Universität Konstanz

Frame Detection in German Political Discourses:

How far can we go without large-scale manual corpus annotation?

Qi Yu, Anselm Fliethmann

1st Workshop on Computational Linguistics for Political Text Analysis
September 6, 2021, Düsseldorf/Online

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation
EXC-2035/1 – 390681379

Cluster of Excellence
The Politics of Inequality

Background: *Frame* in Political Arena

- Entman (1993): To frame is to **select** some aspects of a perceived reality and make them more **salient** in a communicating text

Examples: when talking about refugees and migration…

Unaccompanied, allegedly underage foreigners (UMA) abuse the immigration and asylum law.
(AfD, election program 2017)

→ **legal frame**: **refugees as abuser of asylum law**

The demand (of skilled workers) will continue to grow in the next few years due to our economic development and the declining number of young people entering the labor market.
(CDU, election program 2017)

→ **economic frame**: **migration (of skilled workers) as a gain of the domestic economy**

Earlier NLP Studies on Automated Frame Detection

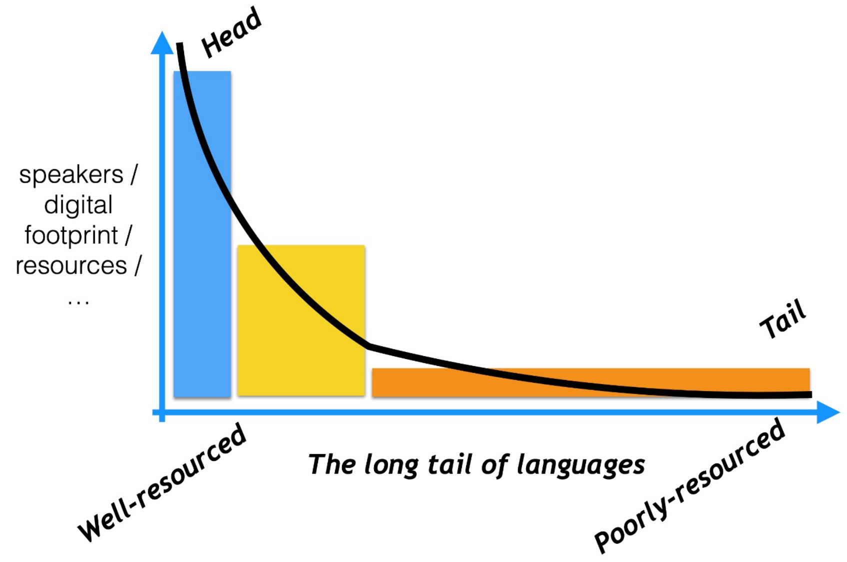
Majority of the earlier studies on automated frame detection:

- use **supervised methods** → relying heavily on manual annotation
  Media Frame Corpus (Card et al. 2015) is the most utilized dataset
- focus on **English** → other languages remain greatly neglected

(Baumer et al. 2015, Naderi & Hirst 2016, Ji & Smith 2017, Johnson et al. 2017, Khanehzar et al. 2019, Liu et al. 2019, Cabot et al. 2020, Mendelsohn et al. 2021)

Restrictions:

- Labeled data is scarce



(Source: Talk by Barbara Plank, LxMLS 2021)

Research Question

Automated frame detection: How far can we go when we don’t have large-scale labeled data?

- Unsupervised methods?
- Other knowledge-based methods?

Roadmap

Roadmap

1. **Data**

2. **Experiments**

   2.1 An unsupervised approach: LDA-based topic modelling

   2.2 A knowledge-based approach: Combining word embeddings with framing keywords

3. **Conclusion and Outlook**

Data

Corpus of German-language news articles on the “European Refugee Crisis” between 2014-2018

newspaper	category	#article	#token
BILD	tabloid; right-wing	12,287	3,554,105
FAZ	broadsheet; right-wing	6,832	3,526,323
SZ	broadsheet; left-wing	4,770	1,893,868

frame differences?

Selection criterium of the articles: ≥1 match with `refugee-keywords` (including inflected forms)
{*Flüchtling, Geflüchtete, Migrant, Asylant, Asylwerber, Asylbewerber, Asylsuchende*}

Corpus cleaning: Omitted articles with frequency of `refugee-keywords` ≤ 0.01

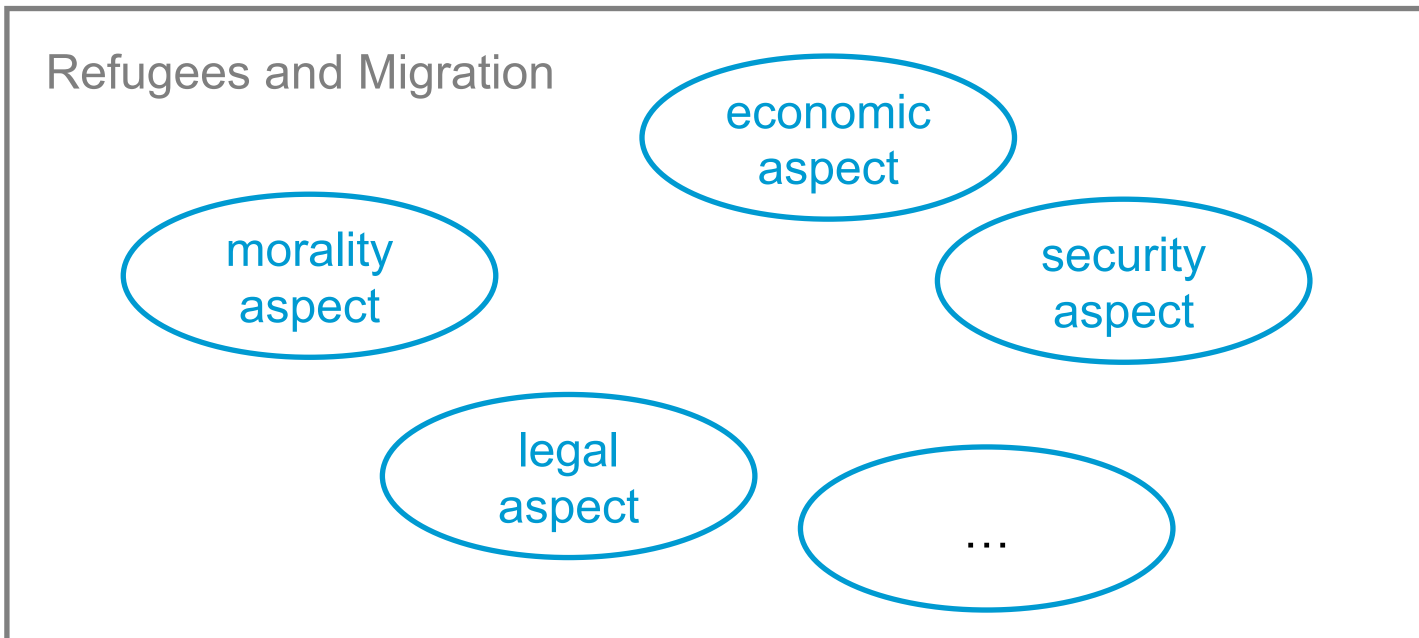
Data availability:
- The corpus cannot be made publicly available due to copyright regulations of the publishers
- However, we release the lexical resource resulting from this paper (details follow shortly)

Roadmap

Experiment 1: Detecting Frames Using LDA-Based Topic Modelling

Rationale:

- Detecting frames of an issue resembles detecting sub-topics within the issue



Experiment 1: Detecting Frames Using LDA-Based Topic Modelling

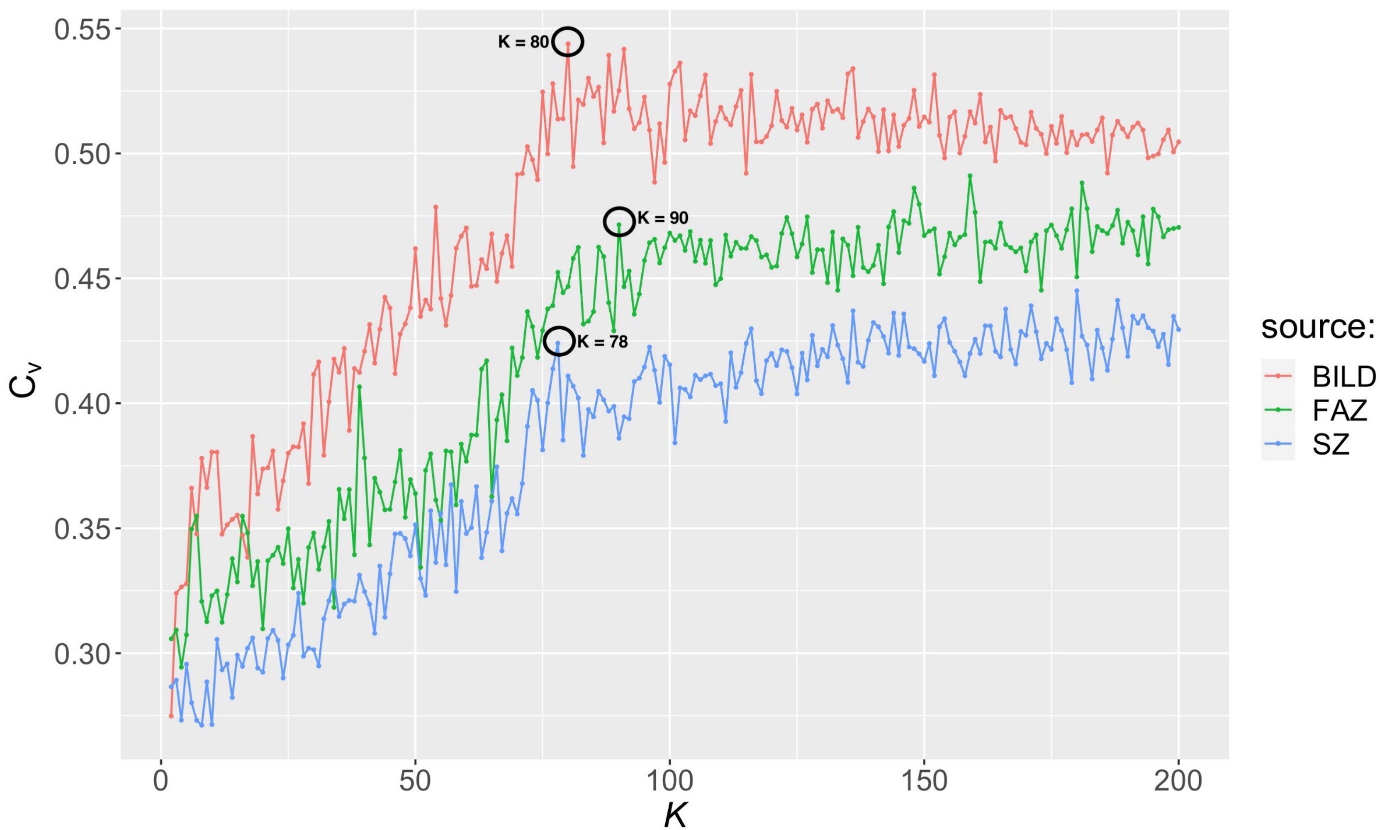
Training:

- One LDA-based model per newspaper using Python library *Gensim* (Řehůřek & Sojka 2010)
- Exclusion of non-informative N-grams ($N \in \{1, 2, 3\}$):
  - N-grams with a document frequency higher than 0.15 and n-grams with a term frequency lower than 5 were excluded
- The `refugee-keywords` were masked in order to eliminate their interference in the topic modelling algorithm, since they appear in all articles

Experiment 1: Detecting Frames Using LDA-Based Topic Modelling

Searching for optimal topic number *K* using coherence score C_v (Röder et al. 2015)**:**

$C_v \in [0\ ,1]$:
The closer the value is to 1, the more coherent the topics are



Experiment 1: Detecting Frames Using LDA-Based Topic Modelling

Result and discussion:

- Low coherence score despite optimized topic numbers (BILD: $C_v = 0.544$, FAZ: $C_v = 0.471$, SZ: $C_v = 0.424$)
- High degree of overlap between the resulting topics:

FAZ	**Topic 77:** Griechenland (Greece), EU (EU), mehr (more), Million_Euro (million Euro), Land (country), Band (band), Europa (Europe), Türkei (Turkey), Integration (integration), Kreis (district) **Topic 80:** Türkei (Turkey), EU (EU), Griechenland (Greece), Ankara (Ankara), Europa (Europe), Brüssel (Brussels), türkisch (Turkish), EU_Staat (EU country), Flüchtlingskrise (refugee crisis), Erdoğan (Erdoğan)

- The high number of K considerably complicates the human interpretation of the overall topic differences between the newspapers

Experiment 1: Detecting Frames Using LDA-Based Topic Modelling

Possible explanation for the poor performance of topic modelling:

- High degree of vocabulary homogeneity among the articles in the dataset

  e.g., *Syrien* (‘Syria’), *Land* (‘country’), *EU* (’EU’):

  highly relevant words for both refugee allocation policies (policy frames) and security issues on the migration route (security frames)

  - Many items show a “stop word”-alike behavior regarding specific issues: ubiquitous, high frequency

    → They may confuse the topic modelling algorithm
  - But unlike stop word, they bear highly relevant semantic information

    → Eliminating them would lead to a huge loss of information

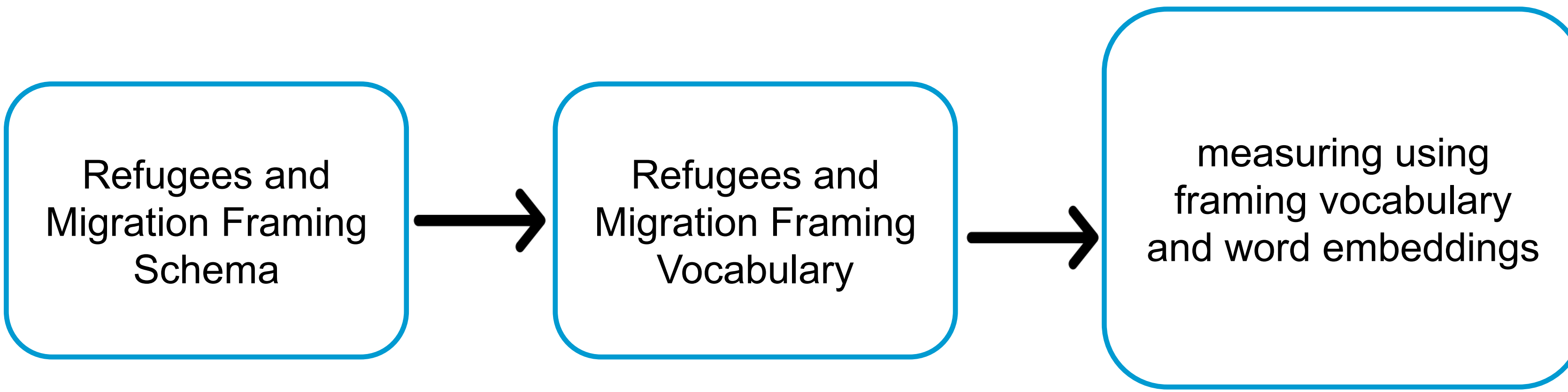
Roadmap

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Rationale:

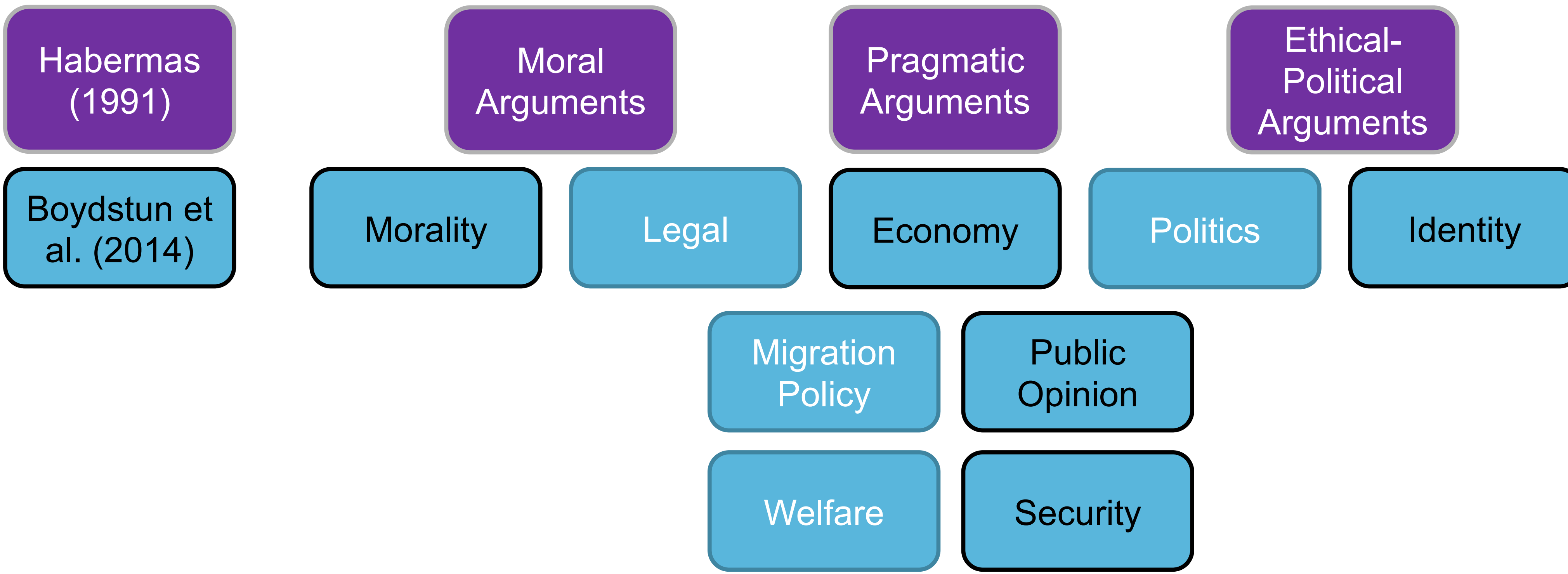
- Earlier studies observed and empirically verified that framing in news is to a large extent a keyword-driven phenomenon (Johnson et al. 2017, Field et al. 2018, Akyürek et al. 2020)

Method overview:



Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Developing *Refugees and Migration Framing Schema*:



Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Creating *Refugees and Migration Framing Vocabulary*:

source 1

Seed words by domain experts
+
GermaNet

source 2

DebateNet-mig15 corpus
(Lapesa et al. 2020)

- 5 domain experts conducted an explorative reading of a small part of articles from the corpus
- The domain experts listed up highly relevant words for each frame category in the schema
- The resulted seed word lists were then expanded by synonyms found with GermaNet

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Creating *Refugees and Migration Framing Vocabulary*:

source 1

Seed words by domain experts
+
GermaNet

source 2

DebateNet-mig15 corpus
(Lapesa et al. 2020)

- **DEbateNet-mig15:** 3,442 text passages from the German newspaper *Die Tageszeitung* (TAZ) of 2015 that are annotated as *claims* (= statements made by political actors)

- For each of the 8 high-level category *C* in the annotation schema of DEbateNet-mig15, top 200 words *w* with highest pointwise mutual information (PMI) to *C* were extracted:

$$PMI(C,w) \equiv log\frac{P(C,w)}{P(C)P(w)} = log\frac{P(w|C)}{P(w)}$$

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Creating *Refugees and Migration Framing Vocabulary*:

source 1	source 2
Seed words by domain experts + *GermaNet*	*DebateNet-mig15* corpus (Lapesa et al. 2020)

Why is DEbateNet-mig15 justified as source for creating framing vocabulary, though the annotated texts there are claims instead of frames?

- The claims are categorized based on the aspect(s) they emphasize
  → the categorization resembles frames to a large extent
- The data is in German language and arises from the same political issue as the one under investigation in our study (cf. Media Frames Corpus, Gun Violence Frame Corpus etc.)

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Creating *Refugees and Migration Framing Vocabulary*:

source 1

Seed words by domain experts
+
GermaNet

source 2

DebateNet-mig15 corpus
(Lapesa et al. 2020)

Manual evaluation of the resulting word lists:
Exclude items that are too general for detecting specific frame categories
(e.g., *Einwanderung* ‘migration’)

Refugees and Migration Framing Vocabulary
https://github.com/qi-yu/refugees-and-migration-framing-vocabulary

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

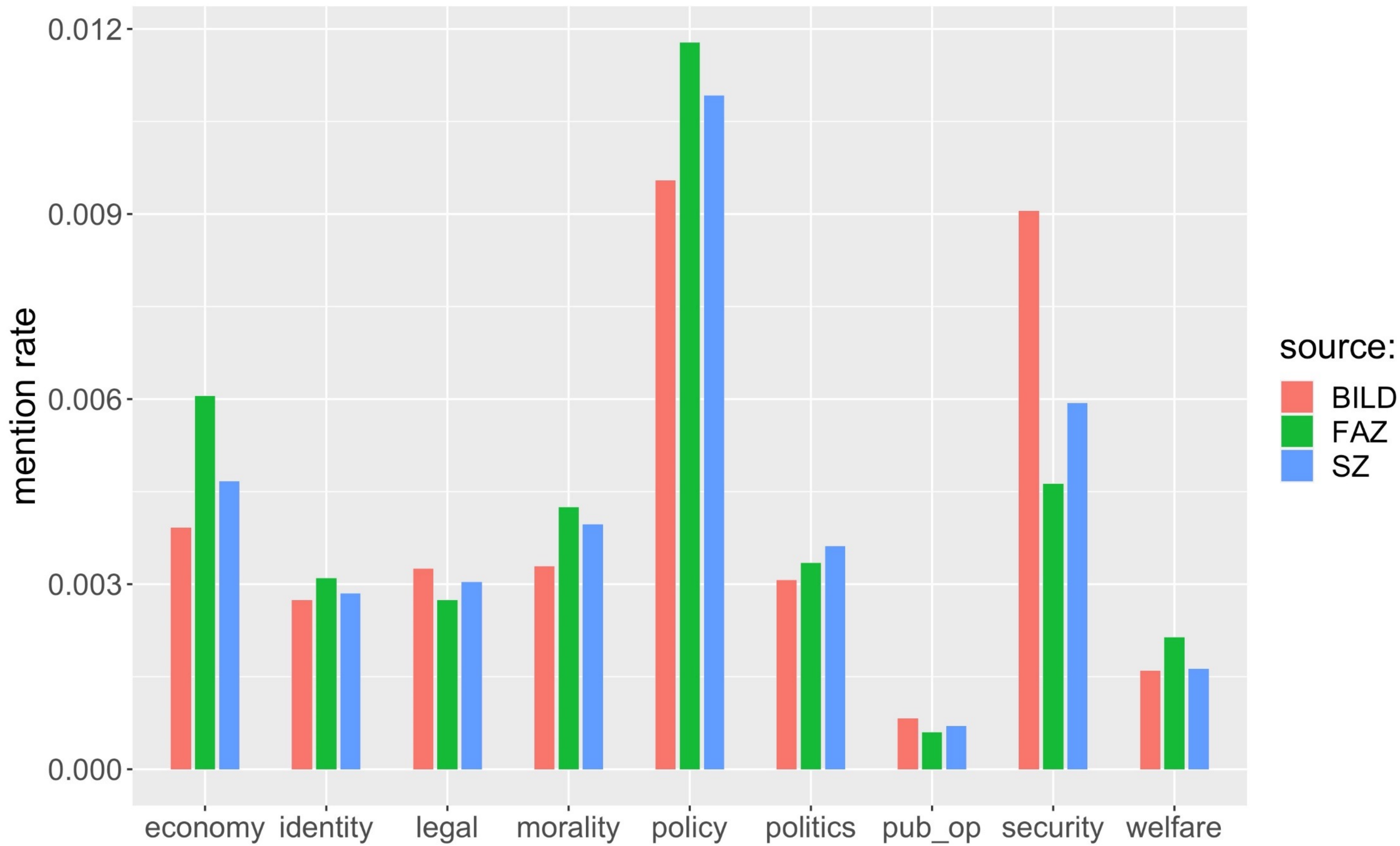
Measure 1: Mention rate of frames

- A frame category is considered to be *mentioned* if at least one of the keywords in the vocabulary $\{w_1, w_2, ..., w_k\}$ of this category is mentioned
- *Mention rate* of frame category *F* in newspaper *N*:

$$mention\_rate_N(F) = \frac{\sum_{i=1}^{k} count_N(w_i)}{count_N(allwords)}$$

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Measure 1: Mention rate of frames



Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Measure 2: Semantic similarity using *word2vec*

- Motivation: to investigate the more subtle attitudinal differences associated to a certain frame in greater depth

- A 300-dimensional *word2vec* model is trained for each newspaper

- For each frame-specific vocabulary list $\{w_1, w_2, ..., w_k\}$, we rank items in the list by their cosine similarity to the `refugee_centroid`, which is computed as the average embedding of all `refugee-keywords`:

  {*Flüchtling, Geflüchtete, Migrant, Asylant, Asylwerber, Asylbewerber, Asylsuchende*}

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Measure 2, Example 1: Semantic similarity using *word2vec*

Frame	BILD	FAZ	SZ
Security	Minderjährige (underage persons) Delikt (offense) Straftäter (perpetrator) Dschihad (Jihad) Gewaltkriminalität (violent crime) Islamist (Islamist) Bürgerkrieg (civil war) Tatverdächtiger (suspect) Schiffsunglück (shipwreck) inhaftieren (imprison)	Minderjährige (underage persons) illegal (illegal) Bürgerkrieg (civil war) Küstenwache (coast guard) Straftat (crime) Kriminalitätsrate (crime rate) Schiffsunglück (shipwreck) Schlepper (human trafficker) Gefängnis (prison) Gefängnisstrafe (imprisonment)	Rettungsmission (rescue mission) Minderjährige (underage persons) Krieg (war) Bürgerkrieg (civil war) illegal (illegal) minderjährig (underage) Schlepper (human trafficker) Straftat (crime) Schutzstatus (protection status) Schiffsunglück (shipwreck)

Roadmap

1. Data
2. Experiments
   2.1 An unsupervised approach: LDA-based topic modelling
   2.2 A knowledge-based approach: combining word embeddings with framing keywords
3. **Conclusion and Outlook**

Conclusion and Outlook

- We showed that topic modelling is insufficient in detecting frames in a dataset with highly homogeneous vocabulary
- We proposed a novel framing schema, the *Refugees and Migration Framing Schema*, which is specifically designed to analyze frames in the context of refugees and migration
- Based on the framing schema, we created the lexical resource *Refugees and Migration Framing Vocabulary*
- We showed that the combination of word2vec and *Refugees and Migration Framing Vocabulary* yielded much more human-explainable and insightful results that topic modelling
- The quality of the handcrafted vocabulary has great impact on the quality of the results
  → Future work: further improve the quality of our vocabulary lists by expanding it with more sophisticated keyword-mining techniques (e.g., Jin et al. 2021)

Universität Konstanz

Thank you!
Questions, comments, ideas?

qi.yu@uni-konstanz.de
anselm.fliethmann@uni-konstanz.de

References

Afra Feyza **Akyürek**, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624.

Eric **Baumer**, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482.

Amber E. **Boydstun**, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues. https://homes.cs.washington.edu/~nasmith/papers/boydstun+card+gross+resnik+smith.apsa14.pdf, last accessed 31 August 2021.

Pere-Lluís Huguet **Cabot**, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4479–4488.

Robert M. **Entman**. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

Anjalie **Field**, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.

Jürgen **Habermas**. 1991. *Erläuterungen zur Diskursethik*. Suhrkamp, Frankfurt am Main.

Yangfeng **Ji** and Noah Smith. 2017. Neural discourse structure for text categorization. *arXiv preprint arXiv:1702.01829*.

Kristen **Johnson**, Di Jin, and Dan Goldwasser. 2017. Modeling of political discourse framing on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 556–559.

Shima **Khanehzar**, Andrew Turpin, and Gosia Mikolajczak. 2019. Modeling political framing across policy issues and contexts. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66.

References

Gabriella **Lapesa**, André Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, Jonas Kuhn, and Sebastian Padó. 2020. DEbateNet-mig15: Tracing the 2015 immigration debate in Germany over time. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 919–927.

Siyi **Liu**, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514.

Julia **Mendelsohn**, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. *arXiv preprint arXiv:2104.06443*.

Nona **Naderi** and Graeme Hirst. 2016. Classifying frames at the sentence level in news articles. In *Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation*, pages 1–9.

Radim **Řehůřek** and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Michael **Röder**, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Universität Konstanz

APPENDIX

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Measure 2, Example 2: Semantic similarity using *word2vec*

Frame	BILD	FAZ	SZ
Economy	Kredit (credit) Arbeitsvertrag (working contract) Bildungsniveau (level of education) Integrationskurs (integration course) Anstellung (employment) Wirtschaftsflüchtling (economic refugee) Studium (academic studies) Deutschkurs (German course) Berufsausbildung (vocational training) Hilfsmittel (aid)	Wirtschaftsflüchtling (economic refugee) Fachkraft (skilled employee) Studium (academic studies) Schulausbildung (school education) Arbeitsstelle (workplace) Arbeitsvertrag (working contract) Berufsausbildung (vocational training) erwerbslos (unemployed) arbeitslos (unemployed) Fachkräfteeinwanderung (skilled employee migration)	Kosten (costs) Wohnung (lodging) Berufsqualifikation (vocational qualification) Ausbildung (training) erwerbstätig (employed) Arbeitslosenquote (unemployment rate) zahlen (pay) Bildungsniveau (level of education) Bleibeperspektive (prospect of staying) qualifiziert (qualified)

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Measure 2, Example 3: Semantic similarity using *word2vec*

Frame	BILD	FAZ	SZ
Policy	Visum (visa) Aufenthaltserlaubnis (residence permit) Ausreise (departure) Integrationskurs (integration course) Sozialhilfe (social care) einstufen (classify) Studium (academic studies) Abschiebung (deportation) Deutschkurs (German course) Sozialleistung (social benefit)	Aufenthaltserlaubnis (residence permit) Visum (visa) Asylverfahren (asylum procedure) Abschiebung (deportation) Balkanroute (Balkan route) Ausreise (departure) Studium (academic studies) Herkunftsland (country of origin) Schulausbildung (school education) Aufenthaltsrecht (right of residence)	Rettungsmission (rescue mission) Abschiebung (deportation) Asylverfahren (asylum procedure) Herkunftsland (country of origin) Wohnung (lodging) Sozialleistung (social benefit) Ausreise (departure) Aufenthaltserlaubnis (residence permit) Balkanroute (Balkan route) Bleibeperspektive (prospect of staying)

Experiment 2: Detecting Frames Using Word Embeddings and Framing Vocabulary

Measure 2, Example 4: Semantic similarity using *word2vec*

Frame	BILD	FAZ	SZ
Morality	Integrationskurs (integration course) Wirtschaftsflüchtling (economic refugee) Hartz IV (Hartz IV) Hilfsmittel (aid) Flüchtlingsversorgung (provisioning for refugees) Arbeitslosengeld (unemployment benefit) menschenwürdig (humane) Wirtschaftsmigrant (economic migrant) Armut (poverty) Ungleichheit (inequality)	Wirtschaftsflüchtling (economic refugee) Fachkräfteeinwanderung (skilled employee migration) Wirtschaftskrise (economic crisis) Integrationskurs (integration course) Quote (quota) Armut (poverty) Wirtschaftsmigrant (economic migrant) Punktesystem (point system) Hartz IV (Hartz IV) menschenwürdig (humane)	Rettungsmission (rescue mission) Flüchtlingsversorgung (provisioning for refugees) Quote (quota) Armut (poverty) Seenotrettungsprogramm (sea rescue program) Leistung (merit) Kontingent (quota) gemeinnützig (non-profit) Wirtschaftsflüchtling (economic refugee) Versorgung (provisioning)