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.

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
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 score3 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 (Pf vs Qf) and WITHIN-AGAINST (Pa vs Qa); and two between-stance: BETWEEN-1 (Pf vs Qa) and BETWEEN-2 (Pa vs Qf).

2.2 Vector Representations

We want to generate vector representations for sets of word instances within a stance (e.g., in Pf). 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 300d 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 600d model trained on the ukWaC corpus (Baroni et al., 2009).

BERT (Devlin et al., 2019). We use contextualized representations generated with the 768d 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 Vp. We include in the vocabulary all nouns and verbs appearing in at least three different sen-

---

1Our code and data are available at https://github.com/ainagari/word2sides.
2We omit the target “Climate Change is a Real Concern” because it only has 26 AGAINST tweets.
3This 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.
4We use the default nltk functions, except for tweets, which we tokenize with nltk’s TweetTokenizer. Lemmatization is done with nltk’s WordNet Lemmatizer.
5As 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 $w_P$ 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, $\cos(w_{P_X}, w_{Q_X})$. 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 $V_{PQ}$:

$$sim(P, Q) = \frac{\sum_{w \in V_{PQ}} \cos(w_P, w_Q)}{|V_{PQ}|}$$

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 ($tfidf$). 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 ($rev-tfidf$). 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, $tfidf$ and $rev-tfidf$ 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
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. 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 ($H_0 : acc = 0.5$). P-values are significant for all datasets together ($p < 0.001$) but not for the set of Twitter datasets ($p = 0.08, \alpha = 0.05$). 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 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>all tf-idf</th>
<th>rev-tf-idf</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemEval2016</td>
<td>0.90 0.85</td>
<td>0.60</td>
</tr>
<tr>
<td>Covid19</td>
<td>0.88 0.81</td>
<td>0.50</td>
</tr>
<tr>
<td>P-stance</td>
<td>1.00 1.00</td>
<td>0.83</td>
</tr>
<tr>
<td>ArgQ</td>
<td>1.00 0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>Global</td>
<td>0.99 0.96</td>
<td>0.90</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>all tf-idf</th>
<th>rev-tf-idf</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) W vs W</td>
<td>0.013 0.10</td>
<td>0.023</td>
</tr>
<tr>
<td>b) B vs B</td>
<td>0.013 0.10</td>
<td>0.023</td>
</tr>
<tr>
<td>c) W vs B</td>
<td>0.047 0.027</td>
<td>0.041</td>
</tr>
</tbody>
</table>

Table 2: Differences in similarity between comparisons.

---

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) ($p < 0.001$).

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


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


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).
<table>
<thead>
<tr>
<th>Data</th>
<th>Target</th>
<th>Sentences</th>
<th>Most different words</th>
<th>Least different words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feminist Movement</td>
<td>779</td>
<td>woman, men, equality woman, men, gender</td>
<td>come, leave, believe go, take, tell</td>
<td></td>
</tr>
<tr>
<td>Hillary Clinton</td>
<td>728</td>
<td>@hillaryclinton, #hillaryclinton, woman @hillaryclinton, #hillaryclinton, campaign</td>
<td>keep, world, go make, take, come</td>
<td></td>
</tr>
<tr>
<td>Donald Trump</td>
<td>447</td>
<td>@realdonaldtrump, trump, #makeamericagreatagain @realdonaldtrump, trump, donald</td>
<td>want, give, take want, one, time</td>
<td></td>
</tr>
<tr>
<td>Atheism</td>
<td>588</td>
<td>religion, #god, believe #freethinker, religion, god</td>
<td>man, think, go take, make, come</td>
<td></td>
</tr>
<tr>
<td>Legalization of Abortion</td>
<td>711</td>
<td>abortion, woman, right abortion, woman, right</td>
<td>think, know, say take, carry, effect</td>
<td></td>
</tr>
<tr>
<td>Face masks</td>
<td>1,361</td>
<td>mask, wear, people wear, mask, people</td>
<td>love, look, shut care, find, care</td>
<td></td>
</tr>
<tr>
<td>Stay at home orders</td>
<td>590</td>
<td>#covid19, #coronavirus, virus #covid19, #coronavirus, virus</td>
<td>day, order, thing let, must, see</td>
<td></td>
</tr>
<tr>
<td>Fauci</td>
<td>1,102</td>
<td>#drfauci, #coronavirus, #covid19 #drfauci, #covid19, #coronavirus</td>
<td>force, work, right leave, history, work</td>
<td></td>
</tr>
<tr>
<td>School closures</td>
<td>865</td>
<td>@imbhupendrasinh, @vijayrupanibjp, school school, kid, @realdonaldtrump</td>
<td>time, do, need come, way, show</td>
<td></td>
</tr>
<tr>
<td>P-stance</td>
<td>Donald Trump</td>
<td>7,953</td>
<td>@realdonaldtrump, #donaldtrump, country @realdonaldtrump, #trump, say</td>
<td>color, head, pay arm, apply, wish</td>
</tr>
<tr>
<td>Bernie Sanders</td>
<td>6,325</td>
<td>@bernesanders, bernie, #democraticdebate @bernesanders, bernie, sander</td>
<td>check, note, ill assume, knock, sick</td>
<td></td>
</tr>
<tr>
<td>Joe Biden</td>
<td>7,296</td>
<td>#democraticdebate, @joebiden, #demdebate #democraticdebate, @joebiden, biden</td>
<td>name, sign, like dirt, tear, air</td>
<td></td>
</tr>
<tr>
<td>Marriage</td>
<td>413</td>
<td>marriage, people, couple marriage, couple, people</td>
<td>union, make, need create, become, thing</td>
<td></td>
</tr>
<tr>
<td>Vow of celibacy</td>
<td>418</td>
<td>celibacy, vow, church celibacy, vow, people</td>
<td>need, take, way nothing, way, time</td>
<td></td>
</tr>
<tr>
<td>Stay-at-home dads</td>
<td>392</td>
<td>home, dad, raise home, dad, men</td>
<td>make, provide, life time, allow, make</td>
<td></td>
</tr>
<tr>
<td>Assisted suicide</td>
<td>392</td>
<td>suicide, assist, people suicide, assist, people</td>
<td>help, take, make death, make, take</td>
<td></td>
</tr>
<tr>
<td>Fast food</td>
<td>416</td>
<td>food, eat, ban food, people, ban</td>
<td>health, make, issue world, make, time</td>
<td></td>
</tr>
<tr>
<td>Urbanization</td>
<td>404</td>
<td>area, urbanization, city urbanization, people, area</td>
<td>space, create, grow population, make, create</td>
<td></td>
</tr>
<tr>
<td>Missionary work</td>
<td>434</td>
<td>people, missionary, work work, people, missionary</td>
<td>make, take, way make, want, need</td>
<td></td>
</tr>
<tr>
<td>Libertarianism</td>
<td>381</td>
<td>libertarianism, government, people libertarianism, government, people</td>
<td>lead, give, provide take, one, work</td>
<td></td>
</tr>
<tr>
<td>Human cloning</td>
<td>416</td>
<td>clone, cloning, human cloning, clone, human</td>
<td>life, need, way make, thing, life</td>
<td></td>
</tr>
<tr>
<td>Blockade of the Gaza Strip</td>
<td>506</td>
<td>strip, gaza, blockade strip, gaza, blockade</td>
<td>stop, right, state state, get, give</td>
<td></td>
</tr>
<tr>
<td>Gender-neutral language</td>
<td>368</td>
<td>language, gender, people language, gender, people</td>
<td>offend, way, time make, feel, way</td>
<td></td>
</tr>
<tr>
<td>Compulsory voting</td>
<td>405</td>
<td>voting, compulsory, vote vote, compulsory, people</td>
<td>make, way, want take, mean, could</td>
<td></td>
</tr>
<tr>
<td>Zero-tolerance policy in schools</td>
<td>454</td>
<td>school, tolerance, student school, student, policy</td>
<td>lead, way, time way, make, time</td>
<td></td>
</tr>
<tr>
<td>Payday loans</td>
<td>442</td>
<td>loan, people, need loan, money, people</td>
<td>situation, take, need take, make, give</td>
<td></td>
</tr>
<tr>
<td>Whaling</td>
<td>423</td>
<td>whale (N), whaling, whale (V) whale (N), whale (V), whaling</td>
<td>help, way, need part, need, world</td>
<td></td>
</tr>
<tr>
<td>Capital punishment</td>
<td>467</td>
<td>punishment, capital, death capital, punishment, crime</td>
<td>justice, make, serve way, give, time</td>
<td></td>
</tr>
<tr>
<td>Cosmetic surgery for minors</td>
<td>494</td>
<td>minor, surgery, child surgery, minor, decision</td>
<td>thing, involve, give need, adult, cause</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Target</td>
<td>Sentences</td>
<td>Most different words</td>
<td>Least different words</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>-----------</td>
<td>----------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>School uniform</td>
<td>474</td>
<td>school, student, uniform school, uniform, student</td>
<td>stop, take, allow make, give, feel</td>
<td></td>
</tr>
<tr>
<td>Foster care</td>
<td>529</td>
<td>child, kid, care child, parent, care</td>
<td>may, service, find become, make, put</td>
<td></td>
</tr>
<tr>
<td>Polygamy</td>
<td>493</td>
<td>polygamy, legalize, marriage polygamy, marriage, woman</td>
<td>make, take, one way, make, time</td>
<td></td>
</tr>
<tr>
<td>Prostitution</td>
<td>499</td>
<td>prostitution, legalize, prostitute prostitution, legalise, woman</td>
<td>give, allow, want choice, involve, want</td>
<td></td>
</tr>
<tr>
<td>Zoos</td>
<td>395</td>
<td>animal, zoo, live animal, zoo, habitat</td>
<td>life, allow, make provide, keep, take</td>
<td></td>
</tr>
<tr>
<td>The right to keep and bear arms</td>
<td>407</td>
<td>keep, bear, arm keep, bear, weapon</td>
<td>law, take, remove person, must, take</td>
<td></td>
</tr>
<tr>
<td>Social media</td>
<td>330</td>
<td>medium, people, allow medium, people, allow</td>
<td>create, make, lose see, world, time</td>
<td></td>
</tr>
<tr>
<td>Multi-party system</td>
<td>390</td>
<td>system, people, multiparty party, system, government</td>
<td>bring, need, allow choose, population, thing</td>
<td></td>
</tr>
<tr>
<td>Nuclear weapons</td>
<td>542</td>
<td>weapon, country, use weapon, country, war</td>
<td>maintain, keep, life mean, make, world</td>
<td></td>
</tr>
<tr>
<td>Homeschooling</td>
<td>395</td>
<td>child, homeschooling, school child, homeschooling, education</td>
<td>give, time, keep help, teacher, way</td>
<td></td>
</tr>
<tr>
<td>Telemarketing</td>
<td>437</td>
<td>telemarketing (N), telemarketing (V), telemarketers</td>
<td>telemarketing (V), telemarketing (N), telemarketers money, work, time</td>
<td></td>
</tr>
<tr>
<td>Entrapment</td>
<td>400</td>
<td>law, crime, entrapment crime, entrapment, commit</td>
<td>get, make, allow place, time, know</td>
<td></td>
</tr>
<tr>
<td>Homeopathy</td>
<td>352</td>
<td>medicine, homeopathy, remedy homeopathy, medicine, people</td>
<td>harm, condition, placebo treat, cause, allow</td>
<td></td>
</tr>
<tr>
<td>Intelligence tests</td>
<td>462</td>
<td>intelligence, people, person test, child</td>
<td>way, base, focus show, type, know</td>
<td></td>
</tr>
<tr>
<td>Austerity regime</td>
<td>412</td>
<td>regime, austerity, economy regime, austerity, debt</td>
<td>spend, time, make reduce, pay, allow</td>
<td></td>
</tr>
<tr>
<td>Child actors</td>
<td>435</td>
<td>actor, child, use actor, child, use</td>
<td>take, show, play take, make, lead</td>
<td></td>
</tr>
<tr>
<td>Mandatory retirement</td>
<td>475</td>
<td>retirement, work, worker retirement, workforce, worker</td>
<td>make, position, force keep, provide, give</td>
<td></td>
</tr>
<tr>
<td>Sex selection</td>
<td>400</td>
<td>selection, child, parent selection, baby, sex</td>
<td>allow, could, decide bear, right, way</td>
<td></td>
</tr>
<tr>
<td>Economic sanctions</td>
<td>389</td>
<td>sanction, country, nation sanction, country, people</td>
<td>leader, make, take make, punish, help</td>
<td></td>
</tr>
<tr>
<td>Intellectual property rights</td>
<td>415</td>
<td>property, right, product property, right, people</td>
<td>come, make, time think, time, think</td>
<td></td>
</tr>
<tr>
<td>Use of public defenders</td>
<td>415</td>
<td>lawyer, defender, use defender, lawyer, defend</td>
<td>get, require, way person, mean, allow</td>
<td></td>
</tr>
<tr>
<td>Guantanamo Bay detention camp</td>
<td>444</td>
<td>guantanamo, bay, detection guantanamo, detection, camp</td>
<td>serve, way, use law, make, usa</td>
<td></td>
</tr>
<tr>
<td>Women in combat</td>
<td>370</td>
<td>combat, woman, men combat, woman, men</td>
<td>prohibit, could, make war, may, make</td>
<td></td>
</tr>
<tr>
<td>Naturopathy</td>
<td>536</td>
<td>medicine, naturopathy, treatment naturopathy, medicine, treatment</td>
<td>lead, take, life seek, allow, make</td>
<td></td>
</tr>
<tr>
<td>Church of Scientology</td>
<td>401</td>
<td>scientology, church, ban scientology, church, ban</td>
<td>member, believe, practice need, allow, practice</td>
<td></td>
</tr>
<tr>
<td>Embryonic stem cell research</td>
<td>396</td>
<td>stem, cell (N), cell (V) cell, stem, research</td>
<td>help, need, use people, need, life</td>
<td></td>
</tr>
<tr>
<td>Affirmative action</td>
<td>438</td>
<td>action, people, job action, people, discrimination</td>
<td>way, get, make school, way, work</td>
<td></td>
</tr>
<tr>
<td>Cannabis</td>
<td>543</td>
<td>cannabis, marijuana, legalize cannabis, marijuana, drug</td>
<td>take, time, way may, allow, take</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>Target</td>
<td>Sentences</td>
<td>Most different words</td>
<td>Least different words</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------</td>
<td>-----------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Vocational education</td>
<td>education, school, subsidize education, subsidize, people</td>
<td>418</td>
<td>lead, make, way work, go, give</td>
<td></td>
</tr>
<tr>
<td>Racial profiling</td>
<td>profiling, criminal, people profiling, people, crime</td>
<td>412</td>
<td>make, person, life stop, time, way</td>
<td></td>
</tr>
<tr>
<td>Private military companies</td>
<td>company, ban, government company, government, military</td>
<td>392</td>
<td>could, make, time security, need, might</td>
<td></td>
</tr>
<tr>
<td>Flag burning</td>
<td>burning, flag, burn flag, burning, burn</td>
<td>426</td>
<td>protect, freedom, make lead, protect, state</td>
<td></td>
</tr>
<tr>
<td>Surrogacy</td>
<td>surrogacy, baby, woman surrogacy, woman, surrogate</td>
<td>431</td>
<td>right, become, term give, make, could</td>
<td></td>
</tr>
<tr>
<td>Student loans</td>
<td>student, loan, education loan, student, subsidize</td>
<td>369</td>
<td>everyone, put, make afford, work, make</td>
<td></td>
</tr>
<tr>
<td>Safe spaces</td>
<td>space, people, student space, people, others</td>
<td>388</td>
<td>life, may, thing make, allow, nothing</td>
<td></td>
</tr>
<tr>
<td>Algorithmic trading</td>
<td>trading, people, market trading, computer, market</td>
<td>387</td>
<td>access, allow, base field, risk, lead</td>
<td></td>
</tr>
<tr>
<td>Olympic games</td>
<td>olympic, game, olympics olympic, game, athlete</td>
<td>409</td>
<td>money, world, time give, time, take</td>
<td></td>
</tr>
<tr>
<td>Journalism</td>
<td>journalism, news, subsidize journalism, subsidize, news</td>
<td>357</td>
<td>medium, need, could could, need, support</td>
<td></td>
</tr>
<tr>
<td>Cosmetic surgery</td>
<td>surgery, people, appearance surgery, people, ban</td>
<td>425</td>
<td>make, take, lead feel, need, way</td>
<td></td>
</tr>
<tr>
<td>Targeted killing</td>
<td>target, people, kill target, killing, people</td>
<td>409</td>
<td>use, state, take enemy, take, put</td>
<td></td>
</tr>
<tr>
<td>Organ trade</td>
<td>trade, organ, sell trade, organ, legalize</td>
<td>408</td>
<td>give, death, way need, create, help</td>
<td></td>
</tr>
<tr>
<td>Space exploration</td>
<td>space, exploration, subsidize space, exploration, planet</td>
<td>381</td>
<td>thing, support, country thing, find, use</td>
<td></td>
</tr>
<tr>
<td>Factory farming</td>
<td>farm, factory, food factory, food farming</td>
<td>410</td>
<td>space, allow, keep produce, keep, allow</td>
<td></td>
</tr>
<tr>
<td>Pride parades</td>
<td>parade, pride, gay parade, pride, lgbt</td>
<td>394</td>
<td>right, allow, make way, want, bring</td>
<td></td>
</tr>
<tr>
<td>Collectivism</td>
<td>collectivism, group, people collectivism, people, society</td>
<td>440</td>
<td>need, one, way take, lead, way</td>
<td></td>
</tr>
<tr>
<td>Television</td>
<td>television, people, watch television, news, entertainment</td>
<td>387</td>
<td>way, thing, keep could, way, make</td>
<td></td>
</tr>
<tr>
<td>School prayer</td>
<td>school, prayer, religion prayer, school, religion</td>
<td>424</td>
<td>allow, take, person part, time, place</td>
<td></td>
</tr>
<tr>
<td>Autonomous cars</td>
<td>car, road, drive car, road, drive</td>
<td>445</td>
<td>cause, way, need take, use, time</td>
<td></td>
</tr>
<tr>
<td>Holocaust denial</td>
<td>holocaust, denial, deny holocaust, denial, deny</td>
<td>456</td>
<td>speech, allow, go allow, world, say</td>
<td></td>
</tr>
<tr>
<td>Executive compensation</td>
<td>executive, compensation, company executive, company, compensation</td>
<td>375</td>
<td>give, deserve, lead level, work, allow</td>
<td></td>
</tr>
<tr>
<td>Three-strikes laws</td>
<td>law, strike, crime law, strike, people</td>
<td>490</td>
<td>take, make, need give, put, allow</td>
<td></td>
</tr>
<tr>
<td>Atheism</td>
<td>atheism, god, religion atheism, religion, people</td>
<td>360</td>
<td>base, allow, make way, provide, lead</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>wikipedia, subsidize, information wikipedia, wikipedia, subsidize</td>
<td>395</td>
<td>could, need, take provide, way, give</td>
<td></td>
</tr>
<tr>
<td>Judicial activism</td>
<td>judge, law, activism judge, activism, law</td>
<td>385</td>
<td>use, need, way allow, rule, could</td>
<td></td>
</tr>
</tbody>
</table>

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.”