



Viewpoint diversity in news recommendation: Theories, Models, and Tasks to support democracy

Myrthe Reuver
CLTL, VU Amsterdam

Who, what?



Myrthe Reuver, PhD candidate at CLTL at VU Amsterdam.

→ Before that: ResMA in Computational Linguistics @ Radboud Nijmegen, MSc in Cognitive Science & AI @ Tilburg University

→ Supervisors: Antske Fokkens (CLTL @ VU), Suzan Verberne (LIACS @ Leiden).

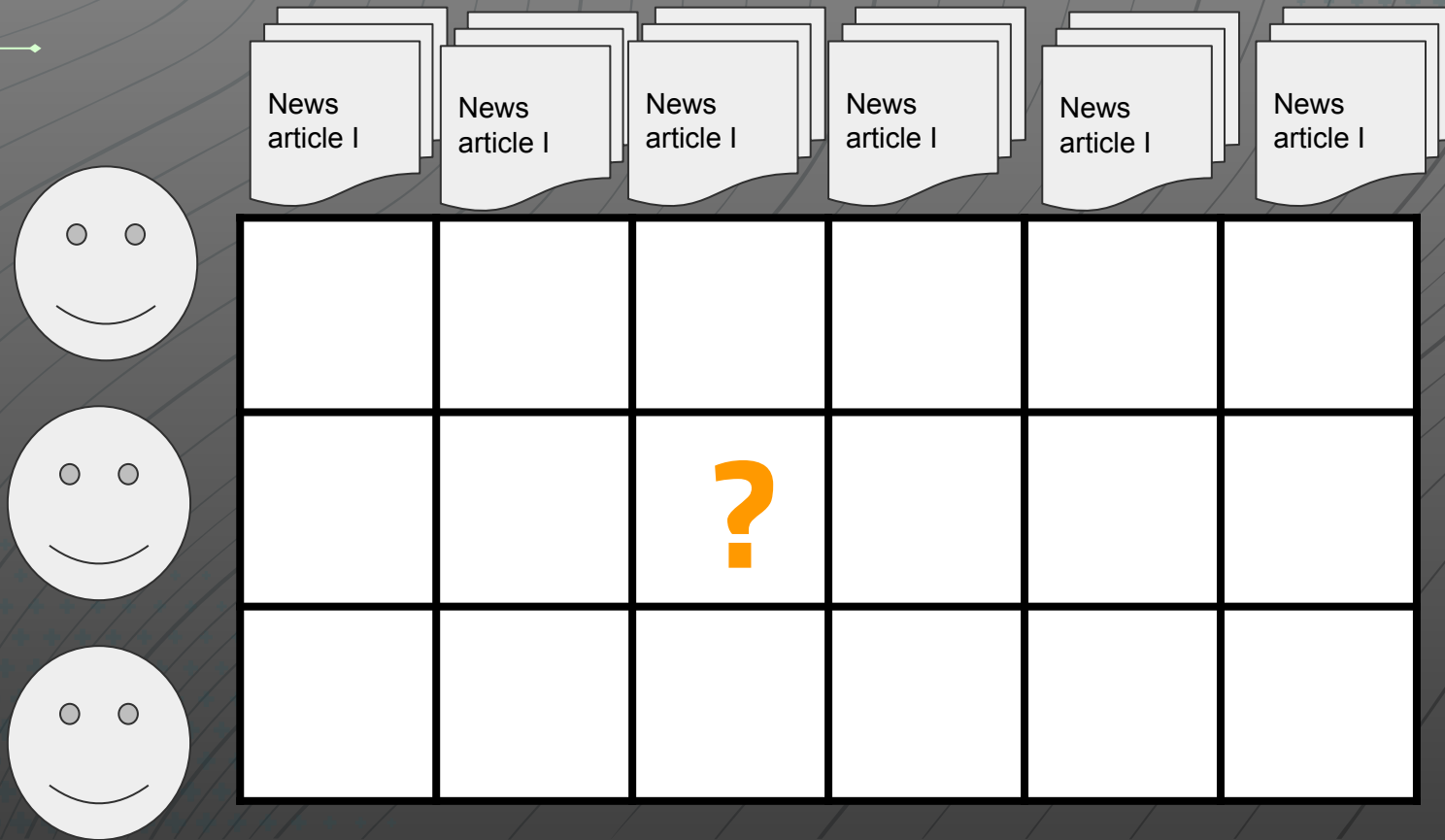


Computational linguist in an **interdisciplinary project** on diversity in news recommendation.

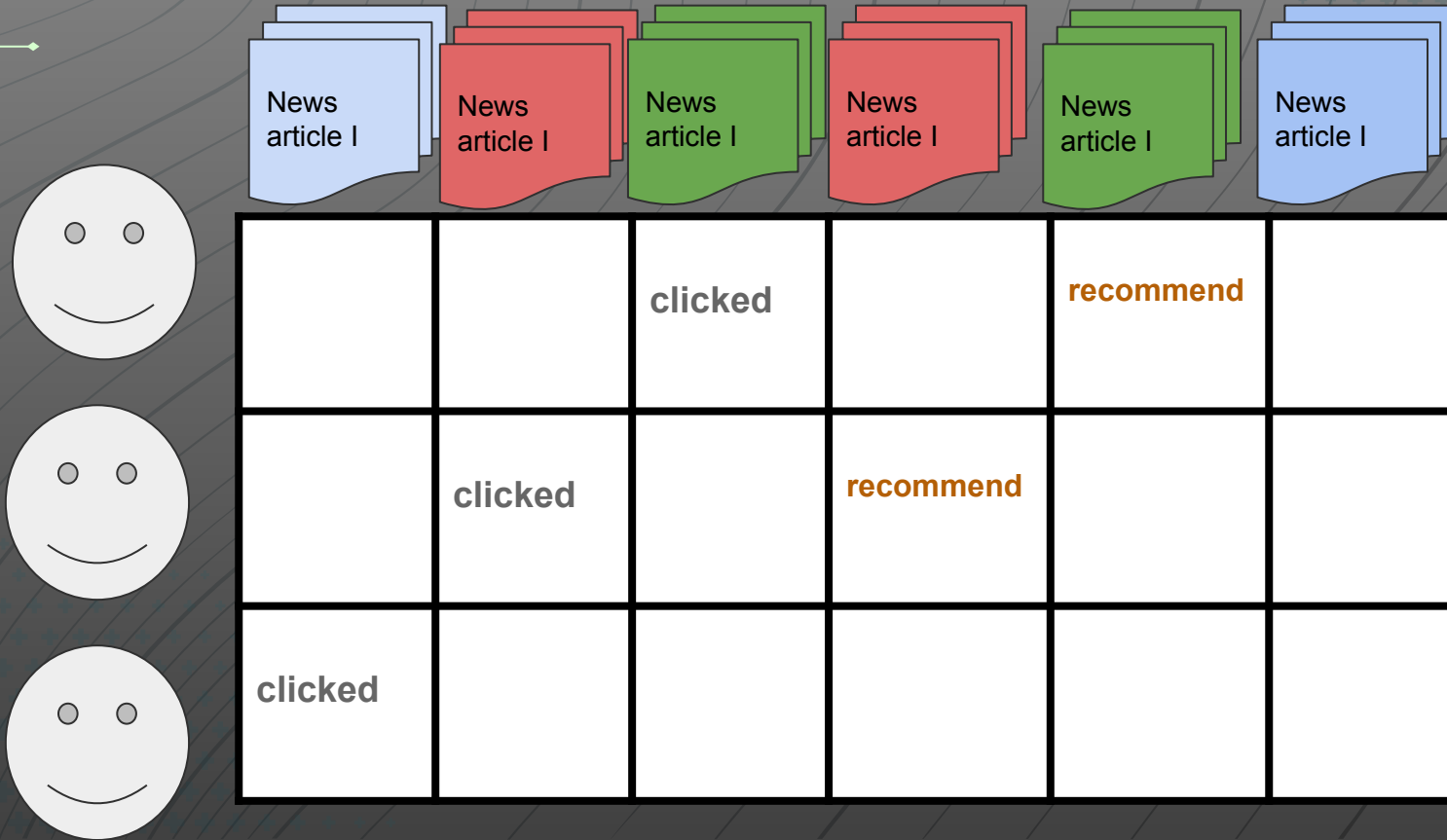
Social scientists, philosophers, and RecSys/computer scientists.

Part I: Context & Goal

What is a [news] recommender system?



What is a [news] recommender system?





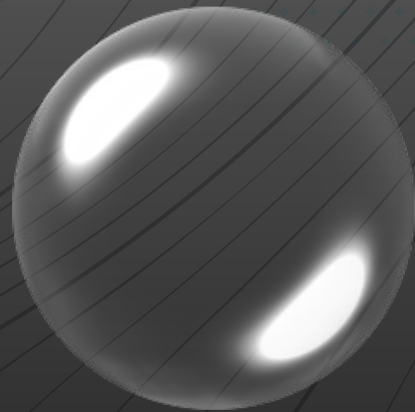
Optimizing in News Recommendation

Usually in RecSys: **click-accuracy** (as proxy for user interest).
Consequence: Showing users more of the same.

Could lead to 'filter bubbles' → potentially problematic for democracy.

But **why** ? And how can computational linguistics/NLP help?

My journey started with **asking experts (social scientists and theorists): What is democracy? What is the (complex) problem here?**



Models of Democracy



- Theoretical models that define a functioning democracy.
- Deliberative model: democracy requires **rational debate**, and actors encountering **a diverse set of viewpoints and ideas** on (societal) issues.
- Helberger (2019) connected this and other models to news recommendation. Supporting a deliberative model of democracy = recommenders promoting **rational rather than emotional content**, and **diverse viewpoints** on issues.

Vrijenhoek et al. (2021)'s Metrics: NLP operationalization needed

- Eval metrics & ideas on 'diversity' in RecSys implicitly still have user satisfaction as ultimate goal

Solution: Vrijenhoek's four new evaluation & optimization metrics:

- 1) **Fragmentation:** shared public sphere
- 2) **Representation:** diverse actors and opinions
- 3) **Alternative Voices:** non-mainstream opinions
- 4) **Calibration:** personalization
- 5) **Affect:** emotional content



Connection to theoretical models:

- Supporting **deliberative debate** requires low affect, low fragmentation
- **critical model** (where viewpoints clashing is considered healthy) requires high affect.

How to operationalize 'different viewpoints' with NLP?

Step 1: Conceptualization of “viewpoint diversity” (distinct from operationalization, following Jacobs & Wallach, 2020)

- Rather vague, e.g. “a wide range of perspectives on a given issue” (Griswold, 1998)

Step 2: operationalization in earlier work as relationship between topics, humans (users or authors), and other complex concepts like stance or sentiment:

- Ren et al (2016): tuple consisting of an entity, a topic related to this entity & sentiment towards topic
- Thonet (2016, 2017): stance: "the standpoint of one or several authors on a set of topics."

Reuver et. al. (2021): looking beyond individual tasks and datasets

NLP task-ification less useful for complex societal problems [Reuver et. al. 2021a]

- **fragmentation of literature and ideas:** for “viewpoints”: a large set of tasks is relevant, with each their own definitions, datasets, shared tasks, and benchmarks.
 - *stance detection, perspectives, Key Point Analysis*
- **definitions & artificial task setting:** Often, these tasks are **not connected to (social science) theory or real-world context**, but rather aimed at what is easy to measure or capture in text data.
- **evaluation** across different tasks is difficult: different metrics, benchmark datasets, etc.

Reuver et. al. (2021) “No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems” Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation.

Thinking beyond Task: the context

Theoretical models of democracy (such as the deliberative model) helps think of nuanced aspects going beyond task:

- Our ultimate goal is not e.g. detecting stances, but **supporting democracy**. That might mean some stances, viewpoints, or ideas (attacking democracy, or inherently violent stances) should **not** be recommended.
- We also need to not only detect stances, but **find diverse, different, or opposing** viewpoints and ideas.

Different tasks for “different viewpoint”

- Frames (Mulder et. al., 2021)
- Stances (Reuver et. al, 2021)
- Ideology (left-wing/right-wing, conservative/liberal, etc.)
- Perspective (content+stance, Fokkens et. al., 2017)
- Values of argument (e.g. “Economic Prosperity” vs “Mental Health”, Liscio et. al., 2021)
- Morals from Moral Foundation Theory (Kobbe et. al, 2020)
- Types of argument (“Moral”, “Civic”, “Economic”, Baden, 2017; Draws et. al. 2022)

.. Combination of the above → e.g. multidimensional, see Draws et. al. 2022 combining stance + argument type, or stance + moral foundation (Kobbe et. al., 2020), or stance + sentiment score specifically in news articles (Alam et. al., 2022)

Part II: Stances

What is [going on with] stance?

Stance detection, common definition: classification task (on texts, often tweets) with labels Pro, Con, Neutral towards an issue or topic

"Abortion is a sin, and should never be practiced."

Topic: **Abortion**, Stance: **Con**

Why stances?

- Built upon the linguistic phenomenon of **actors communicating their evaluation of targets**; placing **themselves and their targets on “dimensions in the sociocultural field”** (Du Bois, 2007).
- Directional (pro/con)
- Immediate **connected to (deliberative) debate and democratic decision making** (agree/disagree with laws, proposals, etc)



Reimers et. al. [2019]: cross-topic stance classification

Train: 7 topics, test: 8th topic

Fine-tuning BERT (base & large)

Findings:

- avg. F1 (10 seeds) = 0.633
- +0.20 over reference model (LSTM)
- Results are *“very promising and stress the feasibility of the task”* (Reimers et al. 2019, p. 575)



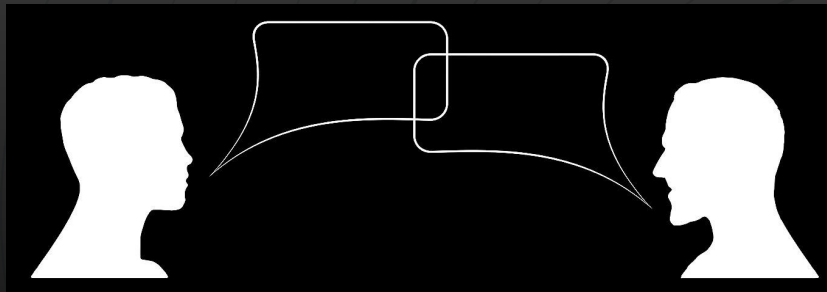
Marco Verch @ Flickr, Creative Commons 2.0.
<https://foto.wuestenigel.com/businessman-walking-from-a-to-b-point/>

Dataset: UKP Dataset (Stab et. al., 2018)

25,492 arguments on 8 topics, in 3 classes:

- For or against “the use, adoption, or idea” of the topic, or no argument
- 8 controversial debate topics from the internet: *abortion, cloning, death penalty, gun control, marijuana legalization, minimum wage, nuclear energy and school uniforms.*

Gerd Altmann, Pixabay licence.
<https://pixabay.com/illustrations/feedback-exchange-of-ideas-debate-2466829/>



Reproduction [Reuver et. al. 2021b]

- Important for science, and NLP specifically! (Fokkens et al., 2013; Belz et al., 2021).
- **Systematic reproduction:** 3 dimensions of reproduction (Cohen et. al., 2018): numeric values, findings, conclusions.
- Non-deterministic results of BERT:
 - **Standard deviation (SD) over seeds;**
 - value is reproduced if it falls **within 2 SDs.**

Reuver, M. E., Verberne, S., Vallejo, R. M., & Fokkens, A. (2021, November). Is Stance Detection Topic-Independent and Cross-topic Generalizable?-A Reproduction Study. In *Proceedings of the 8th Workshop on Argument Mining*,

Mean (stdv) over 10 seeds	F1
Reimers et. al. (2019)	
LSTM (baseline)	.424
BERT-base	.613 (-)
BERT-large	.633 (-)
Reuver et. al (2021)	
SVM+tf-idf (baseline)	.517
Reproduction BERT-base	.617 (.006)
Reproduction BERT-large (all)	.596 (.043)
BERT-large - 5 good seeds	.636 (.007)

Take-aways:

- BERT-large under-performs in 50% of seeds
- SVM+tf-idf model outperforms the LSTM reference model from the original study (F1 of .517 > .424)

Cohen et. al. (2018)'s 3 dimensions of reproducibility:

1. (numeric) values:

✓ Within 2 standard deviations (BERT-large = large SD)

2. findings (relationship between variables, e.g. model & result):

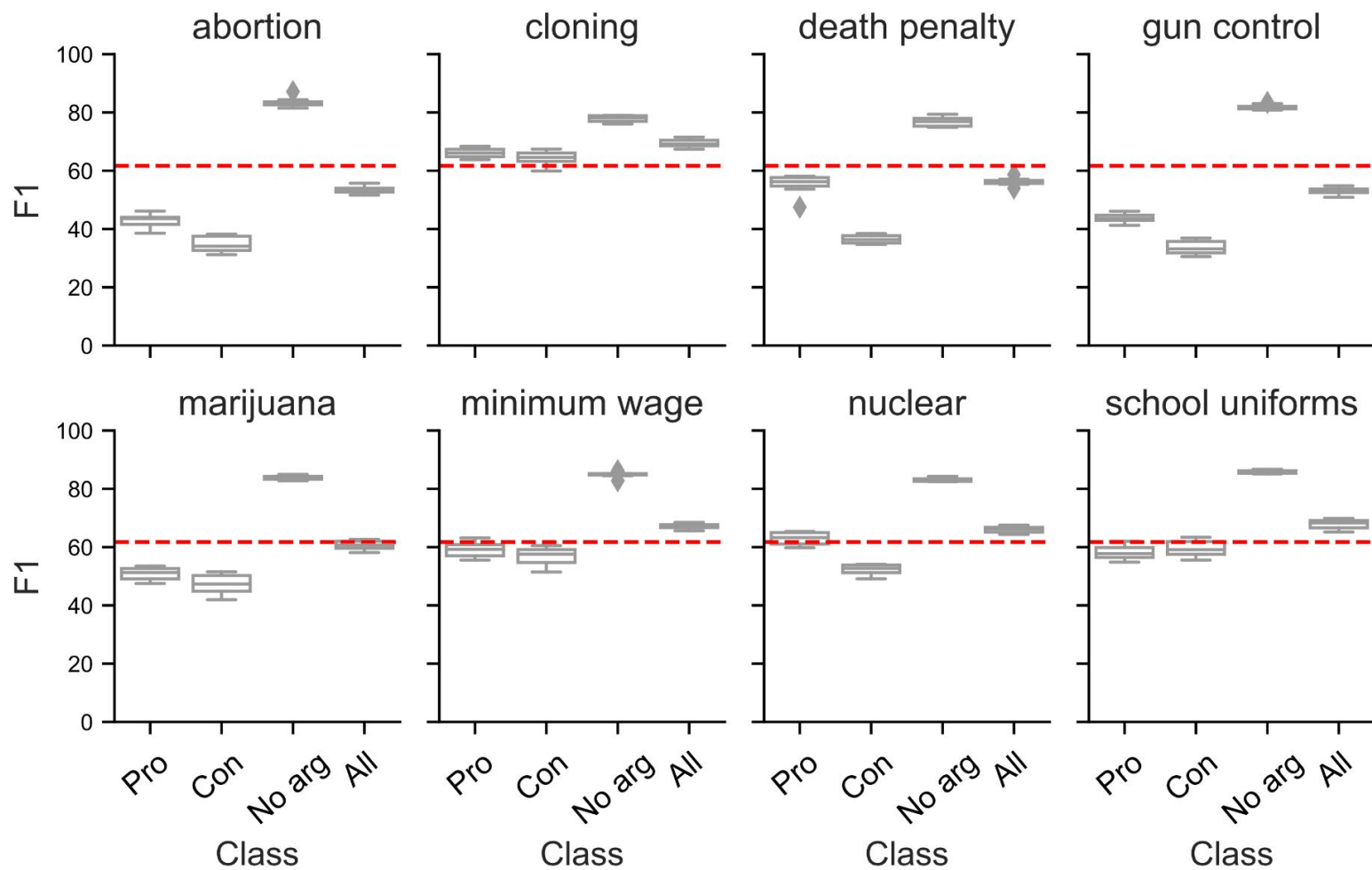
✓ baseline < BERT-base < BERT-large,

✗ .20 improvement over non-BERT model (LSTM) is not the same with our reference model (SVM+tf-idf);

3. conclusion(s):



How feasible is cross-topic? Let's investigate some more, especially on topics.



--- BERT-base F1 (mean)

What does this mean?

Successful reproduction of cross-topic stance classification (Reimers et. al., 2019)
on most dimensions, **but:**

- random seed does matter for BERT-large;
- reference model/baseline matters.

Topic matters! Stance not as topic-independent as seems with one averaged F1 metric reported.

- See also: Thorn Jakobsen et. al. (2021)

A **class/topic** interaction effect in stance



OpenClipArt, Public domain

Time to (re)investigate topic similarity, (socio)-cultural and lexical topic specificity? When can we cross to new topics?

Part III: Current Roadmap

Problems with current Pro/Con stance operationalization

Limited cross-topic generalization

→ less useful because new topics appear in the news (see: Reuver et. al., 2021).

Viewpoints are complex, changing, and contextual (Joseph et. al., 2021).

Risk of oversimplification with Pro/Con stances, e.g. context and underlying arguments are missing (Scott et. al., 2021)

Focus on United States context & Twitter in datasets

Current Road I: Task Reformalization

Recent **task reformalization** into “Same Side Stance” of pairs (Stein et. al., 2020)

- claimed to reduce model’s leaning on topic-specific pro- and con-vocabulary;
- Allows for identifying a *different* stance;
- Recommendation and (user) exploration of “difference”;
- Allows for research direction of “same stance, different frame” etc.

Current Road II: Dataset Creation

What do we need?

- a large dataset
- in Dutch
- Relevant topics, labels, and annotators for the Dutch political context
- Focused on the news
- Carefully and fairly annotated



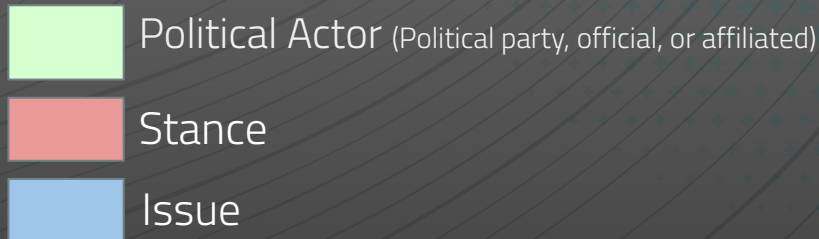
Dutch stance dataset: “beyond gun control”

With: Kasper Welbers, Wouter van Atteveldt, Antske Fokkens, Mariken van der Velden and Felicia Locherbach → most VU Social Science department

Dutch stance dataset on sentences from news texts on the 2020 Dutch elections

Stances in the news on four Issues:
Immigration, Climate measures, taxes, and European Union membership.

Aim: diversity of stances, actors, issues in news recommendation



VVD komt in
opstand tegen
stikstofplannen
eigen minister

TASK 1: (political) actors

Pre-annotated with actor
dictionary + potential
reference words

? Go to next unit

Vrijwel niets staat een nieuw kabinet nog in de weg. De coalitiefracties steunen het regeerakkoord, dat vandaag wordt gepresenteerd. Vanaf morgen kan premier Rutte zijn vierde regeringsploeg gaan smeden. Laurens Kok Hanneke Keultjes Den Haag Gisteren lekten er opnieuw voornemens uit die VVD, D66, CDA en ChristenUnie met elkaar hebben afgesproken. Volgens bronnen zijn de partijen van plan het minimumloon te verhogen en willen ze de lasten met '3 tot 4 miljard euro' verlichten. De details zijn nog niet duidelijk. Die worden in de loop van de dag bekendgemaakt, tijdens de presentatie in het Tweede Kamergebouw. Eerst moeten vanochtend nog wat vuiltjes worden weggewerkt. Alle vier de fracties gingen gisteren akkoord met het onderhandelingsresultaat dat maandagavond werd bereikt. Toch zijn er nog een paar punten die besproken moeten worden, al sluiten betrokkenen uit dat het nu alsnog mis kan gaan. Kamerleden van de vier partijen kwamen gisteren ieder in eigen kring op een geheime locatie bijeen om het eindresultaat te bespreken. Pas nadat de deuren dicht waren gegaan, werd de pers ingelicht. Bij het CDA kringelde als

Add labels (referent) to grey candidate items

TASK 2: issue detection

Validated dictionary
Dictionary expansion with
FastText embeddings

28 / 217

De verlaging van de btw op groente en fruit, de afschaffing van het leenstelsel voor studenten en een hardere aanpak van illegale migranten: het zijn plannen uit het regeerakkoord die op brede steun kunnen rekenen bij de Nederlandse kiezer, zo blijkt uit een representatieve peiling van het actualiteitenprogramma EenVandaag.

gaat deze zin over belastingheffing?

← nee ↑ ik weet niet ja →


TASK 3: stance

53 / 217

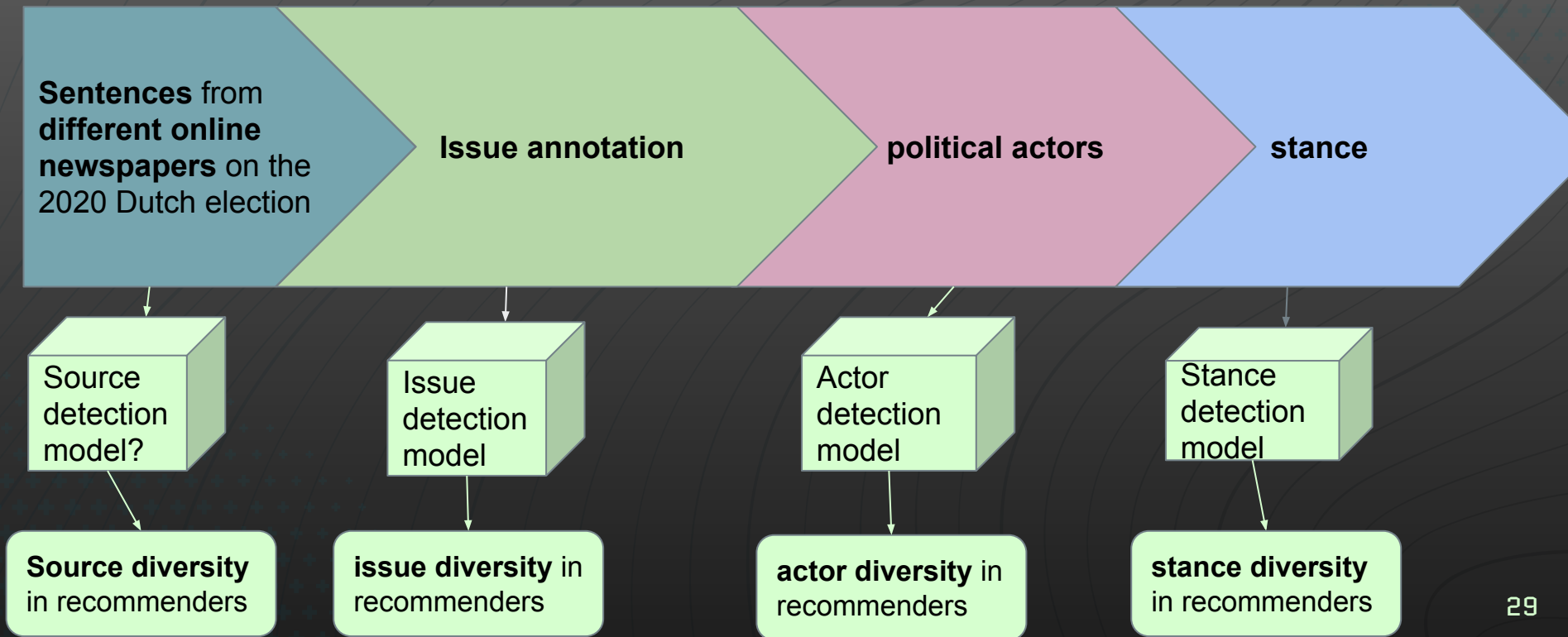
Zo vinden VVD, D66, CDA en ChristenUnie dat er rekeningrijden moet komen, waardoor automobilisten betalen naar gebruik.

Welk standpunt over belastingheffing drukt deze zin uit?

← voor ↑ geen standpunt tegen →



“Dutch Election Stance Dataset” pipeline → stacked annotation tasks



Wanna swipe?

Annotating Dutch Stances
with AnnoTinder
(CCSAnnotator)



Conclusion / Summary

My current research: operationalization of viewpoint

- specifically addressing gaps in the current NLP (stance) literature for the “diverse news recommendation” use case.



- These gaps are:
 - A lack of useful task definitions
 - A lack of useful data for the Dutch context and our use case
 - Lack of cross-topic generalizable operationalizations
- Current roadmap:
 - exploring alternative task definitions (“different stance”)
 - Carefully designing a Dutch dataset
 - Explicitly testing cross-topic generalizability

Questions or ideas?



@myrthereuver



myrthe.reuver[at]vu.nl

References

Reuver, M. & Mattis, N. "Implementing Evaluation Metrics Based on Theories of Democracy in News Comment Recommendation (Hackathon Report)" Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation (co-located at EACL 2021, online). Association of Computational Linguistics, pp. 134–139.

Reuver, M., Fokkens, A. & Verberne, S. (2021). No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems. In: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation (co-located at EACL 2021, online). Association of Computational Linguistics, p. 45–55.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2020) "Data and its (dis) contents: A survey of dataset development and use in machine learning research", NeurIPS 2020 Workshop on Machine Learning Retrospectives, Surveys, & Meta-Analyses.

Vrijenhoek, S., Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a mission: assessing diversity in news recommendations. In the Proceedings of the SIGIR Conference on Human Information Interaction and Retrieval (CHIIR), 173–183.
<https://doi.org/10.1145/3406522.3446019>.

Mean (stdv) over 10 seeds	F1	P pro	P con	R pro	R con
Reimers et. al. (2019)					
LSTM (baseline)	.424	.267	.389	.281	.403
BERT-base	.613 (-)	.505 (-)	.531 (-)	.470 (-)	.576 (-)
BERT-large	.633 (-)	.554 (-)	.584 (-)	.505 (-)	.560 (-)
Reuver et. al (2021)					
SVM+tf-idf (baseline)	.517	.418	.460	.414	.423
Reproduction BERT-base	.617 (.006)	.519 (.011)	.538 (.007)	.464 (.029)	.581 (.019)
Reproduction BERT-large (all)	.596 (.043)	.483 (.057)	.527 (.057)	.464 (.058)	.516 (.063)
BERT-large - 5 good seeds	.636 (.007)	.532 (.014)	.578 (.016)	.515 (.016)	.567 (.022)

Results: further details

Model	UKP Dataset				
	F1	P pro	P con	R pro	R con
mean (stdev) 10 seeds					
Reimers et al. (2019) biclstm+BERT	.424	.267	.389	.281	.403
Reimers et al. (2019) BERT base	.613 (-)	.505 (-)	.531 (-)	.470 (-)	.576 (-)
Reimers et al. (2019) BERT large	.633 (-)	.554 (-)	.584 (-)	.505 (-)	.560 (-)
SVM+tf-idf	.517	.418	.460	.414	.423
Reproduction BERT-base	.617 (.006)	.519 (.011)	.538 (.007)	.464 (.029)	.581 (.019)
Repr. BERT-large - all seeds	.596 (.043)	.483 (.057)	.527 (.057)	.464 (.058)	.516 (.063)
Repr. BERT-large - 5 evenly performing seeds	.636 (.007)	.532 (.014)	.578 (.016)	.515 (.016)	.567 (.022)

- BERT-large under-performs in 50% of seeds
- SVM+tf-idf model outperforms the LSTM reference model from the original study (.517 > .424)