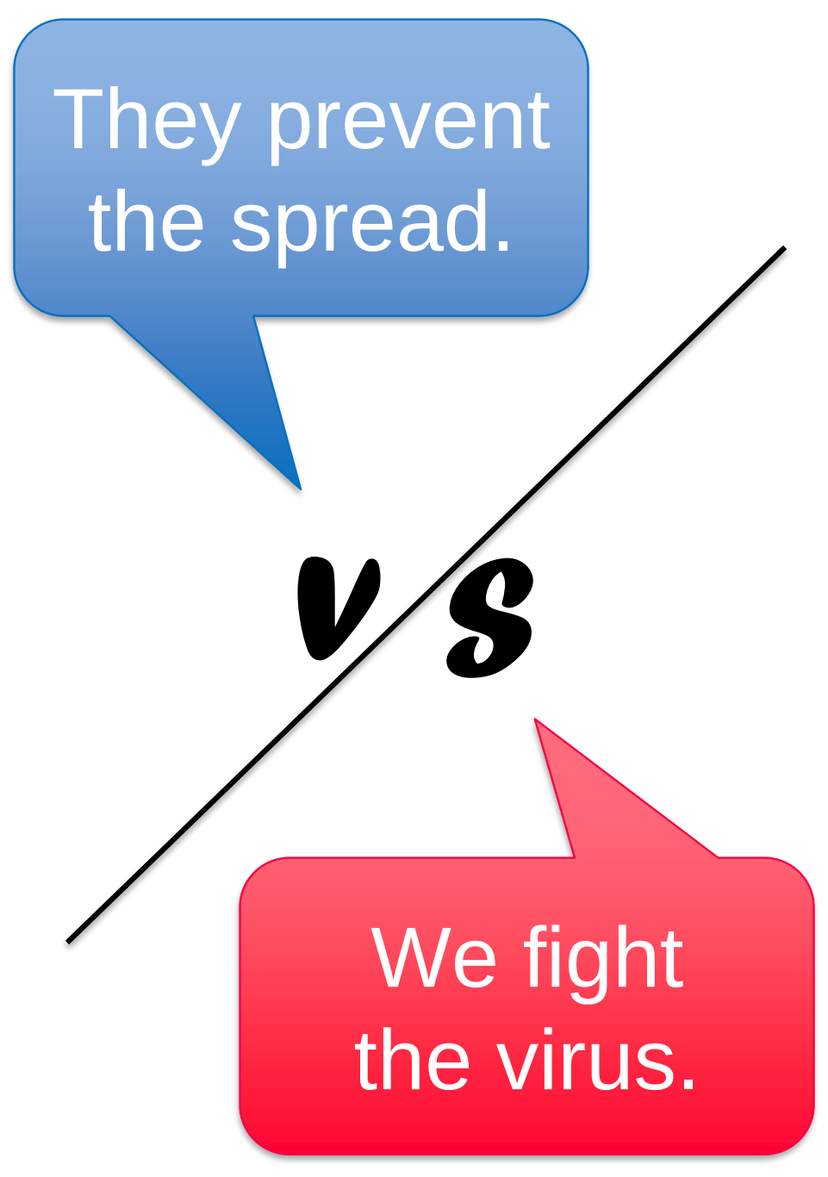
TU GRAZ SCIENCE FOR FUTURE

2022: DIGITAL VISIONS

Unsupervised Transformer-Based Differentiation of Corpus Semantics

Motivation

In a polarized world, understanding differences in argumentation is of utmost importance. Due to the nature of language, the extraction of those differences is challenging. For instance, texts can be similar in content but framed differently. Moreover, supervised machine learning approaches suffer from a lack of labeled training data.



RQ: How to differentiate the framing embedded within text representations in an unsupervised and interpretable manner?

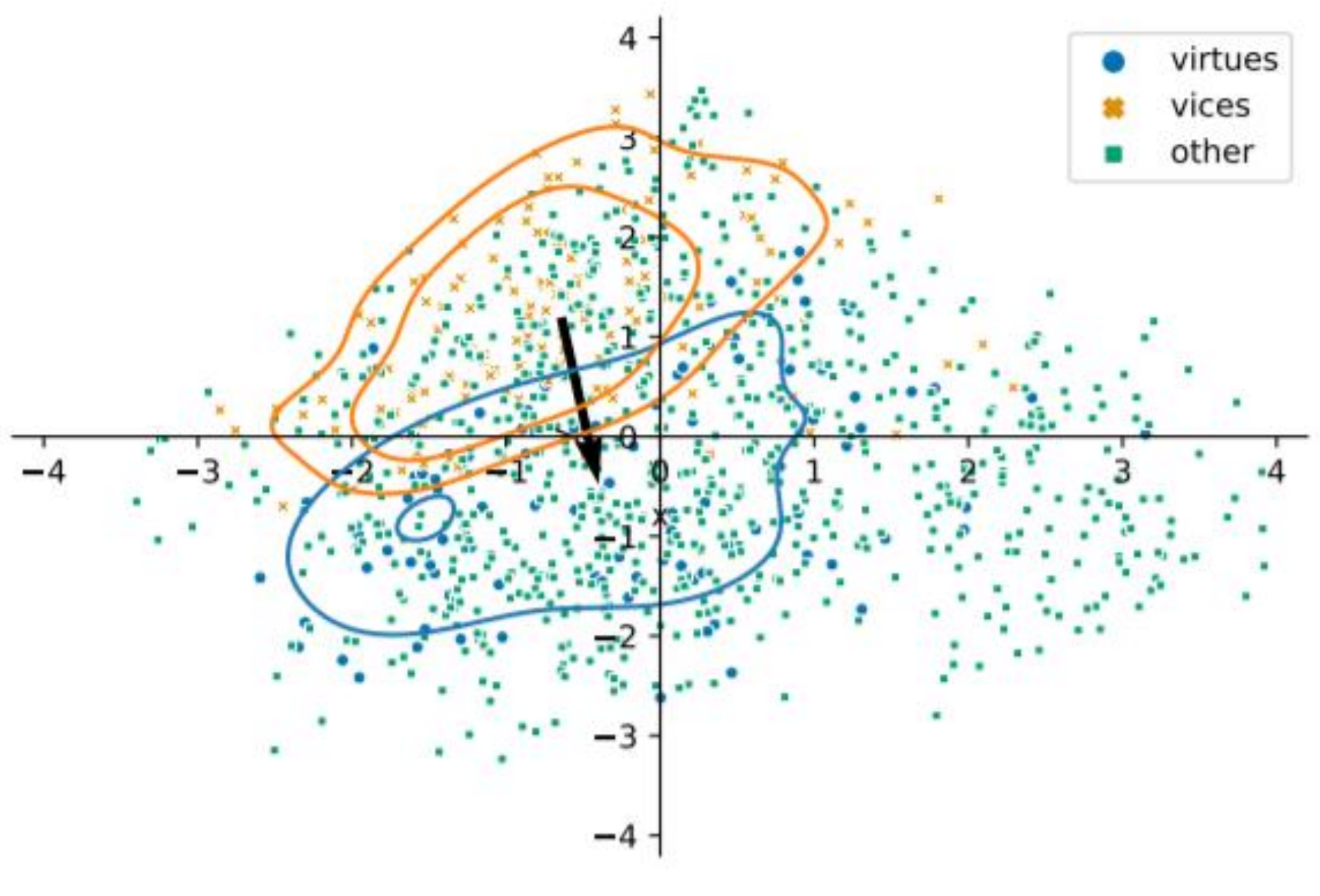
Methods

We investigate unsupervised approaches based on embeddings and transformers for frame extraction to enable the differentiation between two corpora.

We consider five approaches for unsupervised text differentiation. Each method relies on a pre-trained language model. The extraction of results is performed in an unsupervised way (i.e., does not require labeled data). The extracted data is then interpreted by humans.

Table 1: Overview of Approaches

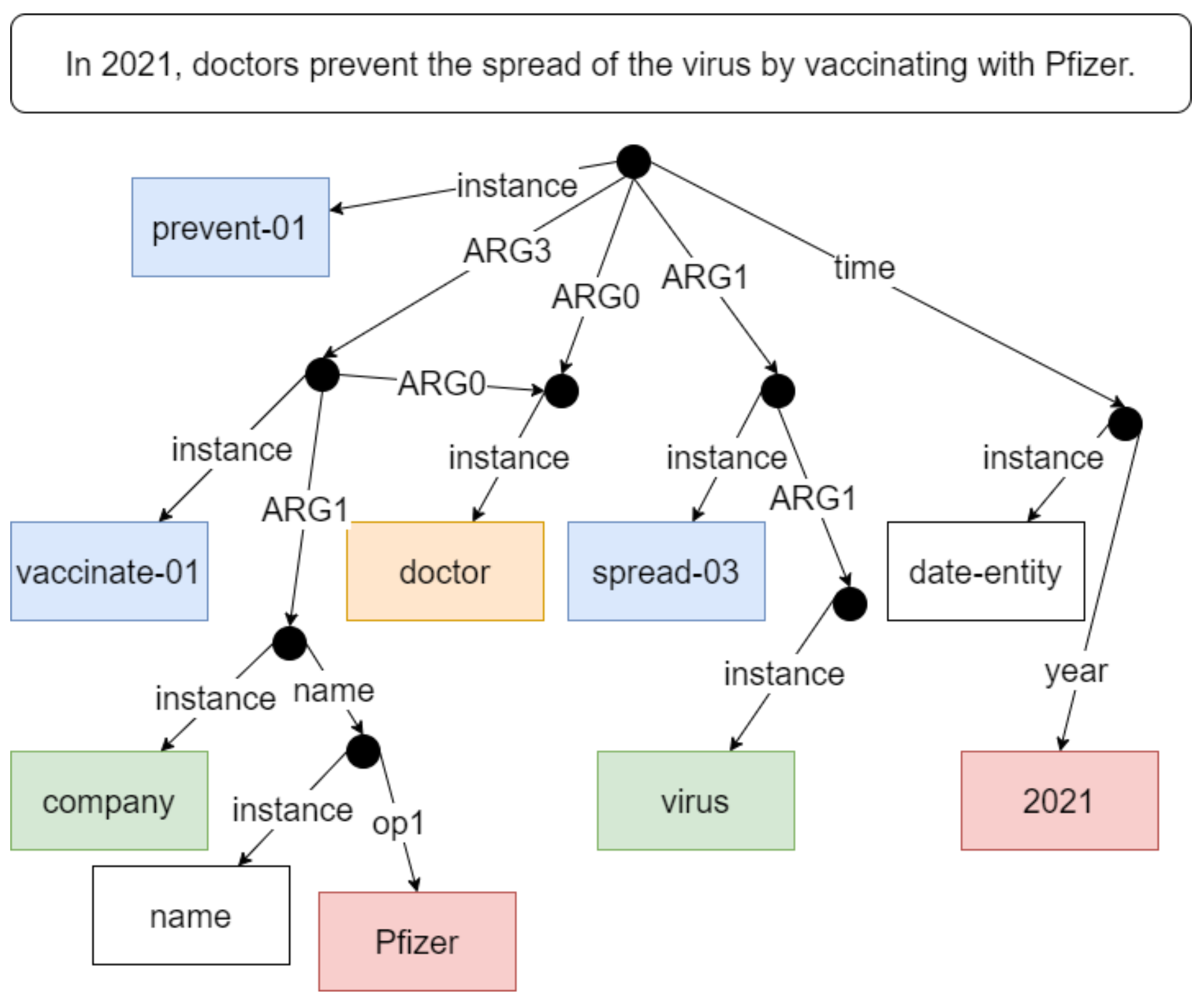
Type	Embeddings[1]	Graphs[2]	Classification	Alignment	Words
Method	FrameAxis + Dictionary	AMR: Abstract meaning representations	Zero-Shot (+ Prompt Tuning)	Paraphrase Similarity	Topics / Keywords
Models	Encoders (BERT) or Word2Vec	Seq2Seq (BART)	NLI on Textual Entailment Task	Sentence BERT	KeyBERT, BERTopic
Data	Continuous (axis)	Discrete (frequency)	Probability (logits)	Cluster Distance	Word Lists/ Mappings
Extract	Bias + Intensity (e.g., Morals)	(Semantic) Frames + Roles	(Verbalized) Labels	Scores (Representativeness)	Top Words (+ Weights)



(a) Care Axis

Embeddings provide a natural notion to contrast and discriminate between texts. As embeddings are high-dimensional, an axis is defined by two poles (e.g., by a set of words or their centroid). However, if a dictionary approach is used, there could be a significant overlap in a projected 2D space (as seen on the left). For instance, consider the word “unwounded” along the care/harm axis (i.e., could be assigned to both).

Graph representations include many nuances of text such as roles (e.g., agent and patients). However, there is no obvious way to aggregate such representations for a whole corpus. A simple method is to consider tuples or triples concerning the graph edges. On the right, there is an AMR representation, which also includes multiple text simplifications.


In 2021, doctors prevent the spread of the virus by vaccinating with Pfizer.

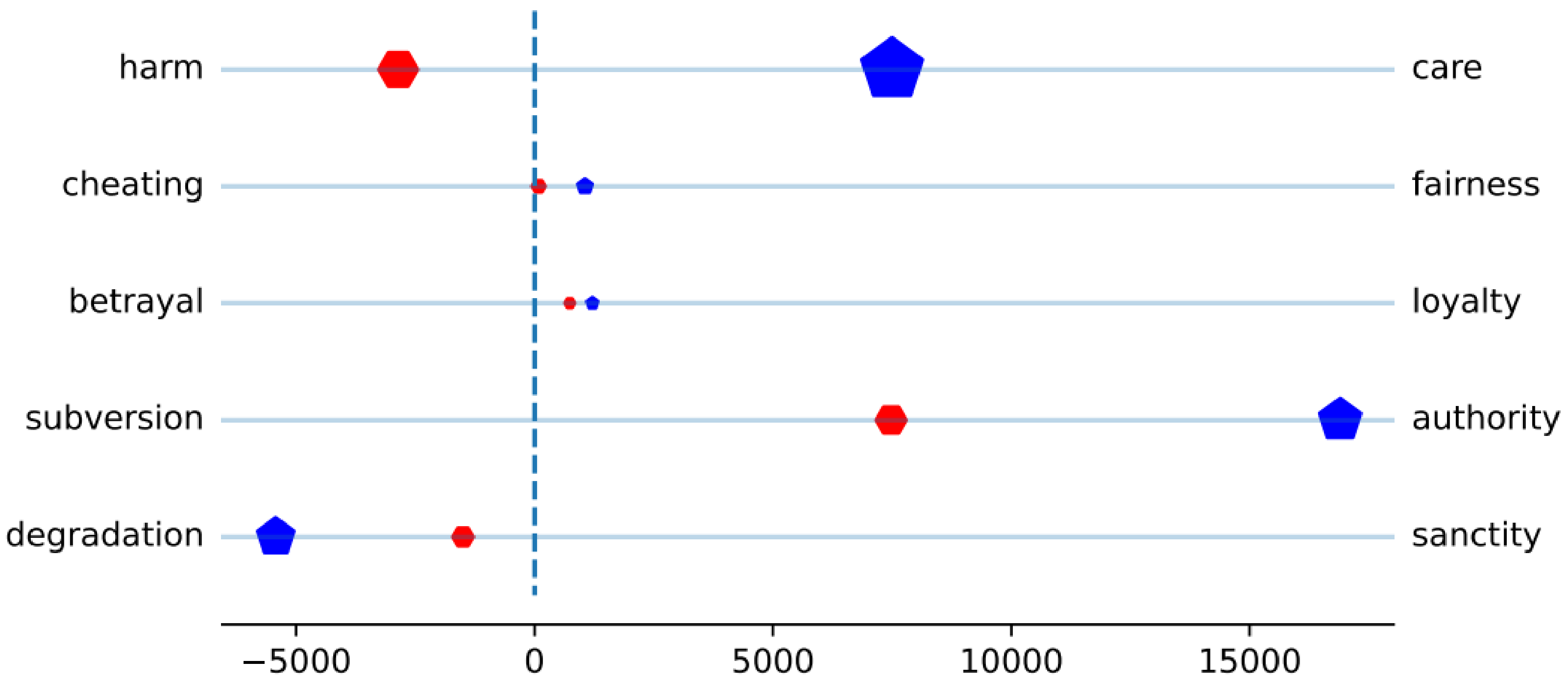
Problem Formulation

Given pair of distinct corpora D_1 and D_2, we want to find a transformation $f(\cdot)$ such that their corresponding aggregation $\bigcup$ maximizes their difference in terms of interpretability:

$$find\ f(\cdot)\ s.t.\ \max\left|\bigcup_{d_i \in D_1} f(d_i) - \bigcup_{d_j \in D_2} f(d_j)\right|$$

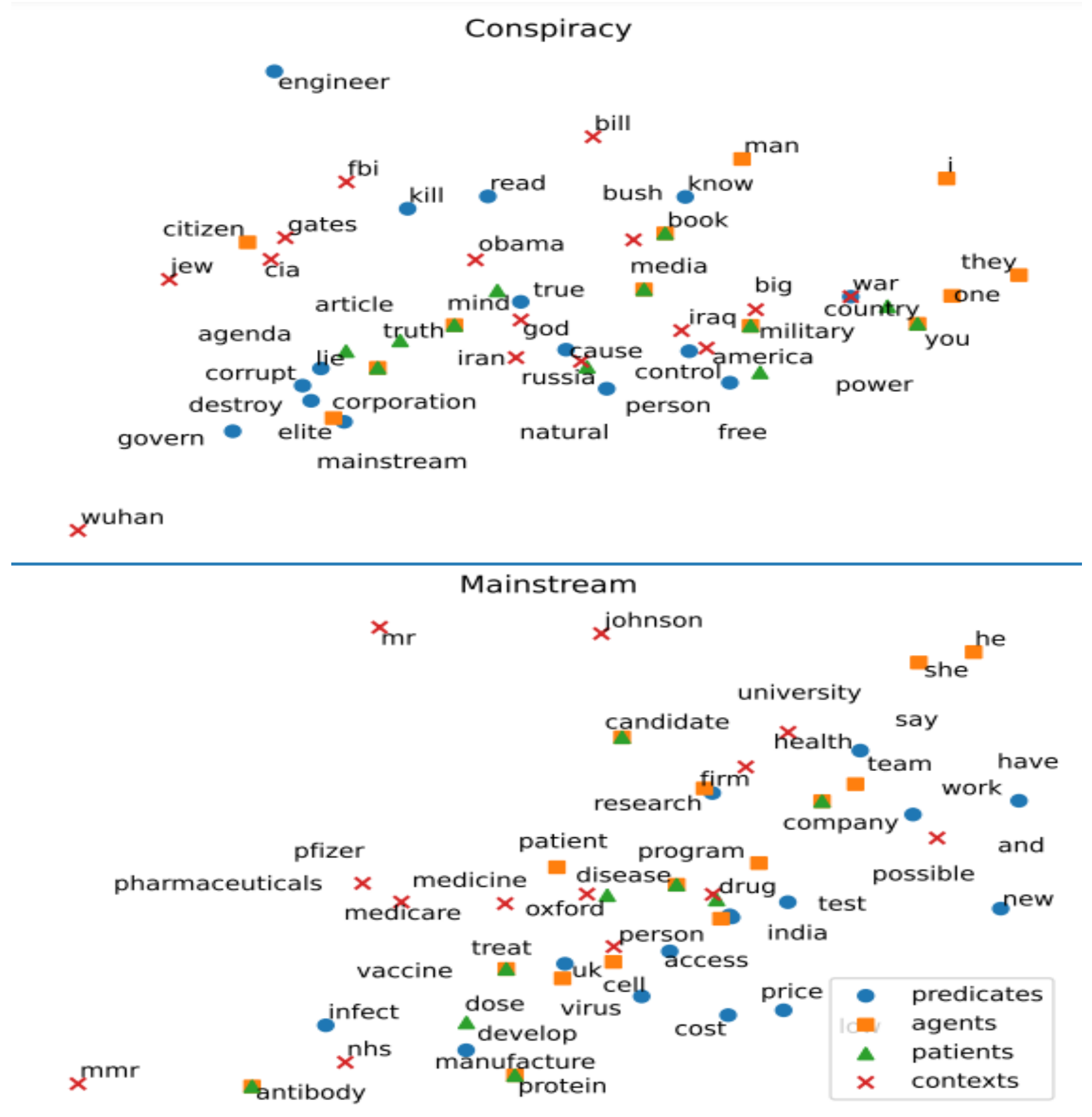
Experiments and Results

In our experiment, we focus on embedding and graph-based approaches.



Embeddings have been used to characterize the moral values of politicians and other politically active people[1]. Similarly (as seen above), we can visualize the moral framing of conspiracy (red) vs. mainstream (blue) media. The bias details the alignment and is projected onto the X-axis. The intensity details the amount of the values that are invoked and is represented by the size.

Using scatterplots, embeddings provide a more advanced visualization (e.g., in comparison to word clouds) of words within a given corpus and can be applied to other methods. On the right, we contrast the frames and concepts extracted from AMR. Specifically, we consider over-representativeness between conspiracy and mainstream as measured by the log-odds ratio.



Discussion and Future Work

Transformer-based models provide novel ways to differentiate between text corpora in an unsupervised and interpretable manner. These methods either improve upon previous methods (e.g., BERTopic) or enable new ways for text analysis (e.g., zero-shot learning).

- Leveraging the performance of pre-trained transformer models provides novel ways to contrast text corpora.
- The approaches provide vastly different representations.
- The evaluation is a key challenge, as human judgment (i.e., interpretability) is subjective, as well as generalizability as certain methods might be better suited for certain corpora.
- Combining multiple approaches seems promising.
- This research will enable social scientists to better analyze vast amounts of text data (e.g., in terms of polarization or conspiracy theories).

For future work, we plan to conduct more experiments with the other methods and contrast their results.

References

[1] Reiter-Haas, M., Kopeinik, S., & Lex, E. (2021, May). Studying Moral-based Differences in the Framing of Political Tweets. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15, pp. 1085-1089).

[2] Reiter-Haas, M., Klösch, B., Hadler, M., & Lex, E. (2022). AMR-based Framing Analysis of COVID-19 Narratives: Conspiracy versus Mainstream Media. In Review.

Markus Reiter-Haas
Institute of Interactive Systems and Data Science
https://socialcomplab.github.io/

INFORMATION, COMMUNICATION & COMPUTING
Fields of Expertise TU Graz