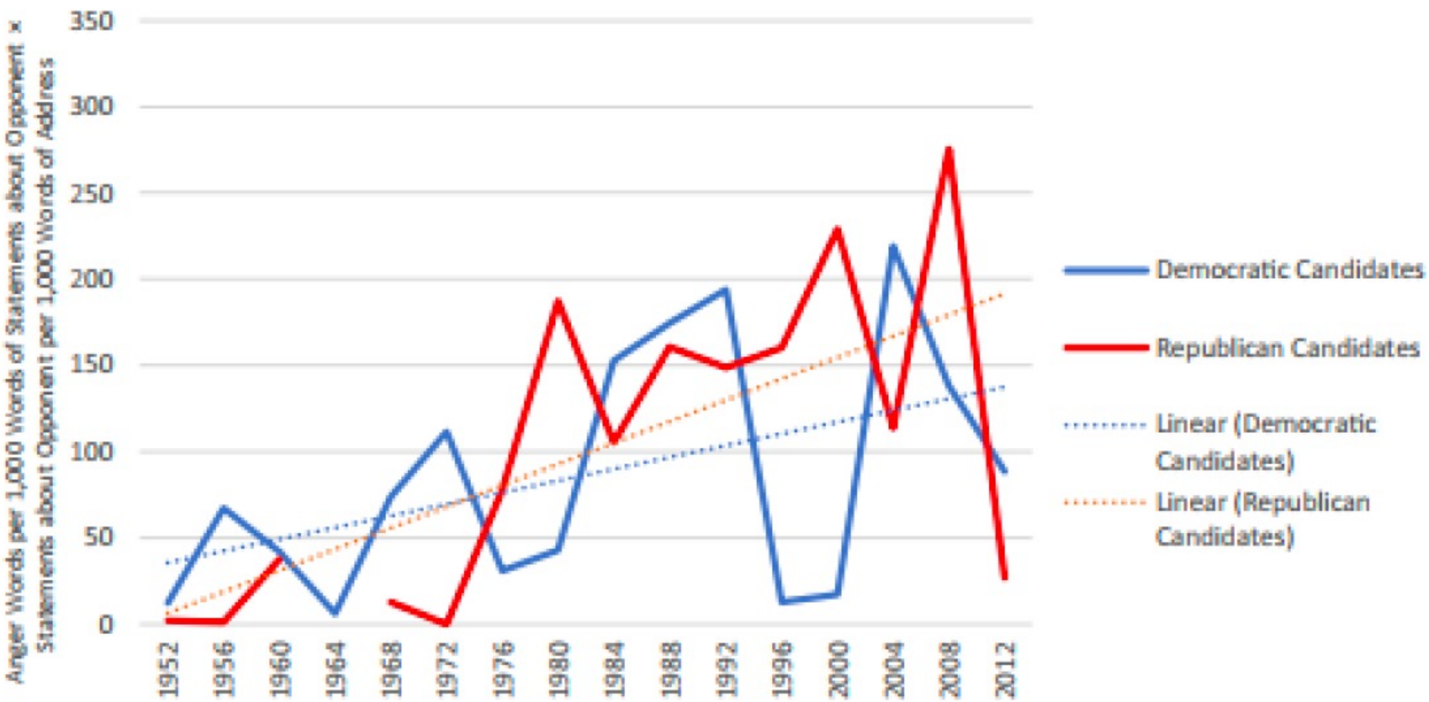


FIGURE 11. Standardized Volume of Anger Content in Statements about Opponent, Democratic and Republican Candidates, 1952–2012.

Rhodes and Vayo 2019

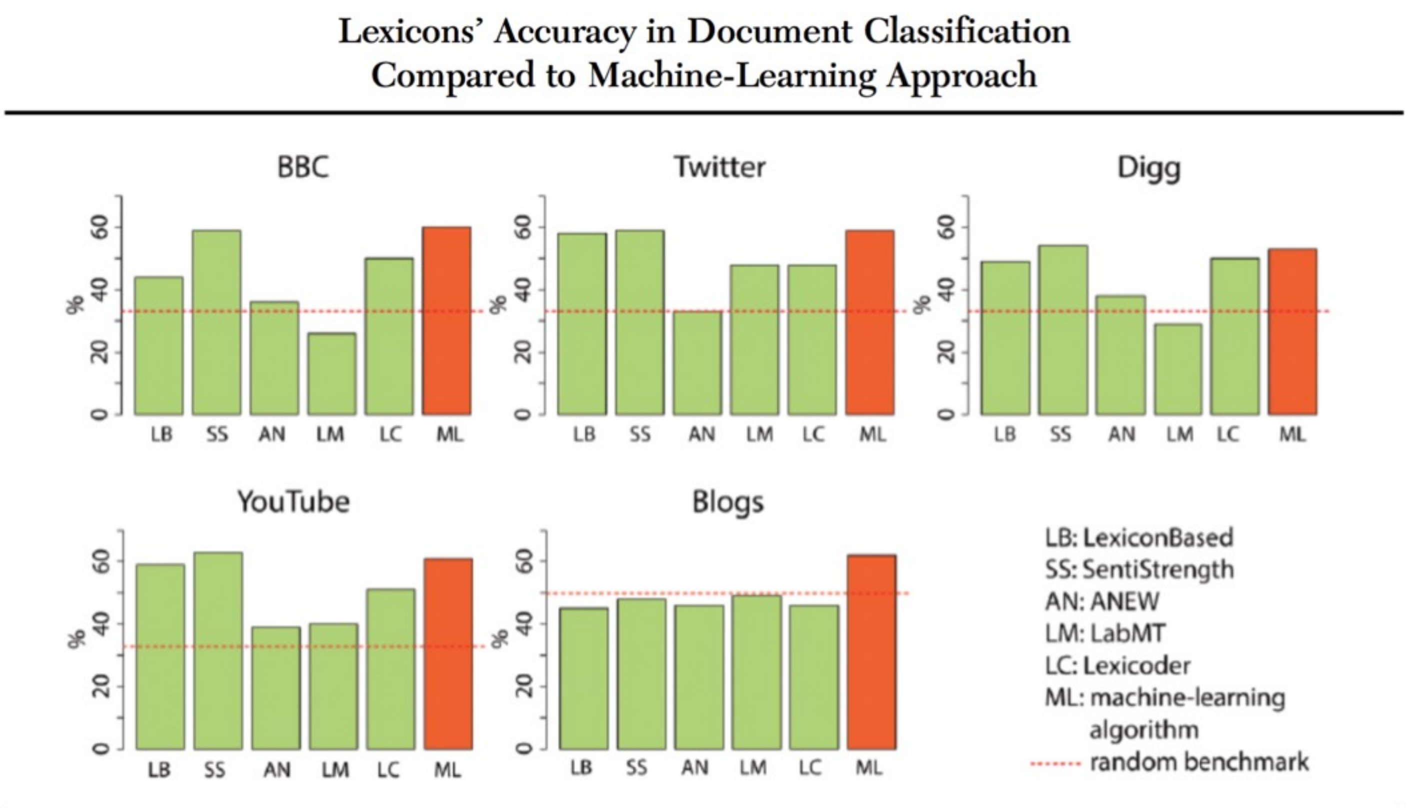
Potential advantage: multi-lingual

APPENDIX B
DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS

	NL	UK	GE	IT
Core	elit*	elit*	elit*	elit*
	consensus*	consensus*	konsens*	consens*
	ondemocratisch*	undemocratic*	undemokratisch*	antidemocratic*
	ondemokratisch*			
	referend*	referend*	referend*	referend*
	corrupt*	corrupt*	korrupt*	corrot*
	propagand*	propagand*	propagand*	propagand*
	politici*	politici*	politiker*	politici*
	bedrog	*deceit*	täusch*	ingann*
	bedrieg	*deceiv*	betrüg*	
			betrug*	
	verraa	*betray*	*verrat*	tradi*
	verrad			
	schaam*	shame*	scham*	vergogn*
			schäm*	
	schand*	scandal*	skandal*	scandal*
	waarheid*	truth*	wahrheit*	verità
	oneerlijk*	dishonest*	unfair*	disonest*
			unehrlich*	
Context	establishm*	establishm*	establishm*	partitocrazia
	heersend*	ruling*	*herrsch*	
	capitul*			
	kapitul*			
	kaste*			
	leugen*		lüge*	menzogn*
	lieg*			mentir*

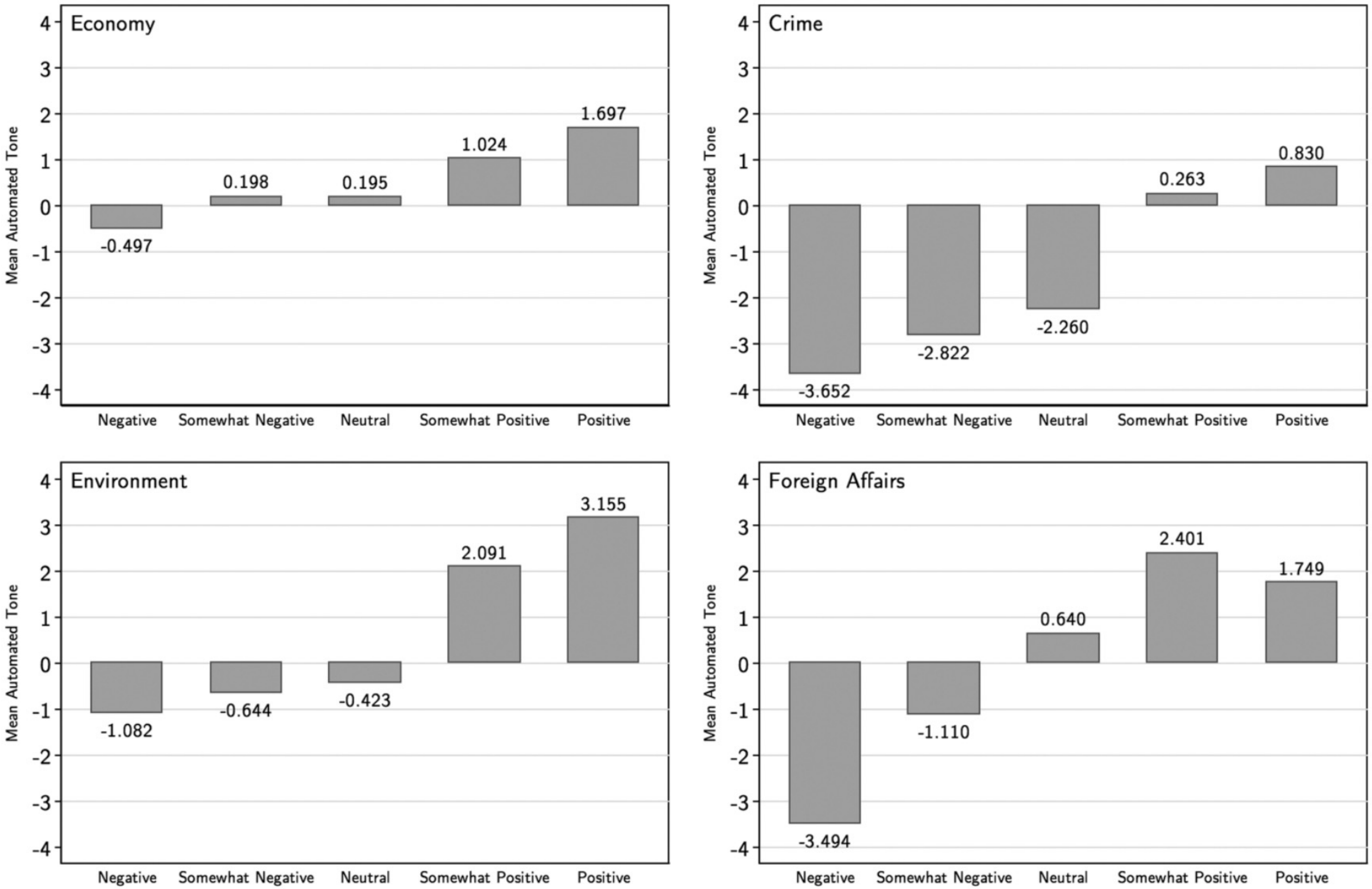
Rooduijn and Pauwels 2011

Potential disadvantage: context specific



Gonzáles-Bailón and Paltoglou 2015

Potential disadvantage: topic specific



Young and Soroka 2012

Key issues

- Validity: Is the dictionary‘s category scheme valid?
- Recall: Does the dictionary identify all my content?
- Precision: Does the dictionary identify only my content?

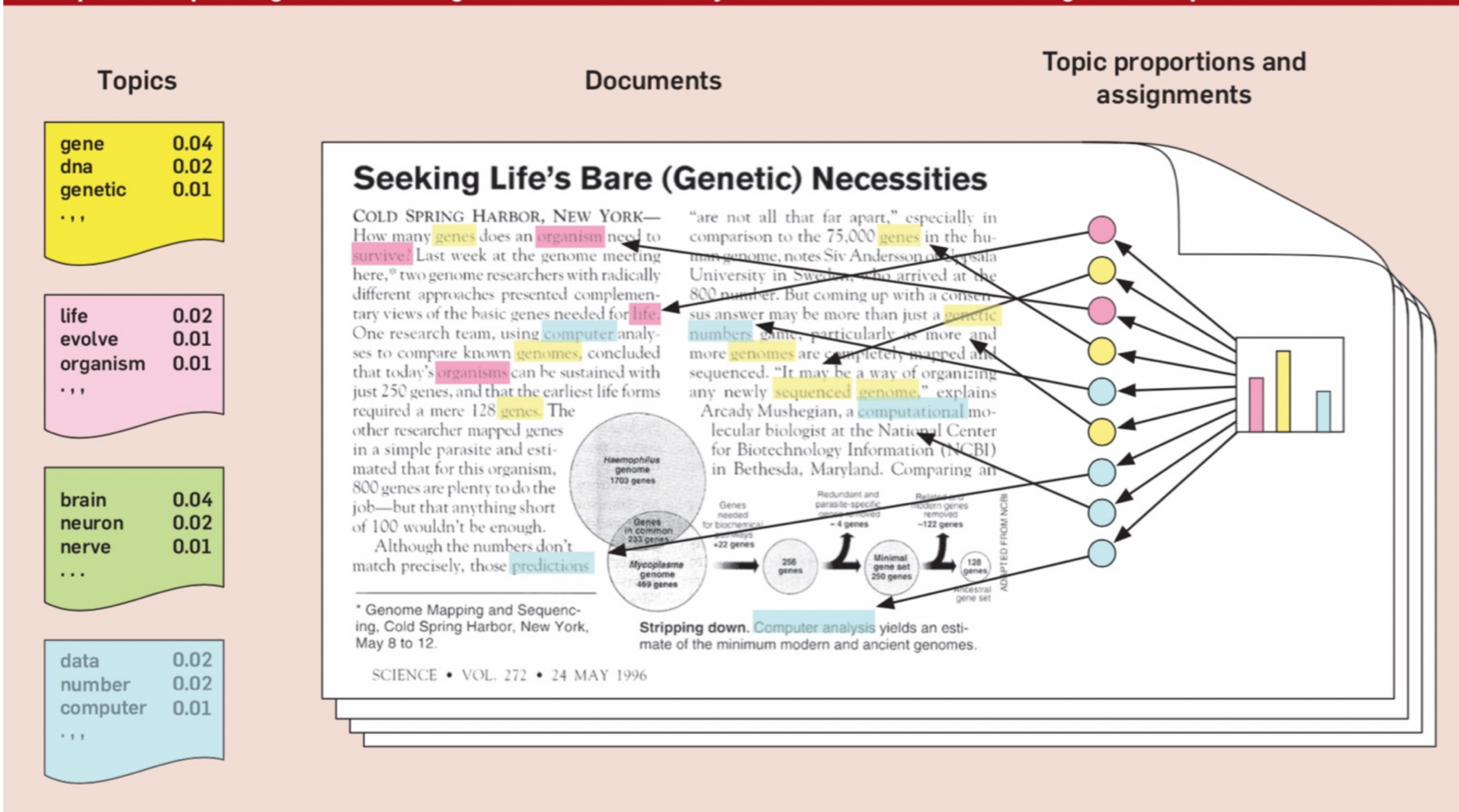
CODING

Topic Models

- Algorithms for discovering the main themes in an unstructured textual corpus
- No prior information about content needed, no training set
  - You only need a decision on k (number of topics)
- Latent Dirichlet Allocation (LDA): assumes that topics are not correlated
- A generative model about how the texts in a corpus where created:
  - Each topic is a distribution over a fixed vocabulary
  - Each text is a collection of words, generated from a multinomial distribution for each topic

LDA

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



CODING

Dictionaries can work fine, but…

- Domain-specific
  - Low agreement with each other (van Atteveldt et al. 2021)
  - Choose your dictionary → choose your results (Pipal et al. 2022)

- Solution: domain-specific dictionaries, e.g. Rauh 2018, Rheault et al. 2016
  - Cost?

- Topic-specific?
  - Sentiment scores differ per topic (e.g. Young and Soroka 2012)
  - Polariy of a word might differ between topics
  - Many textual sources are not just about one topic (e.g. leader speeches)

Joint Sentiment Topic Model

- Branch of sentiment-topic models that extend on LDA

→ LDA assumes texts have topic distributions and draw words from the topics

- Sentiment-topic models add extra layer. Document has a topic distribution **and** sentiment distribution for every sentiment category

Joint Sentiment Topic Model

Example: Smartphone review

- 4/5 stars: generally positive
- Positive about screen, battery life, and camera
- Negative about plastic cover

Sentiment distribution overall positive. Within positive sentiment topics most likely to be screen battery life, and camera. Within negative sentiment most likely cover.

Joint Sentiment Topic Model / reversed JST

Estimate topics and sentiment simultaneously (Lin et al. 2012)

- **JST**: Mixture of sentiment in text, topics clustered within sentiment

  → **overall text sentiment**

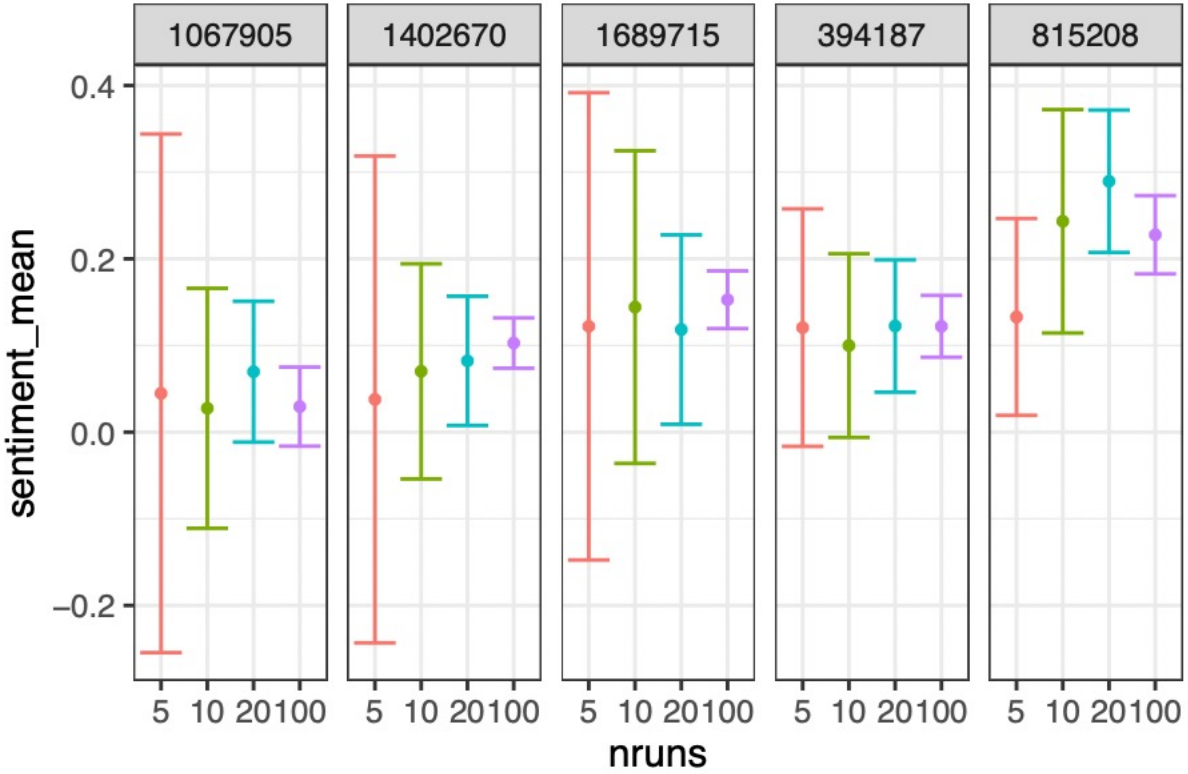
- **reversed JST**: Mixture of topics in text, sentiment clustered within topics

  → **topic-specific sentiment**

Estimation JST/rJST

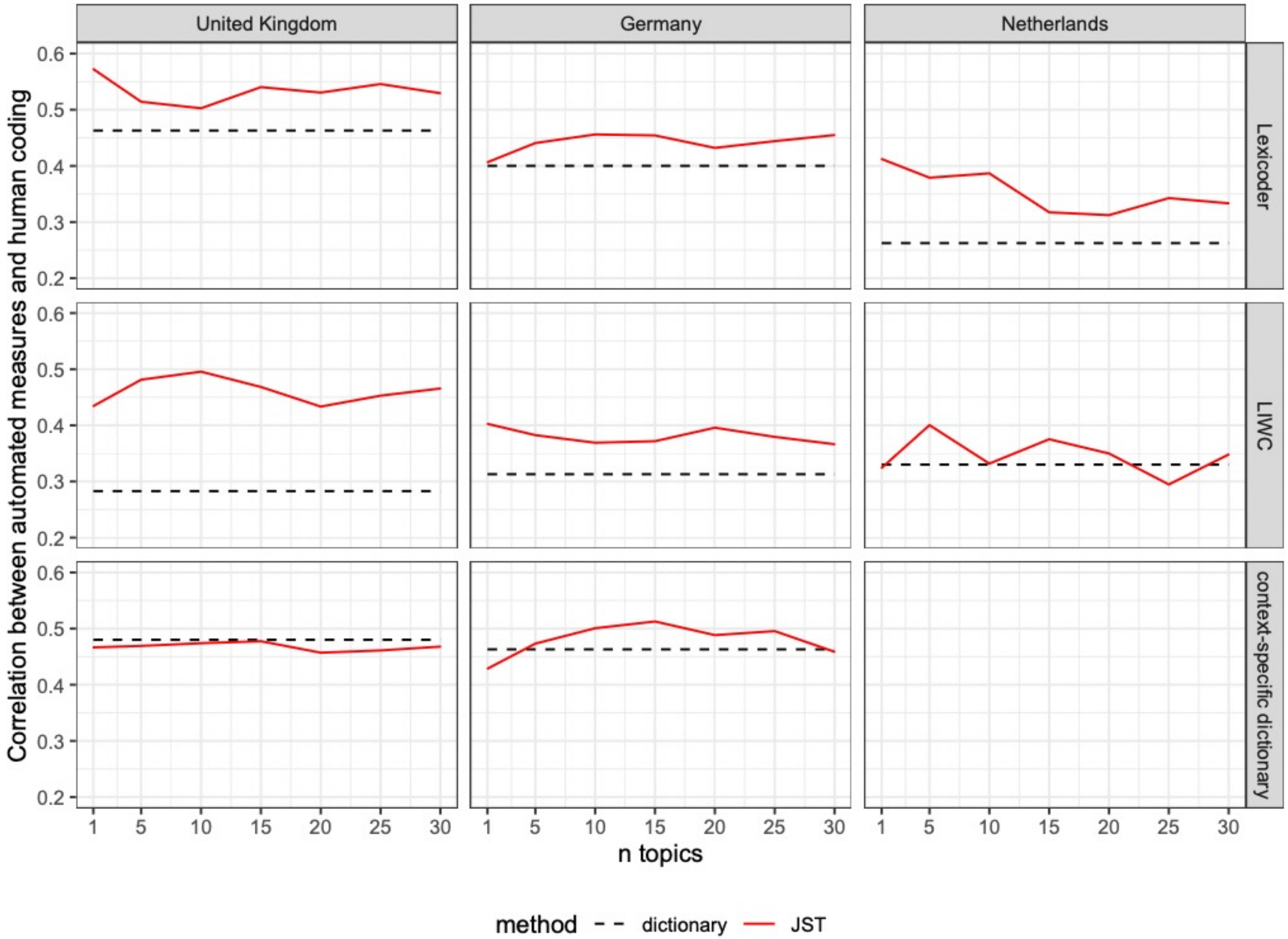
- R-package **sentitopics**
- Prior information from sentiment dictionary (semi-supervised)
- Choose k topics in advance

Variation across model runs



Pipal et al. 2022

Validation with human coding (JST)



Pipal et al. 2022

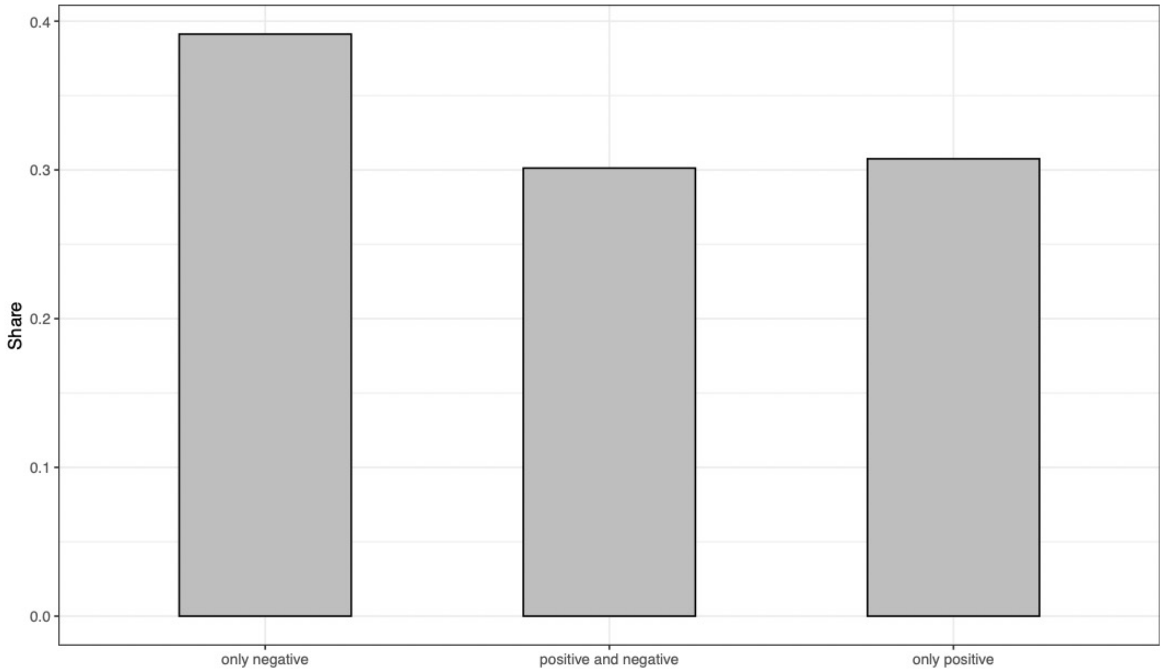
Face validity (rJST)

Armed Forces/Security		
Neutral	Positive	Negative
armi	forc	defenc
afghanistan	secur	royal
defenc	arm	ship
arm	oper	ministri
militari	continu	capabl
personnel	must	navi
troop	train	aircraft
soldier	support	forc
regiment	remain	procur
veteran	also	air
afghan	well	equip
royal	reserv	raf
command	regular	mod
deploy	task	base
ministri	commit	strateg
serv	now	carrier
civilian	effort	shipbuild
battalion	serv	helicopt
war	number	arm
british	howev	militari

European Union		
Neutral	Positive	Negative
european	european	eu
treati	europ	european
union	countri	leav
europ	britain	union
eu	british	agreement
constitut	union	negoti
foreign	germani	uk
articl	franc	deal
maastricht	french	brexit
singl	german	withdraw
referendum	state	vote
negoti	eastern	trade
parliament	nato	remain
commiss	world	custom
british	foreign	singl
institut	presid	futur
vote	eu	exit
veto	western	relationship
sovereignti	join	citizen
council	itali	border

Pipal et al. 2022

Discriminant validity (top 100 rJST words)



Pipal et al. 2022

CODING