

Knowledge discovery from patient forums

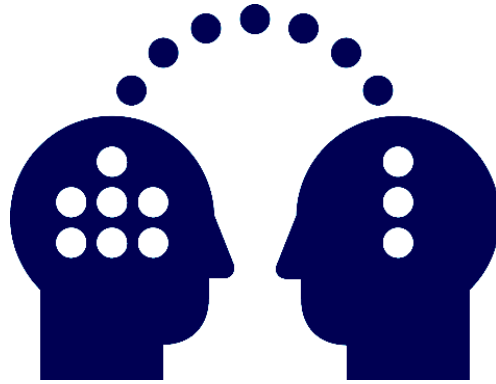
Anne Dirkson

12 June 2019

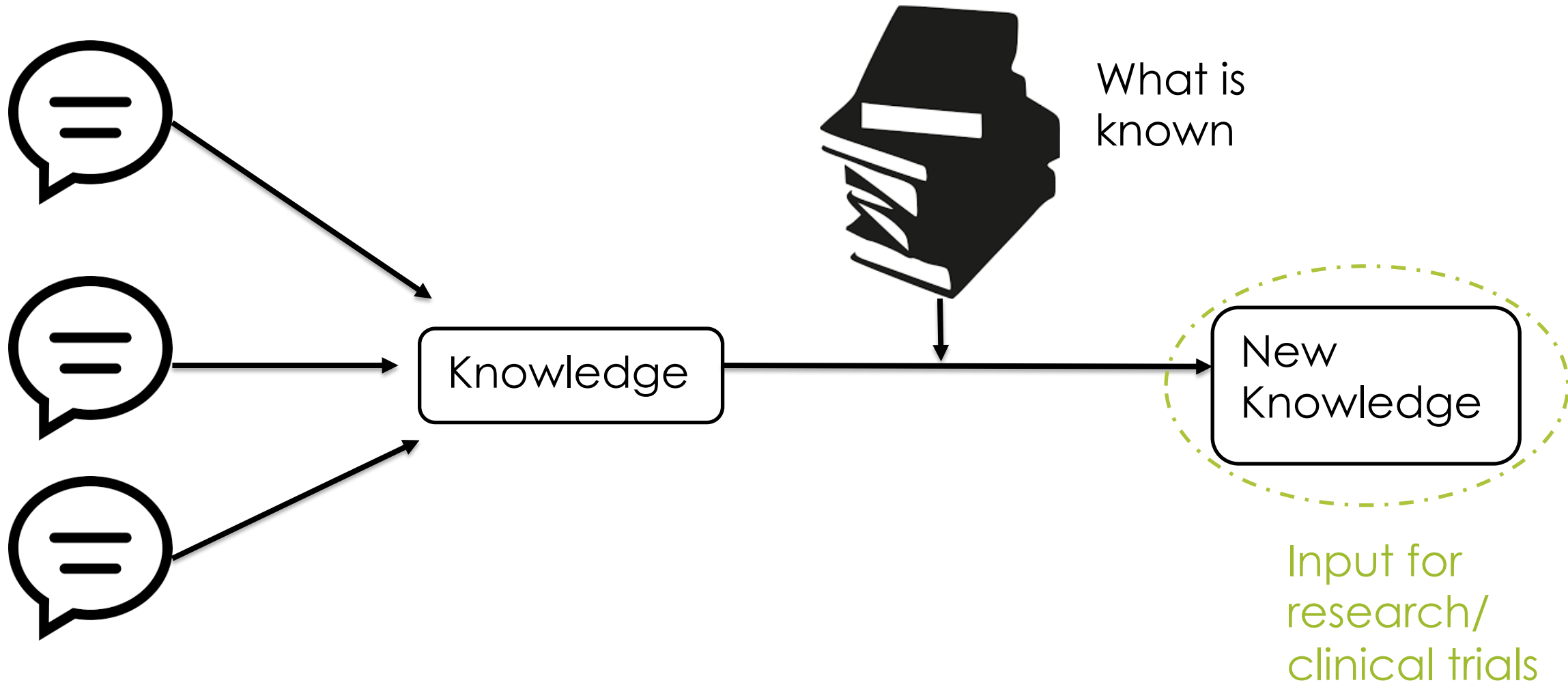


**Universiteit
Leiden**
The Netherlands

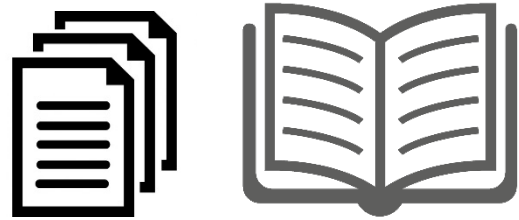
Patient forums are a knowledge gold mine



Medical anecdotes to new knowledge



biomedical literature



patient forums

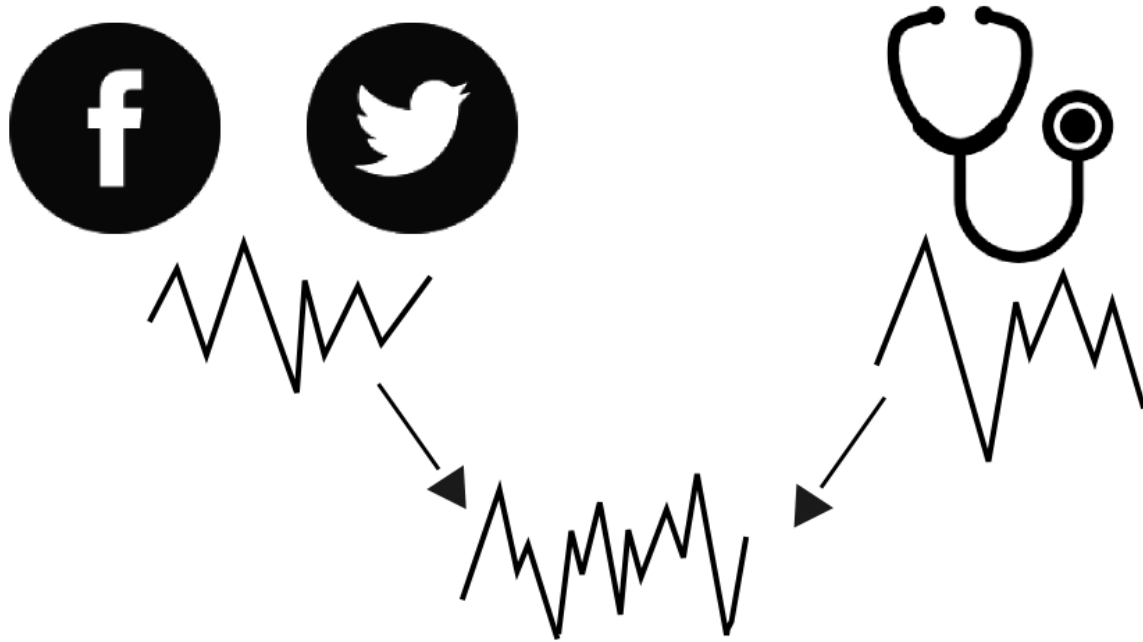


clinical records

Advantages

- + Uncensored
- + Unprompted
- + Volume
- + Available

Patient forums are very noisy





insomnia vs. can't sleep
ablation vs. remove



Challenges for spelling correction

- Lack of domain-specific resources
- Language is dynamic
- Key medical terms are frequently misspelt⁴

Current methods do not suffice

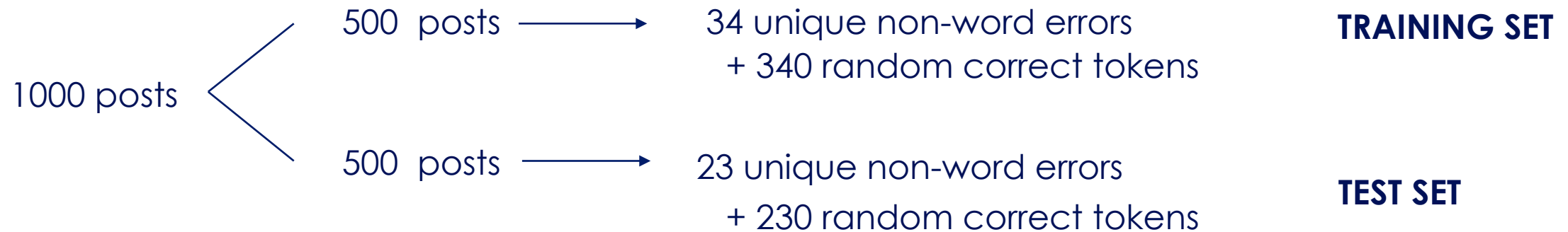
- Traditional methods are unsupervised but rely on dictionaries for detection
- Modern methods are supervised and rely on training data
- **State of the art for social media:** ⁵
 -  Relies on manually created dictionary to detect mistakes
 -  Uses language model of generic Twitter data to correct them

Research questions

1. To what extent can **corpus-driven spelling correction** reduce the out-of-vocabulary rate in medical social media text?
2. To what extent can our **corpus-driven spelling correction** improve accuracy of health-related classification tasks with social media text?

Our data

- Facebook forum for GIST patients
- 36.722 posts

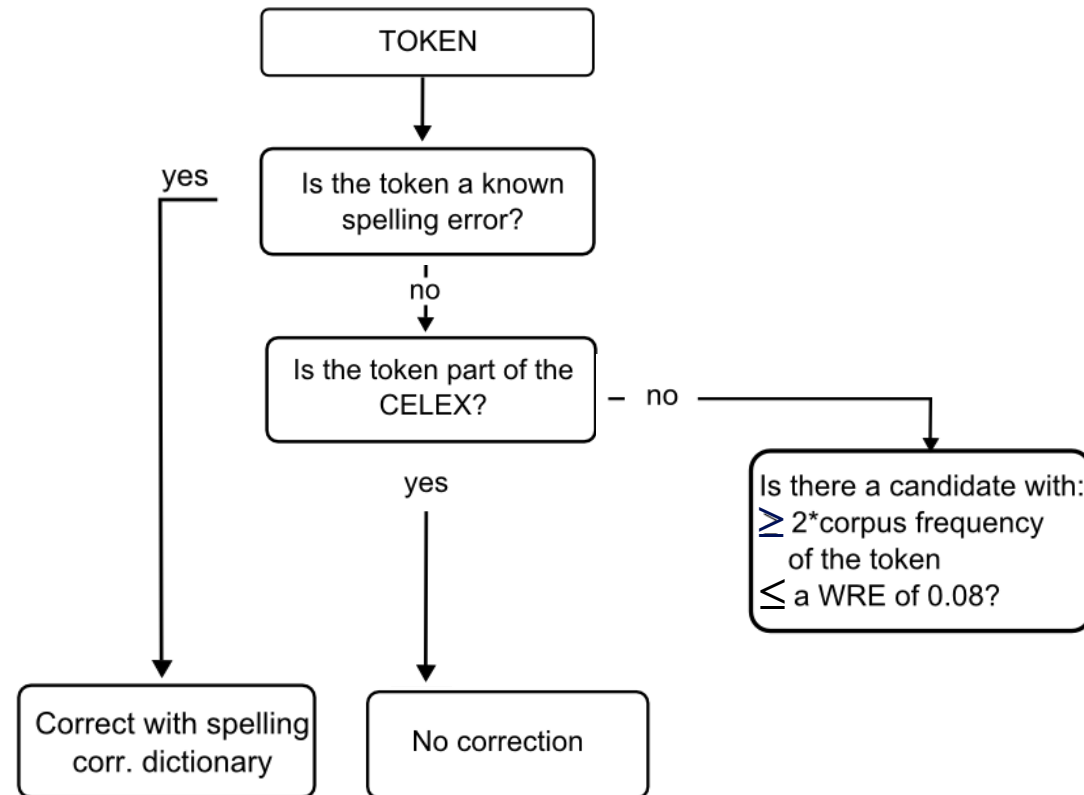


Spelling correction

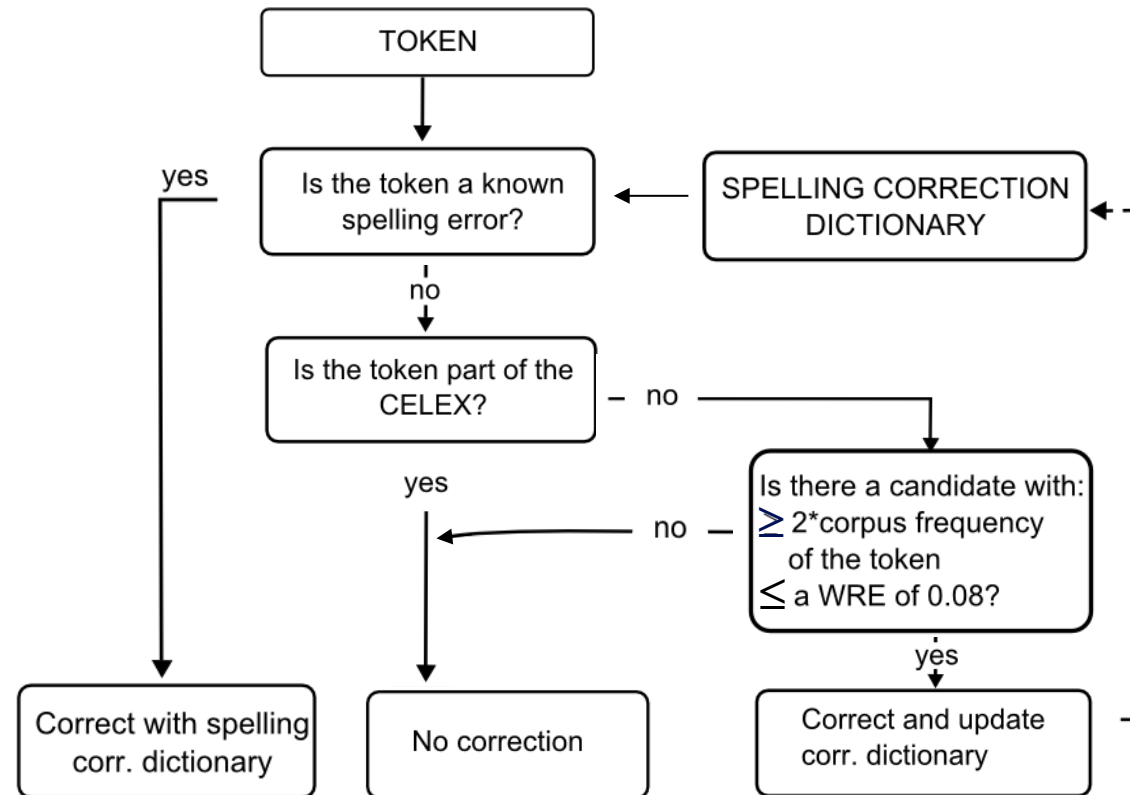
Absolute Edit Dist.	Relative Edit Dist.	Weighted Absolute Edit Dist.	Weighted Relative Edit Dist ⁶	Sarker	TISC
56.6%	56.6%	54.7%	62.3%	20.8%	24.5%

Mistake	Correction						
Gleevac	Gleevec	Gleevec	Gleevec	Gleevec	Gleevec	Colonic	Gleevac
Stomack	Stomach	Stomach	Stomach	Smack	Stomach	Smack	Smack
Resecteded	Resected	Resected	Resurrected	Resected	Resected	Rusticated	Resecteded
Sutant	Sutent	Mutant	Mutant	Sutant	Sutant	mutant	dunant

Unsupervised data-driven spelling correction



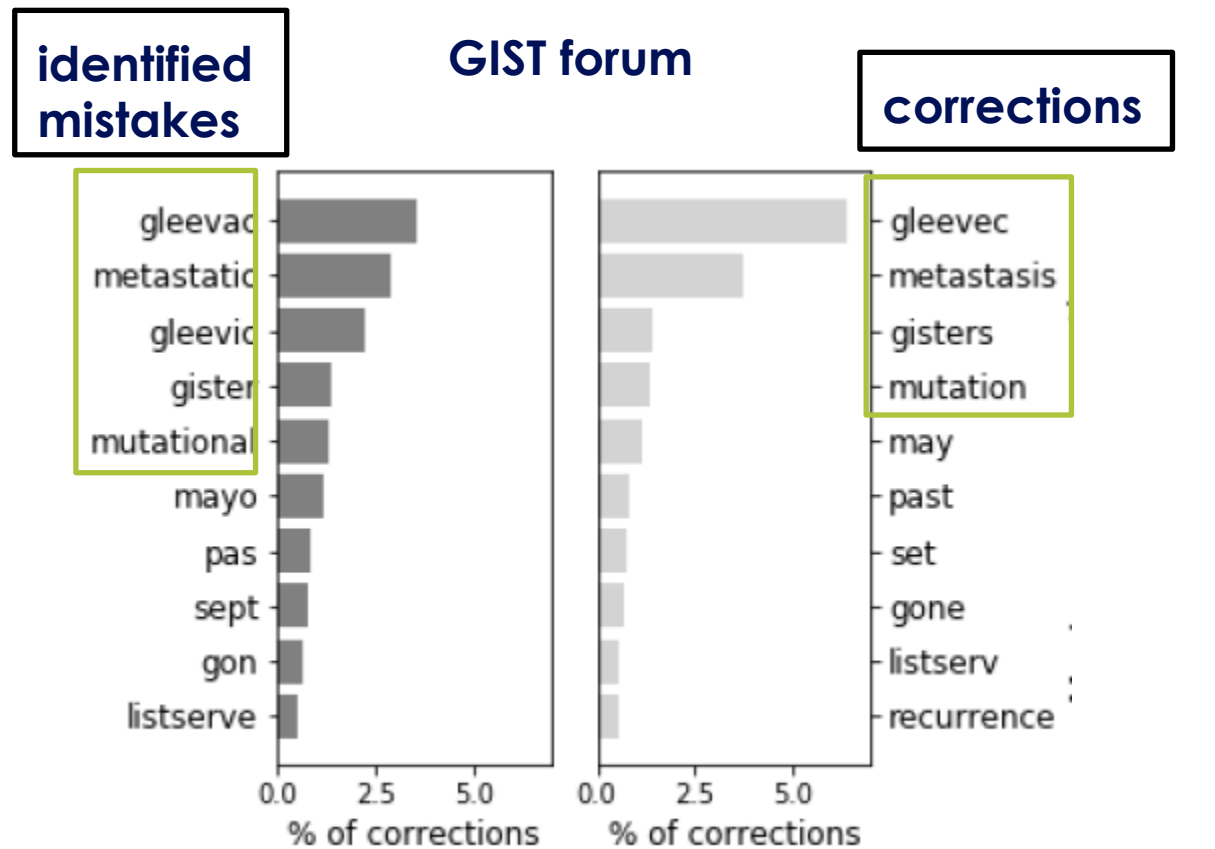
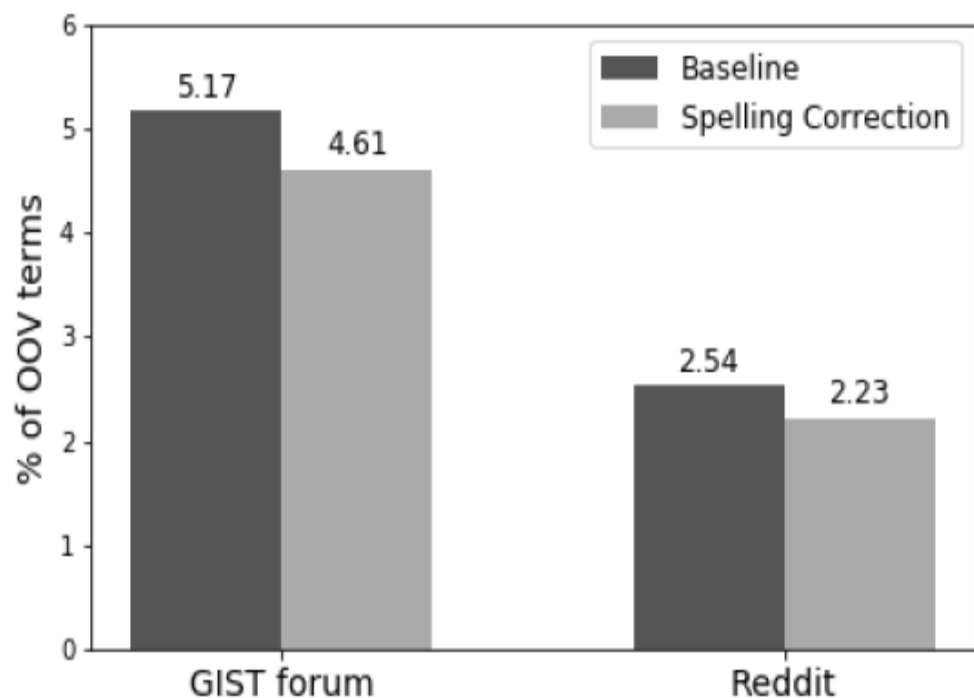
Unsupervised data-driven spelling correction



Spelling mistake detection

	F_{0.5}	F₁	Recall	Precision
CELEX	0.551	0.634	1.0	0.464
Decision process	0.888	0.871	0.844	0.900

Internal validation on a second cancer forum



health-related

Manual error analysis of 50 most frequent OOV

	GIST forum	Reddit
Spelling error	3	1
Real Word	11	21
Abbreviation	14	9
Slang	6	13
Name of person or hospital	14	2
Drug name	1	4
Not English	1	
	50	50

Classification task evaluation

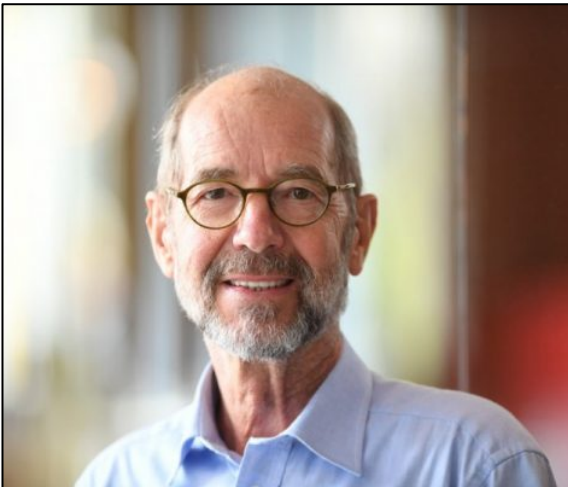
Data Set	Size	Change in F1	% of words corrected
Task 1 SMM4H	16,141	+0.006	1.1
Task 4 SMM4H Flu vaccine	6,738	+0.001	0.47
Flu Vaccination	3,798	+0.002	0.83
Twitter Health	2,598	+0.010*	0.64
Task 4 SMM4H Flu infection	1,034	+0.011	0.29
Zika Conspiracy Tweets	588	-0.011	1.1

Generic social media normalization

	F1	Precision	Recall
State of the art	0.836	0.880	0.796
Our method	0.522	0.646	0.577

Current work

1. More data
2. Improve generalization
3. Error analysis
4. Use of context of error to improve correction





**Universiteit
Leiden**
The Netherlands



a.r.dirkson@liacs.leidenuniv.nl



github.com/AnneDirkson



www.annedirkson.nl