One Word, Two Sides: Traces of Stance in Contextualized Word Representations

Aina Garí Soler, Matthieu Labeau, Chloé Clavel
LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{aina.garisoler,matthieu.labeau,chloe.clavel}@telecom-paris.fr

Abstract

The way we use words is influenced by our opinion. We investigate whether this is reflected in contextualized word embeddings. For example, is the representation of "animal" different between people who would abolish zoos and those who would not? We explore this question from a Lexical Semantic Change standpoint. Our experiments with BERT embeddings derived from datasets with stance annotations reveal small but significant differences in word representations between opposing stances.

1 Introduction

Our opinions are reflected in the way we talk. People with opposing stances on a particular topic may use different words when talking about it. For example, only people against the use of face masks during the COVID-19 pandemic would sometimes refer to them as "muzzles". In this paper, however, we do not investigate *what* words are used by each side. Instead, we compare how speakers who disagree on a subject use the *same* words. Specifically, we want to know whether contextual models capture a difference between the representation of a word (e.g., "mask") when it is used by people who are in favor *vs.* against a certain target (e.g., the compulsory use of face masks).

We address this question from the perspective of Lexical Semantic Change (LSC). Work on LSC typically tries to detect word meaning changes across two or more periods of time (Tahmasebi et al., 2021), but its techniques have also been employed to identify synchronic differences in word usage, for instance across different ages, genders, professions (Gonen et al., 2020), domains (Yin et al., 2018; Schlechtweg et al., 2019), or cultures (Garimella et al., 2016). As opposed to related studies that investigate LSC between different viewpoints (Azarbonyad et al., 2017; Rodriguez et al., 2021), our goal is not to explore the usage of specific words, and we do not evaluate our method based on the ranking of words by meaning stability. We rather want to determine whether vector representations reflect a higher similarity in word usage within a stance than between different stances (see example in Figure 1). We explore this question relying on datasets annotated with stance information. Before that, we test different context-sensitive embedding models on a simulated scarce-data setting. This allows us to select a robust representation type that can identify the words that are used most differently between stances.

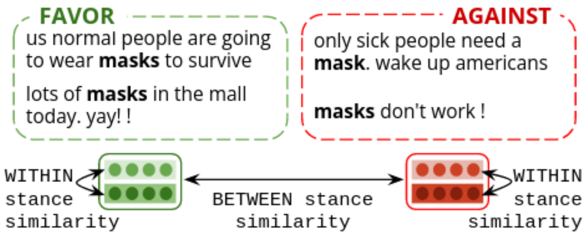

Figure 1: Example instances of "mask" from the Covid19 stance dataset (Glandt et al., 2021). We compare the within- and the between-stance usage similarity.

Our long-term goal is to detect differences in word usage between speakers in a conversation, which could point to their level of conceptual alignment (Stolk et al., 2016); that is, the extent to which dialog participants "mean the same things when using the same words" (Schober, 2005). In this study we present a first step in this direction. Representations that are sensitive to opinion differences could be useful to identify disagreements and misalignment in dialog.

2 Methodology

In this section we introduce the data and the models used in our experiments. We also describe our

Proceedings of the 29th International Conference on Computational Linguistics, pages 3950–3959
October 12–17, 2022.

similarity measure and the criteria for evaluation.[1]

2.1 Data

The datasets we use are in English and contain stance information in the form of sentences that are labeled as being in FAVOR or AGAINST a specific target. We exclude sentences with no (clear) stance (NONE), when present. **SemEval2016** (Mohammad et al., 2016b,a) contains tweets on six varied targets. We use 3,253 sentences.[2] **Covid19** (Glandt et al., 2021) is another dataset with 3,918 tweets centered on four targets related to the COVID-19 pandemic. **P-stance** (Li et al., 2021) is a large dataset containing 21,574 tweets about three US politicians. Finally, IBM-ArgQ-Rank-30kArgs (Gretz et al., 2020), hereafter **ArgQ**, is a collection of arguments on 71 targets which are annotated for stance, stance clarity and argument quality. We use 29,972 arguments that have a clear stance (with a confidence score[3] above 0.6, following Bar-Haim et al. (2020)).

We want to organize the data in a way that allows us to investigate whether instances of the same word have a higher similarity within a stance than between stances. To this end, we preprocess and organize the data as follows.

Preprocessing The ArgQ dataset was originally intended for argument quality detection, and several arguments mention their stance explicitly. To mitigate the potential biases that this could cause, we apply a strategy that we call *sentence trimming* which automatically omits this part of a sentence. We describe it in detail in Appendix A. Then we tokenize, postag and lemmatize sentences in all datasets. [4]

Sentence Sets For a given target, we randomly split the sentences of each stance (f or a) into two equally-sized sets P and Q. With these sets, we run four *comparisons*, two within-stance: WITHIN-FAVOR (P_f vs Q_f) and WITHIN-AGAINST (P_a vs Q_a); and two between-stance: BETWEEN-1 (P_f vs Q_a) and BETWEEN-2 (P_a vs Q_f).

2.2 Vector Representations

We want to generate vector representations for sets of word instances within a stance (e.g., in P_f). For example, we want to obtain one representation of the word "woman" from sentences in favor of the "Feminist Movement" (SemEval2016) and compare it to the representation of "woman" in sentences expressing a stance against this target.

In LSC detection, static embeddings tend to perform better than contextualized ones (Schlechtweg et al., 2020). A typical approach is to learn static embeddings separately for each time period, corpus or viewpoint, and then compare them either by aligning them (Hamilton et al., 2016) or with a nearest-neighbors-based approach (Gonen et al., 2020). In these studies, even in those dealing with short-term change detection (Stewart et al., 2017; Del Tredici et al., 2019), it is common to have a fairly large amount of instances of a given word available. However, the number of available sentences per word within a stance in our data is limited.[5] We therefore experiment with three different types of contextualized embeddings:

À la carte embeddings (ALC) (Khodak et al., 2018) have been used to detect differences in word usage across viewpoints (Rodriguez et al., 2021). The model consists in applying a linear transformation to the averaged pre-trained embeddings of the context words surrounding the target word. We use an ALC model relying on 300*d* GloVe embeddings (Pennington et al., 2014) trained on 840B tokens from Common Crawl.

Context2vec (c2v) (Melamud et al., 2016) is a biLSTM model that generates embeddings for the context surrounding a word. It is optimized so that the representation of a context is similar to that of potential filler words. We use a 600*d* model trained on the ukWaC corpus (Baroni et al., 2009).

BERT (Devlin et al., 2019). We use contextualized representations generated with the 768*d* `bert-base-uncased` model. We explain how we choose the best layer in Section 2.3.

We denote the vocabulary of a sentence set (e.g. P) as V_P. We include in the vocabulary all nouns and verbs appearing in at least three different sen-

[1] Our code and data are available at `https://github.com/ainagari/1word2sides`.

[2] We omit the target "Climate Change is a Real Concern" because it only has 26 AGAINST tweets.

[3] This score reflects the extent to which annotators agreed on the stance of an argument. It is calculated as a weighted average of the annotators' decisions and it ranges from 0 to 1.

[4] We use the default `nltk` functions, except for tweets, which we tokenize with `nltk`'s TweetTokenizer. Lemmatization is done with `nltk`'s WordNet Lemmatizer.

[5] As an example, Schlechtweg et al. (2020) have an average of 788 instances per lemma and time period; and Gonen et al. (2020) study words that appear at least 200 times in their corpus. In our data, the average amount of instances of a word in one side of a comparison is 14.

tences in P. In tweets, mentions and hashtags are treated as nouns. Stopwords are excluded. We treat all instances of a lemma with a specific part of speech (PoS) as the same word. We extract a vector representation $\mathbf{w}_P$ for every word w in V_P. For c2v and BERT, this is done by averaging the representations of all w instances in P.

2.3 Testing Representations

Before our experiments on stance, we first identify the vector representations that are best suited to reflect lexical semantic similarity between small sets of sentences. Following Schlechtweg and Schulte im Walde (2020), we use SemCor (Miller et al., 1993), a sense-annotated corpus, to create a dataset that simulates lexical semantic change. We additionally control for the amount of sentences available for each lemma. The process of creation of this dataset is explained in more detail in Appendix B.

The dataset consists of 576 lemmas: 245 nouns, 241 verbs, 69 adjectives and 21 adverbs. For every lemma we have two sets of 25 instances each, P and Q. To simulate situations of scarce data, we create X-sized subsets of P and Q (P_X, Q_X). We experiment with different values of X ($X \in \{3, 5, 10, 20, 25\}$). As in Schlechtweg and Schulte im Walde (2020), we determine the "true" semantic distance between two groups P_X and Q_X by calculating the Jensen-Shannon divergence (JSD) between their sense distributions.

Similarity predictions for a word w are obtained by simply calculating the cosine similarity between the representations of that lemma in each sentence set, $cos(\mathbf{w}_{P_X}, \mathbf{w}_{Q_X})$. We report the Kendall's tau-b correlation coefficient between JSD and the similarities predicted by each representation type. Results of this experiment are presented in Section 3.1.

2.4 Similarity Calculation

To calculate the global similarity in word usage for a comparison between two sets of sentences P and Q, we first identify the words that are common in both sets, $V_P \cap V_Q$. $V_P \cap V_Q$ contains words that are not necessarily central to the target that is being discussed. We therefore calculate a similarity based only on a subset of $V_P \cap V_Q$, which we call V_{PQ}. The similarity score is the average cosine similarity of all words in V_{PQ}:

$$sim(P, Q) = \frac{\sum\limits_{w \in V_{PQ}} cos(\mathbf{w}_P, \mathbf{w}_Q)}{|V_{PQ}|} \tag{1}$$

This similarity measure is intended to reflect the extent to which words are used in the same way and in the same senses in two sentence sets. We experiment with three definitions of V_{PQ}. In all of them, we take care of using the same amount of words for all four comparisons within a target. In *all*, we include the top k most frequent words in $V_P \cap V_Q$, where k corresponds to the smallest size of $V_P \cap V_Q$ available for that target. Frequency is determined from the union of sentences in P and Q. We also use the top 10 words in $V_P \cap V_Q$ with highest tf-idf scores in that target (*tf-idf*). Tf-idf scores are calculated on the ensemble of stance datasets, treating all sentences about the same target as one document. Finally, we also use the 10 words in $V_P \cap V_Q$ with lowest tf-idf (*rev-tf-idf*). This subset contains words that are less relevant to the target, and therefore we expect BETWEEN- and WITHIN-stance similarities to have closer values in this setting. Note that 25% of comparisons (in SemEval2016 and ArgQ) have less than 20 words in common. In these cases, *tf-idf* and *rev-tf-idf* are partially calculated with the same words.

2.5 Evaluation

We expect WITHIN-stance comparisons to exhibit a higher average similarity than BETWEEN-stance comparisons. To measure the extent to which this holds, we use pairwise accuracy: we check for how many (WITHIN, BETWEEN) comparison pairs the BETWEEN comparison has a lower similarity. With 4 comparisons per target, our experiments involve a total of 332 (WITHIN, BETWEEN) pairs. Results on stance data are presented in Section 3.2.

3 Results

3.1 Selecting a Representation Type

Results on SemCor are shown in Figure 2. In plots *a* and *b*, we see the correlations obtained by the different representation types on various amounts of data (X). Naturally, performance is worse with lower values of X. This is especially the case of ALC embeddings, which at X=25 continue to improve. In the case of c2v and BERT, however, we do not observe big improvements after X=10. In this scarce-data setting, the performance of ALC

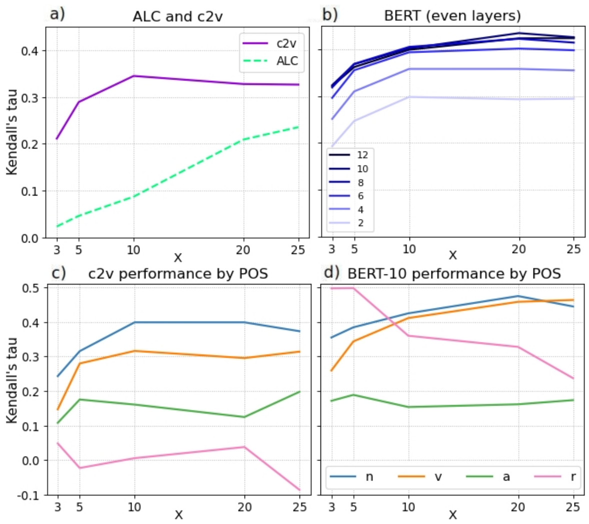

Figure 2: *a* and *b*: Kendall's tau obtained by different vector representations on SemCor. We only include even layers for BERT for better readability. *c* and *d*: Performance of c2v and BERT (10th layer) by PoS.

embeddings is much lower than that of c2v and BERT. Overall, BERT representations from the 10th layer work best. We therefore use embeddings from this layer for our experiments on stance data. We also look at the performance of the best two models (c2v and the 10th layer in BERT) by PoS (plots *c* and *d*): we find that nouns and verbs, the PoS included in our stance experiments, are generally better represented. We also make interesting observations regarding the other PoS. Despite the lower performance, adjective representations seem to be less affected by a smaller number of sentences. When it comes to BERT adverb representations, similarity estimations are more reliable at lower values of X. These differences in PoS should be taken into account when deriving type-level vectors from BERT representations.

3.2 Results on Stance

Pairwise accuracy obtained with the 10th BERT layer with different definitions of V_{PQ} is found in Table 1. We see that, especially for *all* and *tf-idf*, pairwise accuracy is remarkably high in all datasets. This shows that contextualized word representations from BERT reflect differences in the way words are used between two opposing stances.

When using the 10 words with lowest tf-idf (*rev-tf-idf*) performance decreases, but is still high in P-stance and ArgQ. We run chi-square goodness-of-fit tests on *rev-tf-idf* predictions to determine their likelihood under the null hypothesis (H_0 : acc

Dataset	all	tf-idf	rev-tf-idf
SemEval2016	0.90	0.85	0.60
Covid19	0.88	0.81	0.50
P-stance	1.00	1.00	0.83
ArgQ	1.00	0.98	0.95
Global	0.99	0.96	0.90

Table 1: Pairwise accuracy by dataset and with different V_{PQ}. *Global* corresponds to all datasets put together.

	all	tf-idf	rev-tf-idf
a) W VS W	0.013	0.010	0.023
b) B VS B	0.013	0.010	0.023
c) W VS B	0.047	0.027	0.041

Table 2: Differences in similarity between comparisons.

$= 0.5$). P-values are significant for all datasets together ($p < 0.001$) but not for the set of Twitter datasets ($p = 0.08$, $\alpha = 0.05$).[6] It seems BERT representations do, to some extent, encode differences in words that are less relevant to the target. However, if for some reason not all words can be used (if there are too many), then it is preferable to select a subset carefully (e.g. with tf-idf).

We also examine the words that have the highest and the lowest similarities in BETWEEN comparisons; we provide this information in Appendix C. The words that are used most differently between stances tend to be nouns that are central to the topic (e.g. "religion" in "Atheism"), while the most similar words are often non-topical ("man" or "take"). In the middle of the distribution, in targets with a small common vocabulary (<30) we find words that are relevant to the topic, but in a less obvious way (e.g. "world" and "community" for the target "Missionary work"). In targets with a larger vocabulary we find a combination of relevant and non-relevant words.

We investigate how large the differences in similarity are between WITHIN (W) and BETWEEN (B) comparisons. We investigate this by looking at the differences in similarity (in absolute value) across comparison pairs: a) between WITHIN-FAVOR and WITHIN-AGAINST (W VS W), b) between BETWEEN-1 and BETWEEN-2 (B VS B), and c) the average difference found in the four WITHIN VS BETWEEN pairings (W VS B). We expect the latter to have a larger difference in similarity than

[6]This could be due to particularities of the language used in Twitter. We leave the use of models specialized on tweets (e.g. BERTweet (Nguyen et al., 2020)) for future work.

a) and b)), where comparisons are of the same type. Results are shown in Table 2. We report the average of these values on all the data. Differences in similarity are quite low overall, indicating that the contrast (i.e., the extent to which WITHIN comparisons display a higher similarity than BETWEEN comparisons) is subtle. Values are, however, between 1.8 and 3.6 times larger for the W vs B comparison pairs. For all V_{PQ} definitions, the difference values in these comparison pairs are significantly different from those in a) and b) ($p < 0.001$).[7]

4 Conclusion and Future Work

We have shown that BERT word representations are sensitive to the opinion expressed in the sentences they are derived from. Differences in similarity found between concurring and conflicting stances are small, but significant; and words with the highest differences tend to be central to the topic. Our approach can serve to identify points of discrepancy with regard to a target, and it can be useful for stance detection and debate analysis. Our experiments on SemCor provide valuable insight on the sufficient amount of word instances needed to obtain quality representations. This is relevant for low-resource LSC and, more generally, for inferring word vectors from little data.

In future work, we plan to apply this methodology to dialog. Sets P and Q would each correspond to the utterances of one speaker in a conversation. The similarity measure would act as an approximation of the conceptual or *stance alignment* between the two participants, indicating whether speakers share opinions and use words in a similar way.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback and insightful comments. This work was supported by the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS).

References

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words Are Malleable: Computing Semantic Shifts in Political and Media Discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1509–1518, New York, NY, USA. Association for Computing Machinery.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From Arguments to Key Points: Towards Automatic Argument Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-Term Meaning Shift: A Distributional Exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying Cross-Cultural Differences in Word Usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance Detection in COVID-19 Tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis. *Proceedings of the*

[7]According to Wilcoxon or paired t-tests depending on normality (determined by Shapiro-Wilk tests).

AAAI Conference on Artificial Intelligence, 34:7805–7813.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Melbourne, Australia. Association for Computational Linguistics.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-Stance: A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A Dataset for Detecting Stance in Tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Pedro L Rodriguez, Arthur Spirling, and Brandon M Stewart. 2021. Embedding regression: Models for context-specific description and inference. Technical report, Working Paper Vanderbilt University.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg and Sabine Schulte im Walde. 2020. Simulating Lexical Semantic Change from Sense-Annotated Data. In *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)*.

Michael F Schober. 2005. Conceptual alignment in conversation. *Other minds: How humans bridge the divide between self and others*, pages 239–252.

Ian Stewart, Dustin Arendt, Eric Bell, and Svitlana Volkova. 2017. Measuring, Predicting and Visualizing Short-Term Change in Word Representation and Usage in VKontakte Social Network. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):672–675.

Arjen Stolk, Lennart Verhagen, and Ivan Toni. 2016. Conceptual alignment: How brains achieve mutual understanding. *Trends in cognitive sciences*, 20(3):180–191.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational approaches to semantic change*. Language Science Press.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The global anchor method for quantifying linguistic shifts and domain adaptation. *Advances in neural information processing systems*, 31.

A Sentence Trimming

Sentence trimming is intended to omit a part of a sentence in the ArgQ dataset where stance is expressed explicitly. These sentences often start with the same words as the target. For example, for the target “Homeschooling should be banned”, we find the sentence “’Homeschooling should not be banned because it is a right for parents to educate their children in their comfort of home”. If the beginning of a sentence contains the same words as the target (with the optional addition of *not* and *n’t*) and is followed by the token *because (of)*, *as*, *since*, a comma or a stop, we omit the first part of the sentence up to and including that token. In the example above, this results in the sentence “it is a right for parents to educate their children in their comfort of home”. This procedure modifies 3,223 sentences. Some sentences with an explicit stance remain, but their number is importantly reduced. These include sentences starting with the target followed by connectors expressing effect (e.g., *so that*, *so as to*), which cannot be easily trimmed into a correct sentence or NP.

B Dataset for Testing Representations

In this section we describe in detail how we collect the data from SemCor (see Section 2.3). We randomly select 50 instances for every lemma that appears at least 50 times in SemCor. These instances are randomly split into two sets of 25 sentences each, P and Q. The X-sized subset of P, P_X, consists of the X first sentences in P. This approach results in a dataset with rather low JSD, especially for larger values of X. For example, for $X = 25$, the mean JSD is 0.22 and only 2% of lemmas have JSD > 0.5. To have a stronger representation of high JSD values, we maximize JSD for certain lemmas. We do this for a subset of the lemmas for which it is possible to find a P-Q split with zero sense overlap, such that JSD $= 1$. Enforcing these splits for $\sim$17% of all lemmas, the mean JSD for $X = 25$ goes up to 0.33.

C Highest- and Lowest-Similarity Words

Table 3 contains, for every target in our study, the words that differed the most and the least between FAVOR and AGAINST statements. Interestingly, among the top five most different words across all targets, we find a majority of nouns (85.9% nouns and 14.1% verbs). In the bottom five, instead, verbs are more common (38.1% nouns and 61.9% verbs).

Data	Target	Sentences	Most different words	Least different words
SemEval 2016	Feminist Movement	779	woman, men, equality woman, men, gender	come, leave, believe go, take, tell
	Hillary Clinton	728	@hillaryclinton, #hillaryclinton, woman @ hillaryclinton, #hillaryclinton, campaign	keep, world, go make, take, come
	Donald Trump	447	@realdonaldtrump, trump, #makeamericagreatagain @realdonaldtrump, trump, donald	want, give, take want, one, time
	Atheism	588	religion, #god, believe #freethinker, religion, god	man, think, go take, make, come
	Legalization of Abortion	711	abortion, woman, right abortion, woman, right	think, know, say take, carry, effect
Covid19	Face masks	1,361	mask, wear, people wear, mask, people	love, look, shut care, find, care
	Stay at home orders	590	#covid19, #coronavirus, virus #covid19, #coronavirus, virus	day, order, thing let, must, see
	Fauci	1,102	#drfauci, #coronavirus, #covid19 #drfauci, #covid19, #coronavirus	force, work, right leave, history, work
	School closures	865	@imbhupendrasinh, @vijayrupanibjp, school school, kid, @realdonaldtrump	time, do, need come, way, show
P-stance	Donald Trump	7,953	@realdonaldtrump, #donaldtrump, country @realdonaldtrump, #trump, say	color, head, pay arm, apply, wish
	Bernie Sanders	6,325	@berniesanders, bernie, #democraticdebate @berniesanders, bernie, sander	check, note, ill assume, knock, sick
	Joe Biden	7,296	#democraticdebate, @joebiden, #demdebate #democraticdebate, @joebiden, biden	name, sign, like dirt, tear, air
ArgQ	Marriage	413	marriage, people, couple marriage, couple, people	union, make, need create, become, thing
	Vow of celibacy	418	celibacy, vow, church celibacy, vow, people	need, take, way nothing, way, time
	Stay-at-home dads	392	home, dad, raise home, dad, men	make, provide, life time, allow, make
	Assisted suicide	392	suicide, assist, people suicide, assist, people	help, take, make death, take, make
	Fast food	416	food, eat, ban food, people, ban	health, make, issue world, make, time
	Urbanization	404	area, urbanization, city urbanization, people, area	space, create, grow population, make, create
	Missionary work	434	people, missionary, work work, people, missionary	make, take, way make, want, need
	Libertarianism	381	libertarianism, government, people libertarianism, government, people	lead, give, provide take, one, work
	Human cloning	416	clone, cloning, human cloning, clone, human	life, need, way make, thing, life
	Blockade of the Gaza Strip	506	strip, gaza, blockade strip, gaza, blockade	stop, right, state state, get, give
	Gender-neutral language	368	language, gender, people language, gender, people	offend, way, time make, feel, way
	Compulsory voting	405	voting, compulsory, vote vote, compulsory, people	make, way, want take, mean, could
	Zero-tolerance policy in schools	454	school, tolerance, student school, student, policy	lead, way, time way, make, time
	Payday loans	442	loan, people, need loan, money, people	situation, take, need take, make, give
	Whaling	423	whale (N), whaling, whale (V) whale (N), whale (V), whaling	help, way, need part, need, world
	Capital punishment	467	punishment, capital, death capital, punishment, crime	justice, make, serve way, give, time
	Cosmetic surgery for minors	494	minor, surgery, child surgery, minor, decision	thing, involve, give need, adult, cause

Data	Target	Sentences	Most different words	Least different words
ArgQ	School uniform	474	school, student, uniform school, uniform, student	stop, take, allow make, give, feel
	Foster care	529	child, kid, care child, parent, care	may, service, find become, make, put
	Polygamy	493	polygamy, legalize, marriage polygamy, marriage, woman	make, take, one way, make, time
	Prostitution	499	prostitution, legalize, prostitute prostitution, legalize, woman	give, allow, want choice, involve, want
	Zoos	395	animal, zoo, live animal, zoo, habitat	life, allow, make provide, keep, take
	The right to keep and bear arms	407	keep, bear, arm bear, keep, weapon	law, take, remove person, must, take
	Social media	330	medium, people, allow medium, people, allow	create, make, lose see, world, time
	Multi-party system	390	system, people, multiparty party, system, government	bring, need, allow choose, population, thing
	Nuclear weapons	542	weapon, country, use weapon, country, war	maintain, keep, life mean, make, world
	Homeschooling	395	child, homeschooling, school child, homeschooling, education	give, time, keep help, teacher, way
	Telemarketing	437	telemarketing (N), telemarketing (V), telemarketers telemarketing (V), telemarketing (N), telemarketers	allow, need, take money, work, time
	Entrampment	400	law, crime, entrapment crime, entrapment, commit	get, make, allow place, time, know
	Homeopathy	352	medicine, homeopathy, remedy homeopathy, medicine, people	harm, condition, placebo treat, cause, allow
	Intelligence tests	462	intelligence, people, person person, test, child	way, base, focus show, type, know
	Austerity regime	412	regime, austerity, economy regime, austerity, debt	spend, time, make reduce, pay, allow
	Child actors	435	actor, child, use actor, child, use	take, show, play take, make, lead
	Mandatory retirement	475	retirement, work, worker retirement, workforce, worker	make, position, force keep, provide, give
	Sex selection	400	selection, child, parent selection, baby, sex	allow, could, decide bear, right, way
	Economic sanctions	389	sanction, country, nation sanction, country, people	leader, make, take make, punish, help
	Intellectual property rights	415	property, right, product property, right, people	come, make, time time, take, think
	Use of public defenders	415	lawyer, defender, use defender, lawyer, defend	get, require, way person, mean, allow
	Guantanamo Bay detention camp	444	guantanamo, bay, detection guantanamo, detection, camp	serve, way, use law, make, usa
	Women in combat	370	combat, woman, men combat, woman, men	prohibit, could, make war, may, make
	Naturopathy	536	medicine, naturopathy, treatment naturopathy, medicine, treatment	lead, take, life seek, allow, make
	Church of Scientology	401	scientology, church, ban scientology, church, ban	member, believe, practice need, allow, practice
	Embryonic stem cell research	396	stem, cell (N), cell (V) cell, stem, research	help, need, use people, need, life
	Affirmative action	438	action, people, job action, people, discrimination	way, get, make school, way, work
	Cannabis	543	cannabis, marijuana, legalize cannabis, marijuana, drug	take, time, way may, allow, take

Data	Target	Sentences	Most different words	Least different words
ArgQ	Vocational education	418	education, school, subsidize education, subsidize, people	lead, make, way work, go, give
	Racial profiling	412	profiling, criminal, people profiling, people, crime	make, person, life stop, time, way
	Private military companies	392	company, ban, government company, government, military	could, make, time security, need, might
	Flag burning	426	burning, flag, burn flag, burning, burn	protect, freedom, make lead, protect, state
	Surrogacy	431	surrogacy, baby, woman surrogacy, woman, surrogate	right, become, term give, make, could
	Student loans	369	student, loan, education loan, student, subsidize	everyone, put, make afford, work, make
	Safe spaces	388	space, people, student space, people, others	life, may, thing make, allow, nothing
	Algorithmic trading	387	trading, people, market trading, computer, market	access, allow, base field, risk, lead
	Olympic games	409	olympic, game, olympics olympic, game, athlete	money, world, time give, time, take
	Journalism	357	journalism, news, subsidize journalism, subsidize, news	medium, need, could could, need, support
	Cosmetic surgery	425	surgery, people, appearance surgery, people, ban	make, take, lead feel, need, way
	Targeted killing	409	target, people, kill target, killing, people	use, state, take enemy, take, put
	Organ trade	408	trade, organ, sell trade, organ, legalize	give, death, way need, create, help
	Space exploration	381	space, exploration, subsidize space, exploration, planet	thing, support, country thing, find, use
	Factory farming	410	farm, factory, food factory, food, farming	space, allow, keep produce, keep, allow
	Pride parades	394	parade, pride, gay parade, pride, lgbt	right, allow, make way, want, bring
	Collectivism	440	collectivism, group, people collectivism, people, society	need, one, way take, lead, way
	Television	387	television, people, watch television, news, entertainment	way, thing, keep could, way, make
	School prayer	424	school, prayer, religion prayer, school, religion	allow, take, person part, time, place
	Autonomous cars	445	car, road, drive car, road, drive	cause, way, need take, use, time
	Holocaust denial	456	holocaust, denial, deny holocaust, denial, deny	speech, allow, go allow, world, say
	Executive compensation	375	executive, compensation, company executive, company, compensation	give, deserve, lead level, work, allow
	Three-strikes laws	490	law, strike, crime law, strike, people	take, make, need give, put, allow
	Atheism	360	atheism, god, religion atheism, religion, people	base, allow, make way, provide, lead
	Wikipedia	395	wikipedia, subsidize, information wikipedia, wikipedia, subsidize	could, need, take provide, way, give
	Judicial activism	385	judge, law, activism judge, activism, law	use, need, way allow, rule, could

Table 3: Words with the highest and lowest differences for every target with representations from the 10th layer of BERT. The two rows for each target correspond to BETWEEN-1 and BETWEEN-2, respectively. Target names in ArgQ have been abbreviated for convenience. For example, the target "Marriage" was originally "We should abandon marriage".