Aligning Multidimensional Worldviews and Discovering Ideological Differences

Jeremiah Milbauer
Language Technologies Institute,
Carnegie Mellon University
jmilbaue@cs.cmu.edu

Adarsh Mathew
Knowledge Lab,
University of Chicago
adarshm@uchicago.edu

James Evans
Knowledge Lab / Sociology,
University of Chicago
jevans@uchicago.edu

Abstract

The Internet is home to thousands of communities, each with their own unique worldview and associated ideological differences. With new communities constantly emerging and serving as ideological birthplaces, battlegrounds, and bunkers, it is critical to develop a framework for understanding worldviews and ideological distinction. Most existing work, however, takes a predetermined view based on political polarization: the "right vs. left" dichotomy of U.S. politics. In reality, both political polarization – and worldviews more broadly – transcend one-dimensional difference, and deserve a more complete analysis. Extending the ability of word embedding models to capture the semantic and cultural characteristics of their training corpora, we propose a novel method for discovering the multifaceted ideological and worldview characteristics of communities. Using over 1B comments collected from the largest communities on Reddit.com representing 40% of Reddit activity, we demonstrate the efficacy of this approach to uncover complex ideological differences across multiple axes of polarization.

1 Introduction and Motivation

> "The limits of my language mean the limits of my world"
>
> *Tractatus Logico-Philosophicus*, 1921, Ludwig Wittgenstein

Media choice, social networking platforms, and collaborative filtering on the internet have enabled individuals to enter "echo chambers" that reflect shared worldviews (Sunstein, 2018; Mutz, 2006; Bishop, 2009). The internet also publicly reveals these communities and their communication for analysts of language, culture and interaction at unprecedented scale. Despite the abundance of such data, however, analysis of worldviews and ideological difference has been dominated by considerations of "polarization" (Boxell et al., 2017; Bail et al., 2018), which impoverishes the comparison of ideologies by reducing them to pairs separated along a singular dimension.

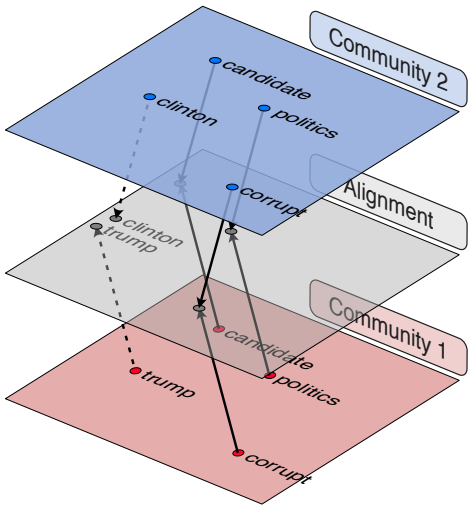

Figure 1: By training the model to align "candidate", "politics", and "corrupt," a hypothesis alignment $f(\text{"trump"}|C_1) \approx f(\text{"clinton"}|C_2)$ emerges.

Here, we draw inspiration from the approach of interpretive anthropology and the focus of cognitive anthropology to represent, investigate and compare worldviews from community discourse. In the *Interpretation of Cultures*, Geertz rendered culture as "a system of inherited conceptions expressed in symbolic forms by means of which men communicate, perpetuate, and develop their knowledge about and attitudes toward life" (Geertz et al., 1973). Combined with cognitive anthropology's concern with how implicit knowledge changes the way people perceive and relate to the world (d'Andrade, 1995), this motivates assessment of worldviews through modern pre-trained natural language models that render words (Mikolov et al., 2013b; Pennington et al., 2014) and phrases (Devlin et al., 2019; Radford et al., 2019) in relation to one another as a function of their proximity in discourse. When pre-trained on the discourse of distinctive communities, these models have begun

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4832–4845
November 7–11, 2021.

to enable a highly resolved evaluation of expressed worldviews – symbol systems that reveal shared patterns of attention and association (Kang and Evans, 2020).

Based in the premise that the language a community uses carries markers of the culture of that community (Webson et al., 2020), recent work has demonstrated the ability of trained embedding models to uncover cultural values (Garg et al., 2018; Xie et al., 2019; Kozlowski et al., 2019). However, these models are limited by requiring significant researcher input to query the model for insights. Recent work has also demonstrated the potential to embed communities themselves Waller and Anderson (2021), but has not extended to the level of a word-level understanding of community worldview.

We instead model community language as a specific instance of an ideological dialect (or, an "ideolect")[1]. Using a similar approach to KhudaBukhsh et al. (2021), which identified single-axis polarized political "languages" on YouTube, we introduce a new method for unsupervised cultural analysis based on multilingual embedding alignment. Our method provides high-accuracy alignment, is the first to analyze multiple facets of ideological polarization, and readily enables analysis in a large multi-community setting – which we demonstrate by identifying multiple axes of ideological differences on Reddit.

As an additional contribution, we publish a Github repository with all the code necessary to replicate this work and apply our methods in new settings.[2] This repository also includes tables of results that were too long to reasonably include in this paper.

2 Unsupervised Cultural Analysis

In this section, we summarize previous approaches to the analysis of cultural values through word embedding models.

2.1 The Queried Approach

Early work on cultural analysis through word embeddings observed that the cultural values of a community or society are embedded within the text produced by that community or society, and are discoverable by word embedding (Garg et al., 2018). These values can then be queried by measuring the distance between wordpairs.

Measuring stereotypes By pre-selecting a set of "entity" words, and "value" words, researchers can measure how attitudes towards the selected entities differ over time, or across communities. This approach works by training an embedding model on a text corpus, and then computing the similarity between each query word and each value word. Garg et al. (2018) use this approach, with occupations comprising the entities and gender or ethnic categories as the values. Each pair thus represents the strength of a particular cultural value or stereotype. However, this approach is limited in that it is only able to discover the specific stereotypes queried by the researchers; it is unable to discover cultural values on its own.

Axes of polarization Another approach to unsupervised cultural analysis is introduced by Kozlowski et al. (2019). In this method, two words representing the opposite poles of a particular cultural value (such as "rich" and "poor") are selected. Entity words, such as the names of different sports, are then projected to an axis drawn between the polar words.

Although such models have the capacity to render worldviews as high dimensional spaces, research typically compares representations only selectively in terms of a modest set of keywords queried and compared between models. In these cases, the keywords are typically manually selected according to a predetermined notion of which words may exhibit polarization, and compared with words that are pre-selected to encode cultural values, essentially producing a cultural relatedness score for a given (Entity,Value) pair in some corpus:

$$\text{Entity}, \text{Value} \rightarrow \text{Score}_C$$

This method can then be used to identify differences between corpora:

$$\text{Entity}, \text{Value} \rightarrow \text{Score}_{C_1} - \text{Score}_{C_2}$$

[1] This is not to be confused with the linguist's notion of an "idiolect", language quirks unique to a person but understood by others, or Wittgenstein's notion of a language uniquely understandable by a single person. Our notion of "ideolect" draws on both: a language shared by an ideological group, which necessarily contains the private worldview of that group and may not be naively decipherable to those outside.

[2] https://github.com/jmilbauer/worldview-ideology

2.2 Toward less supervision

Xie et al. (2019) make progress on this issue by introducing the use of the Moral Foundations Dictionary (Graham et al., 2009) to approximate moral categories. Using a trained embedding model, they assign each word to its nearest cluster of Moral Foundations Words, and measure differences in terms of a word's movement between clusters across distinct corpora. This approach has two key advantages: it does not presuppose the relevant cultural values (the Moral Foundations Dictionary is designed to be comprehensive), and it allows the moral categories to be specific to each corpus's embedding.

With this approach, we are now able to evaluate the relevance of each word to the moral differences between communities C_1, C_2, as:

$$C_1, C_2 \rightarrow \{\text{MoralDifference(w)} | \forall w \in V\}$$

This set of scored words thus represents the cultural differences between two communities. However, it too is limited in expressivity by the reliance on the Moral Foundations Dictionary's list of moral categories.

2.3 Aligning Ideological Dialects

Rather than rely on the previous query-value paradigm, we achieve fully unsupervised cultural analysis through the use of multilingual embedding alignment. We explicitly model corpus-specific ideological dialects using techniques designed for multilingual word embedding alignment, to learn a translation function $\mathcal{F}$ from each embedding to the joint space. Then, for any two corpora, each word has an alignment score:

$$\text{AlignmentScore(w)} = d(\mathcal{F}(w|C_1), \mathcal{F}(w|C_2))$$

This ultimately yields a similar set of scores to the Moral-Foundations approach:

$$C_1, C_2 \rightarrow \{\text{AlignmentScore(w)} | \forall w \in V\}$$

with two important benefits: our model requires no moral supervision, and can discover more than just moral differences. By contrasting semantic models *per se*, we automatically discover ideological differences in a multi-community corpus.

Additionally, for a given word w_1 in C_1, we can compute a the nearest image w_2 in C_2, such that

$$w_2 = \arg\min_w d(\mathcal{F}(w_1|C_1), \mathcal{F}(w|C_2))$$

This represents the hypothesis of a conceptual substitution between communities, yielding a high-resolution comparison of the worldviews, ideologies, and cultural differences between two communities, without any supervision. Figure 1 illustrates this idea in the context of a conservative political community (bottom panel, in red) and a liberal one (top panel, in blue). Worldviews are seen anchored by the words "corrupt", "politics", and "candidate", and an alignment between the semantics of "clinton" and "trump" emerges.

3 Data

Reddit serves as the primary source of data for this project. The platform is structured as a collection of peer-driven communities called "subreddits," ostensibly self-regulated by norms decided upon by members of the subreddit and enforced by moderators. All users are anonymous, can be a part of multiple subreddits, and are free to create their own. As such, user comments serve as a rich source of conversation and discourse across varied interests and topics, organized into communities of self-selected individuals.

The structure of Reddit lends itself to a community-focused analysis of language, with the site's use of self-enforced boundaries allowing us to observe discourse across groups without having to define the notion of a group ourselves. Instead, we rely on every user's own choice about where they wish to engage, and where to post their comments. This multi-community setting has been exploited in the past by researchers, with Tan and Lee (2015) exploring the contours of multi-community engagement and the widening of interests via a user's exploration of different subreddits over time. Rajadesingan et al. (2020) explore the norms of interaction dictated and enforced by multiple "toxic" subreddits, showcasing how self-selection and pre-entry learning play a key role in sustaining these norms. Kumar et al. (2018) explicitly study negative mobilizations between different subreddits as conflict, finding that they tend to occur between communities that are highly similar in content.

We use data from Reddit for the period 2016-2019, and select 32 subreddits from the largest communities to study, representing between 30 and 40% of the site's monthly activity. We rely on the Reddit dumps ingested by Pushshift as described in Baumgartner et al. (2020), which we accessed in January of 2020. These dumps contain comment

Year	Comments	Tokens	GB
2016	242.65 M	7032.87 M	34.14
2017	256.76 M	7429.59 M	36.08
2018	266.12 M	7707.22 M	37.37
2019	288.36 M	7977.69 M	38.84

Table 1: Number of comments, tokens, and gigabytes in the dataset

activity across all of Reddit for each month. Given the delay in ingesting activity across all subreddits, some comments and users can be deleted before ingestion occurs. Additionally, users are given the opportunity to have their data not ingested by submitting an opt-out request. Although the Pushshift dataset includes the users' usernames, we scrub all information aside from the actual text of the post before even any pre-processing occurs. When discussing a specific community, we refer to it as "r/[community name]," as is customary on Reddit.

Table 1 contains information about the size of our dataset after preprocessing.

4 Modeling and Aligning Ideological Dialects

We conceive of the alignment procedure as a matching of "conceptual anchors," designed to align the worldview of two communities. If two communities, $\mathcal{C}_a$ and $\mathcal{C}_b$, have identical worldviews, we expect that structural relations between words will be preserved across the community boundary. However, if they have *different* worldviews, we would expect that the words central to that conflict would *not* align well, even when anchoring words are well-aligned.

On notation For a community called a, we typically use $\mathcal{C}_a$ to indicate the "language" of the community, V_a for its vocabulary, A to represent an embedding matrix trained on $\mathcal{C}_a$, and A_i to represent the word embedding of a word w_i in C_a.

4.1 Foundations

In order to align and compare community-specific models, we turn to the literature on multilingual word embeddings. Broadly speaking, these works aim to learn a single embedding space in which synonymous words in different languages have the same embedding. Approaches to this problem vary, but typically either rely on training with parallel corpora in multiple languages, or aligning embeddings with the help of a multilingual lexicon. In our case, all data collected from different Reddit communities is in English – so we automatically have a complete parallel lexicon. As such, we choose to use the lexicon approach to align our different "ideolects." Furthermore, this approach allows our work to be immediately useful to computational social scientists currently using out-of-the-box word embedding algorithms for cultural analysis.

Most common is the bilingual case; given two languages, $\mathcal{L}_a$ and $\mathcal{L}_b$, we use a bilingual lexicon to learn two transformation functions: $f_{a\to c}$ and $f_{b\to c}$, such that for every word $i \in \mathcal{L}_a$ and every word $j \in \mathcal{L}_b$, $f_{a\to c}(\text{Emb}(i)) = f_{b\to c}(\text{Emb}(j))$. In this bilingual case, it is possible to set c to b, and essentially learn a single transformation from one space to the other. In the multilingual case, one can learn a latent space into which all languages are projected, chose one language as the target for all the other languages' transformations, or learn direct pairwise bilingual transformations.

In this work, we adapt approaches (Ammar et al., 2016; Mikolov et al., 2013a) developed for multilingual alignment to the cultural analysis use-case.

After aligning worldviews with this approach, we then evaluate multiple dimensions of ideological difference by computing misalignment scores across different topics.

4.2 Pre-processing

We treat the posted comments of each community, in each year, as its own corpus. For each community-corpus in the dataset, we tokenize each of the comments posted in the community (without stemming or lemmatization), remove formatting tokens, reduce hyperlinks to just their surface forms, and make all characters lower case. We then run a basic phrase detection algorithm (Mikolov et al., 2013b), implemented in Gensim (Rehurek and Sojka, 2011), to detect common bigrams in each community.

4.3 Training Word Embeddings

We begin by training a word embedding model for each independent community. Here, we use Gensim's implementation of the Skip-gram model (Mikolov et al., 2013b). We train embeddings in both 100 and 300 dimensions; the following experiments were conducted with 100-dimensional embeddings. Given that the data is from a long-tailed forum community on the Internet, we use a maximum vocabulary of 30,000 words. In order to promote the stability of the embedding for

each community-corpus, we over-sample sentences from smaller communities.

4.4 Anchors

In order to train an alignment between two embedding spaces, we must first construct a "bilingual" lexicon to anchor the alignment. All text in our corpora a in English, so we can easily construct an lexicon of size $N = |V|$, using the entire shared vocabulary of two trained embeddings as the anchoring words. However, it should be noted that the goal of our embedding alignment should *not* be maximum accuracy. We intend to use the trained alignment as a tool for cultural analysis by exploring the misaligned words; so we should not attempt to achieve a perfect map.

We experiment with three distinct approaches to construct the bilingual lexicon. The first approach uses the entire shared vocabulary to anchor the alignment. The second approach uses a large set of stopwords – the most frequent 5000 words across the combined corpora. The third approach uses a smaller set of stopwords – 1000.

4.5 Topic Modeling

In order to identify topic areas within which to measure misalignment, we implement a topic assignment procedure, inspired by the success of a simple embedding-based approach for Twitter data in Demszky et al. (2019). We learn word clusters using an embedding model trained on the union of the communities, with the scikit-learn (Pedregosa et al., 2011) implementation of KMeans++ (Arthur and Vassilvitskii, 2006). We then treat each word cluster as a topic.

To validate these topics, we compute the core topics for each community by assigning each comment a topic label, and calculating the association of each topic with each community. For each topic t and community $\mathcal{C}$:

$$\text{Score}(t, \mathcal{C}) = \frac{P(t, \mathcal{C})}{P(t)P(\mathcal{C})} = \frac{P(t|\mathcal{C})}{P(t)}$$

Using these scores, we rank the topics of each community. Table 2 includes examples of some top topics for r/gaming, r/politics, and r/askmen – popular groups that discuss gaming, politics, and mens' issues respectively.

Figure 4 in the Appendix includes a full comparison of topics similarities across communities. An interesting observation from this validation is that communities for which we hypothesize a strong ideological disagreement (such as r/politics and r/the_donald), there is a strong similarity in topic distribution.

r/gaming	Rank 1	doom, halo, zelda, ...
	Rank 2	os, hulu, apple, ...
	Rank 3	graphics, 480, nvidia, ...
r/politics	Rank 3	states, president, ...
	Rank 4	fdr, communist, ...
	Rank 5	sent, emails, scandals, ...
r/askmen	Rank 1	puberty, sex, tinder, ...
	Rank 3	younger, minded, ...
	Rank 4	old, lover, spouse, ...

Table 2: Randomly sampled words from top topics for a small selection of subreddits. Topics consisting primarily of administrative and moderation messages are omitted.

4.6 Alignment

Once anchoring words have been selected (either by using all words, stop words, or non-salient topic words), we can train an alignment between embedding spaces. We choose to treat alignment as a linear transformation, $\mathcal{T}_{a\to b} \in \mathbb{R}^{d\times d}$ from one d-dimensional vector space A to another, B, so $A \cdot \mathcal{T}_{a\to b} = B$. This allows the learned transformation to be both **compositional** and **invertible**:

$$A \cdot \mathcal{T}_{a\to b} \cdot \mathcal{T}_{b\to c} = C$$

$$B \cdot \mathcal{T}^{-1}_{a\to b} = A$$

These properties are important when creating multilingual embeddings, especially for low-resource languages. When there is no bilingual lexicon for a pair $\{\mathcal{L}_a, \to \mathcal{L}_b\}$, we can still learn transformations between them by passing through a high-resource $\{\mathcal{L}_c$ like English:

$$A \cdot \mathcal{T}_{a\to c} \cdot \mathcal{T}_{c\to b} = B$$

In our case, because the linear transformation is an isomorphism, we also think of our work as an extension of the idea of analogies in Mikolov et al. (2013b), but at the community level.

This compositionality also allows us to reduce the number of alignments to train, which is useful when performing experiments at scale. For a set of N communities, describing the entire set requires N^2 alignments. By relying on compositionality, we need only train N transformations: one for each community and the high-resource community. In our dataset, r/AskReddit is the highest resource

community, and thus the most appropriate analog to English in the multilingual setting.

We consider three techniques for alignment: **MultiCCA**, developed by Ammar et al. (2016), a linear equation solver, and an SVD-based approach described in Smith et al. (2017). We experiment using each of these approaches to select linear projections, different anchoring set size. In each case, we begin with two word embedding models trained on different community corpora, $\mathcal{C}_a$ and $\mathcal{C}_b$, each with their own vocabulary V_a and V_b. We then construct the set of potential anchoring words, $D^{a,b} \subset V^{a,b}$, where $V^{a,b} = V_a \cap V_b$. Our first anchoring strategy uses all words in $D^{a,b}$, our second strategy uses only the 1000 most frequent words ($D^{a,b}_{1000}$), and our third strategy uses the 5000 most frequent words ($D^{a,b}_{5000}$). We then construct two training matrices: A' and B', where $A'_i = \text{Emb}_a(D_i)$ and $B'_i = \text{Emb}_b(D_i)$. We then train the alignment using A' and B', then evaluate.

Mode	N	Acc@1	Acc@5	Acc@10
ALL	∼30k	0.6937	0.8335	**0.8709**
SW	1000	0.6583	0.8037	0.8450
	5000	0.6934	0.8306	0.8679

Table 3: Performance of Least-squares alignments for the year 2016, measuring alignment to the shared space.

Mode	N	Acc@1	Acc@5	Acc@10
ALL	∼30k	0.7284	0.8591	**0.8924**
SW	1000	0.6234	0.7735	0.8174
	5000	0.6984	0.8352	0.8718

Table 4: Performance of MultiCCA alignments for the year 2016, measuring alignment to the shared space.

Mode	N	Acc@1	Acc@5	Acc@10
ALL	∼30k	0.4980	0.6462	0.7091
SW	1000	0.4798	0.6434	0.6991
	5000	0.5203	0.6841	**0.7374**

Table 5: Performance of SVD alignments for the year 2016, measuring pairwise alignment.

MultiCCA For communities $\mathcal{C}_a$ and $\mathcal{C}_b$, **MultiCCA** seeks to learn two projections to latent space $\mathcal{C}$: $\mathcal{T}_{a\to c}$ and $\mathcal{T}_{b\to c}$, in order to maximize the correlation of $A \cdot \mathcal{T}_{a\to c}$ and $B \cdot \mathcal{T}_{b\to c}$. From these projections, we then recover the projection of interest $\mathcal{T}_{a\to b}$:

$$\mathcal{T}_{a\to b} = \mathcal{T}_{a\to c} \cdot \mathcal{T}^{-1}_{b\to c}$$

We implement this approach using scikit-learn's `cross_decomposition.CCA` module. (Pedregosa et al., 2011)

Linear Equation Solver For A and B, the linear equation solver aims to learn $\mathcal{T}_{a\to b}$ by solving the equation: $A \cdot \mathcal{T}_{a\to b} = B$. We use NumPy's Least-squares linear equation solver, `linalg.lstsq`. (Harris et al., 2020)

Singular Value Decomposition This method is employed by KhudaBukhsh et al. (2021) (albeit with many fewer anchoring words), and is described in Smith et al. (2017). Alignment is trained directly between community pairs, rather than between each community and a shared space. For this method, the projection is learned by solving $U\Sigma V^T = A^T B$, setting $\mathcal{T}_{a\to b} = UV^T$. We use NumPy's `linalg.svd`. (Harris et al., 2020)

Evaluation For each pair of communities C_a and C_b, and each word $w_i \in V_a, V_b$, we translate w_i from C_a to C_b. Each w_i has an embedding A_i learned from C_a, an embedding B_i learned from C_b, and an image B'_i under alignment, where $B'_i = A_i\mathcal{T}_{a\to b}$. We then find the N nearest-neighbors of B'_i in V_b, using cosine similarity. Acc@N is the proportion of N-nearest-neighbor sets that contain w_i. Tables 3, 4, and 5 contain the results of this evaluation for the year 2016, macro-averaged over each projection learned. Other years are included in the appendix.

Discussion As might have been anticipated, the anchoring method that uses all available words is the most accurate. We also notice a trend of decreasing accuracy from 2016 to 2019, despite the increase in dataset size and therefore embedding stability. This suggests growing semantic differences between Reddit communities over time. For future experiments and evaluation, we use the 5000-anchor **MultiCCA** approach, which we found to empirically provide alignment accuracy without exposing the model to all of the data.

4.7 Comparison with Previous Methods

Unsupervised cultural analysis of this kind is an extremely recent development in the literature. However, previous methods can be adapted to provide a baseline for comparison. For the following comparisons, we select for analysis two communities with both a high degree of moral polarization, and a known axis of polarization: r/politics and

r/the_donald. These communities are highly politically polarized. We perform the comparison with data from the year 2017.

We initially perform a comparison with an approach described by Xie et al. (2019), which identifies changes in moral semantics across corpora. We use the technique to generate a set of misaligned words by identifying words that move from a positive to a negative moral category (and vice versa) between communities. We then rank the words by degree of movement. This method retrieves political words (defined as words falling into political topic clusters) with a **0.2247** MAP.

For both our method and the method described by KhudaBukhsh et al. (2021), we follow the procedure for anchoring and training an alignment. For KhudaBukhsh et al. (2021), this means using SVD with NLTK stopwords (Bird et al., 2009). We then sort the misaligned wordpairs by degree of alignment, and classify a wordpair as political if either of the misaligned words is in one of the political clusters. KhudaBukhsh et al. (2021) achieves **0.3076** MAP; our method achieves **0.3318** MAP.

5 Exploring Worldview and Ideology

In this section, we use our method to perform a number of sociolinguistic explorations.

5.1 Worldview Misalignment

We begin by using the learned projection/alignment to identify "misaligned" words in a political context.

We say that a word is "aligned" when the nearest image of a word w_i from C_1 is itself:

$$i = \arg\min_j d(A_i \cdot \mathcal{T}_{a\to b}, B_j)$$

And "misaligned" when it is not:

$$i \neq \arg\min_j d(A_i \cdot \mathcal{T}_{a\to b}, B_j)$$

We anticipate the words that will ultimately misalign are either words with low quality embeddings (owing to low frequency in the corpus) or words with very polarized meanings across communities.

Our first experiment, analyzing two politically misaligned corpora, is a typical area of inquiry. (KhudaBukhsh et al., 2021; Xie et al., 2019; Webson et al., 2020) We select r/politics ($\mathcal{C}_a$), a general-purpose political discussion board with a strong liberal tendency, and r/the_donald ($\mathcal{C}_b$), a Trump-supporting and aggressively conservative community well known as a breeding grounds for conspiracy theories, including PizzaGate (Kang, 2016). We begin by finding the vocabulary of shared words between r/politics and r/the_donald, and use our alignment algorithm to "translate" each word from r/politics to r/the_donald. Using **MultiCCA**, and r/askreddit ($\mathcal{C}_c$) as the "high-resource" language, the translation is formulated as:

$$\mathcal{T}_{a\to x} \cdot \mathcal{T}^{-1}_{\text{shared}\to x} \cdot \mathcal{T}_{\text{shared}\to y} \cdot \mathcal{T}^{-1}_{b\to y}$$

Using this matrix transformation, we project all shared words from $\mathcal{C}_a$ to $\mathcal{C}_b$. We also repeat this process in reverse.

Querying this model for political words, we find a number of interesting misalignments, including the words which directly define the known axis of polarization: "democrat" and "republican." Table 6 contains a sample of misalignments from r/politics to r/the_donald. This demonstrates the ability of our method to identify the nature of polarization between two communities without any presuppositions about the communities.

5.2 Conceptual Reflections

While the approach described in section 5.1 is able to identify misaligned words and "translate" across the cultural boundary, we also consider another procedure: using the trained embedding alignments to identify the antonyms that describe an axis of semantic reflection between two communities. We use a predetermined set of antonym pairs from Miller (1995), and identify all instances where a word w in $\mathcal{C}_a$ maps to its antonym in $\mathcal{C}_b$.

We apply this approach to the community pair of r/askwomen and r/askmen, forums that discuss womens' and mens' issues, respectively. Table 7 contains top identified antonyms pairs.

Although the list is not exhaustive, we see that the antonym approach quickly identifies the gender axis between the two communities. A weakness of this approach is that many words, such as names and other proper nouns, may not be included in a predetermined set of antonyms.

5.3 Conceptual Homomorphism

There may exist two distinct communities of speakers that have similar worldviews and conceptual structures, but do not talk about the same things. A good example of this are the two communities

r/politics	r/the_donald	Alignment
democrat	republican	0.8562
republican	democrat	0.8501
leftwing	rightwing	0.8307
socialized_medicine	universal_healthcare	0.8041
magas	libtards	0.6578

r/the_donald	r/politics	Alignment
republican	democrat	0.8570
democrat	republican	0.8527
prolife	prochoice	0.8435
foxnews	cnn	0.7960
pocahontas	elizabeth_warren	0.6694

Table 6: Selected words from r/politics and their nearest image under alignment in r/the_donald (left); selected words from r/the_donald and their nearest image under alignment in r/politics (right). Degree of alignment measured in cosine similarity.

r/askwomen	r/askmen	Alignment
son	daughter	0.7675
daughter	son	0.7621
husband	wife	0.7503
father	mother	0.7445
brother	sister	0.7145
girlfriend	boyfriend	0.7032
wife	husband	0.6941
boyfriend	girlfriend	0.6708
uncle	aunt	0.6314

Table 7: Words in r/askwomen that align to their antonym when projected to r/askmen. Degree of alignment measured in cosine similarity.

r/dota2 and r/leagueoflegends. Both of these communities are discussion boards centered around a "MOBA" (Multiplayer Online Battle Arena) video game, and both video games share a great deal of similarity. However, r/dota2 players and r/leagueoflegends players often see each other as rivals or enemies. By using our alignment technique, we demonstrate a use-case for bridging the conceptual gap between two similar communities and finding conceptual homomorphisms.

By aligning the embeddings of two communities $\mathcal{C}_a$ and $\mathcal{C}_b$, we can project words that are in $\mathcal{V}_a$, but not $\mathcal{V}_b$, from A to B, learning a semantic representation for an out-of-vocabulary word unknown to $\mathcal{C}_b$. This projection yields C_b's equivalent of C_a's unique word. This is similar to unsupervised translation.

We then use the projection learned between r/leagueoflegends and r/dota2 to estimate the nearest word within the r/dota2 space for a small set of query words unique to r/leagueoflegends. Table 8 contains some examples of the projections, and Figure 2 provides an additional illustration of the success of the technique in identifying cross-community semantic analogs.[3]

[3]For readers unfamiliar with the games *League of Legends* or *Dota 2*; "/r/summonerschool" is a community for learning to play *League of Legends*; "opgg" is a website used for tracking stats in *League of Legends*, and "dotabuff" is used by *Dota2* players; "Riot Games" and "Valve" are the creators of *League of Legends* and *Dota 2* respectively; "rito" and "volvo" are both joking nicknames for the respective game creators; "Aatrox" and "Bloodseeker" are both blood-themed fighters.

r/LeagueOfLegends	r/Dota2	Alignment
/r/summonerschool	/r/learndota2	0.8420
op.gg	dotabuff	0.8396
rito	volvo	0.8378
riot	valve	0.8003
aatrox	bloodseeker	0.6473

Table 8: Selected words from r/leagueoflegends and their nearest image in r/dota2. Alignment is measured in cosine similarity.

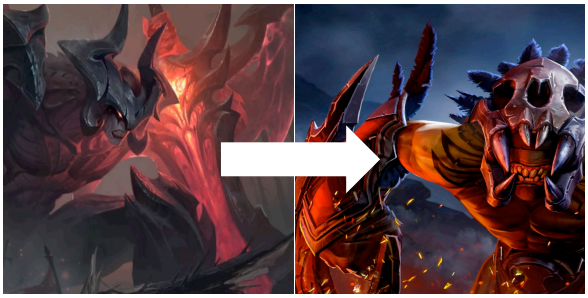

Figure 2: These are not the same! The character on the left, "Aatrox" from *League of Legends*, projects to the character on the right, "Bloodseeker" from *Dota 2*.

5.4 Large-scale Analysis

Finally, we perform a large-scale analysis across all top Reddit communities. Using the topic clusters described in section 4.5, we compute the number of misalignments for each topic cluster.

We are then able to produce pairwise misalignment scores for each pair of communities with respect to each topic cluster, uncovering the multidimensional ideological misalignment across Reddit. These comparisons are numerous; we include two here. Figure 3 demonstrates the degree of misalignment with respect to two political subcategories, corresponding to "Economics" and "Authority".

Despite low KL-divergence in topic distributions for political communities, as shown in Figure 4, they demonstrate strong misalignment on the "Economics" topic. The difference demonstrates our method's ability to resolve specific types of polarization across specific ideological categories, as opposed to previous work that treats political polarization as a single-dimensional problem. Additional topic misalignments are included in 5.

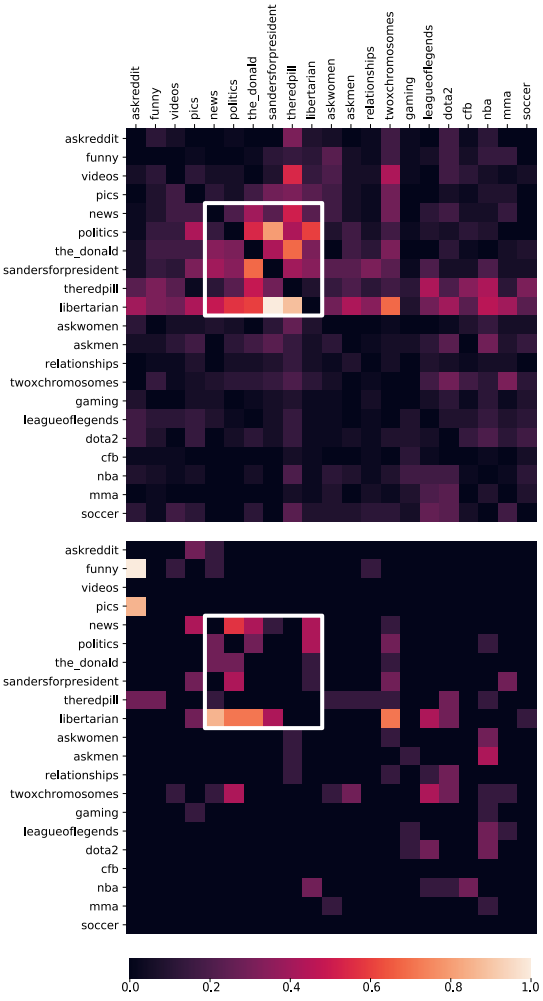

Figure 3: Misalignment frequency within the "Economics" cluster (top), and the "Authority" cluster (bottom). Color corresponds to the relative intensity of misalignment, and the white squares outline political communities.

While significant, an analysis of Reddit communities is only a fraction of what this approach is capable of. Unlike previous methods that rely on calculating all pairwise alignments, the compositional nature of the **MutliCCA** approach we propose only requires learning the alignment between each community's ideological dialect and a central high-resource community. As such, the training time scales linearly with the number of communities analyzed, which makes the study of the potentially large number of ideological communities much more tractable.

6 Conclusion

In this paper we have demonstrated a novel technique for unsupervised cultural analysis by building upon existing work treating word embeddings as tools to explore worldview, as well as work on multilingual embedding alignment. We have shown that our formulation is flexible, and able to operate effectively in a complex multi-community setting.

We have also demonstrated a number of useful applications of the worldview discovery procedure, from the automatic identification of axes of polarization, to the identification of out-of-vocabulary words with similar semantics, to the large-scale analysis of an online social community with multiple dimensions of ideological polarization.

6.1 Future Directions

A key application of this method is in unsupervised cultural analysis, which would allow researchers to explore culture at scale, without using a manual value-querying process that imputes their own beliefs and values into the process. Such advancements may also enable more sophisticated explorations of Internet conflict. With a high-dimensional estimate of ideology for a user and their body of comments, research on Internet conflict can extend beyond high-temperature "confrontation" alone. This would enable analysts to identify and respect "legitimate" conflict—conflict that emerges not from trolling or a clash of moods and personalities (Cheng et al., 2017), but a clash of underlying worldviews.

We believe our method also extends well to the study of academia itself, i.e. the science of science. An unsupervised method to identify terms that translate well into adjacent scientific fields/approaches would make cross- and interdisciplinary studies easier, providing a ready lexicon of ideas which best relate to what you already know. It could also allow us to examine how ideas fare when they are imported into fields adjacent or distant to their point of origin. Even more broadly, our approach could be used to generalize search that takes into account different perspectives on—and different phrasings for—similar underlying concepts and issues.

7 Broader Impacts and Ethical Considerations

We recognize the significant impact that modern natural language processing technology can have on society, and the potential for its abuse. This paper lays the groundwork for a large-scale unsupervised approach to the analysis of culture, which could ultimately lead to technologies capable of effectively forecasting conflict and radicalization in online speech. In the wrong hands, that might inspire information operations that could have a chilling effect on online speech.

But we are optimistic about the future of this approach to cultural (mis)alignment. As demonstrated, it can be used to identify not only disagreement, but where there is undiscovered potential for agreement. We began this paper with a quote: "The limits of my language are the limits of my world." We hope that by building on this technique to reveal both similarities and differences in community worldviews, we can someday expand the limits of everyone's worldview by facilitating mutual understanding, finding ways to resolve ideological tension, and make new knowledge easier to transmit and receive.

Acknowledgements

We would like to thank members of Knowledge Lab, University of Chicago for helpful comments and suggestions. This work is funded in part by DARPA via grants RA-19-01 and HR00111820006 and AFOSR via grant FA9550-19-1-0354.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings.

David Arthur and Sergei Vassilvitskii. 2006. k-means++: The advantages of careful seeding. Technical report, Stanford.

Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bill Bishop. 2009. *The Big Sort: Why the Clustering of Like-minded America is Tearing Us Apart*. Houghton Mifflin Harcourt.

Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. 2017. Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, 114(40):10612–10617.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1217–1230. Association for Computing Machinery.

Roy G d'Andrade. 1995. *The development of cognitive anthropology*. Cambridge University Press.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. 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 2970–3005, 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 *NAACL-HLT (1)*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Clifford Geertz, ifford Geertz, and Basic Books. 1973. *The Interpretation Of Cultures*. Basic Books.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Charles R. Harris, K. Jarrod Millman, St'efan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'ıo, Mark Wiebe, Pearu Peterson, Pierre G'erard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Cecilia Kang. 2016. Fake news onslaught targets pizzeria as nest of child-trafficking. *The New York Times*.

Donghyun Kang and James Evans. 2020. Against method: Exploding the boundary between qualitative and quantitative studies of science. *Quantitative Science Studies*, 1(3):930–944.

Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom Mitchell. 2021. We Don't Speak the Same Language: Interpreting Polarization through Machine Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14893–14901.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 933–943. International World Wide Web Conferences Steering Committee.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119. Curran Associates Inc.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Diana C Mutz. 2006. *Hearing the Other Side: Deliberative Versus Participatory Democracy*. Cambridge University Press.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Samuel L Smith, DHP Turban, S Hamblin, and NY Hammerla. 2017. Offline bilingualword vectors. *Orthogonal Transformations*.

Cass R Sunstein. 2018. *# Republic: Divided democracy in the age of social media*. Princeton University Press.

Chenhao Tan and Lillian Lee. 2015. All Who Wander: On the Prevalence and Characteristics of Multicommunity Engagement. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1056–1066, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms.

Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. 2020. Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling Denotation and Connotation in Vector Spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4090–4105. Association for Computational Linguistics.

Jing Yi Xie, Renato Ferreira Pinto Junior, Graeme Hirst, and Yang Xu. 2019. Text-based inference of moral sentiment change. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4654–4663. Association for Computational Linguistics.

A Appendix

			Linear Equation solver			MultiCCA		
Year	Mode	N	Acc@1	Acc@5	Acc@10	Acc@1	Acc@5	Acc@10
2016	ALL	∼30k	0.6937	0.8335	0.8709	0.7284	0.8591	0.8924
	SW	1000	0.6583	0.8037	0.8450	0.6234	0.7735	0.8174
		5000	0.6934	0.8306	0.8679	0.6984	0.8352	0.8718
2017	ALL	∼30k	0.6720	0.8118	0.8510	0.7092	0.8412	0.8761
	SW	1000	0.6393	0.7829	0.8251	0.6055	0.7532	0.7973
		5000	0.6737	0.8099	0.8484	0.6793	0.8163	0.8542
2018	ALL	∼30k	0.6645	0.8013	0.8402	0.6850	0.8145	0.8509
	SW	1000	0.6336	0.7730	0.8143	0.5992	0.7425	0.7857
		5000	0.6670	0.8003	0.8385	0.6719	0.8057	0.8436
2019	ALL	∼30k	0.6370	0.7809	0.8230	0.6818	0.8165	0.8543
	SW	1000	0.6030	0.7481	0.7923	0.5701	0.7185	0.7467
		5000	0.6401	0.7787	0.8202	0.6505	0.7884	0.8291

Table 9: Performance comparison of yearly alignments by Linear Equation solver & MultiCCA, evaluated based on projection to the shared space.

		2016			2017		
Mode	N	Acc@1	Acc@5	Acc@10	Acc@1	Acc@5	Acc@10
ALL	∼30k	0.4980	0.6462	0.7091	0.5064	0.6683	0.7221
SW	1000	0.4789	0.6434	0.6991	0.4562	0.6130	0.6680
	5000	0.5203	0.6841	0.7374	0.4948	0.6531	0.7063

		2018			2019		
Mode	N	Acc@1	Acc@5	Acc@10	Acc@1	Acc@5	Acc@10
ALL	∼30k	0.4980	0.6562	0.7091	0.4804	0.6382	0.6922
SW	1000	0.4478	0.5993	0.6526	0.4291	0.5792	0.6332
	5000	0.4849	0.6388	0.6909	0.4675	0.6200	0.6729

Table 10: Performance of yearly alignments by SVD, evaluated for community pairs.

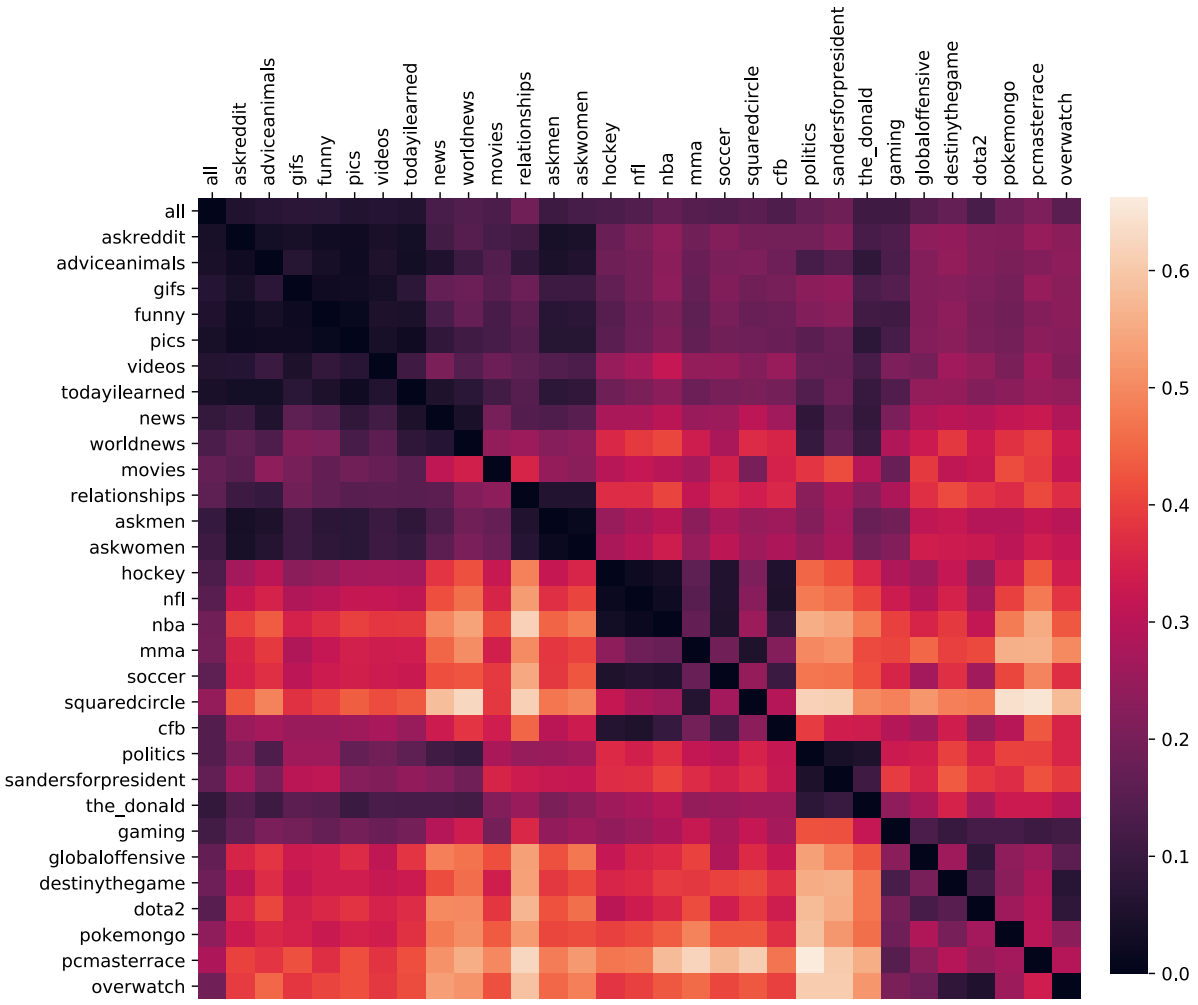

Figure 4: $\mathcal{D}_{KL}(P(t|\mathcal{C}_a)||P(t|\mathcal{C}_b))$, where $\mathcal{C}_a$ is labeled on the y-axis, and $\mathcal{C}_b$ is on the x-axis.

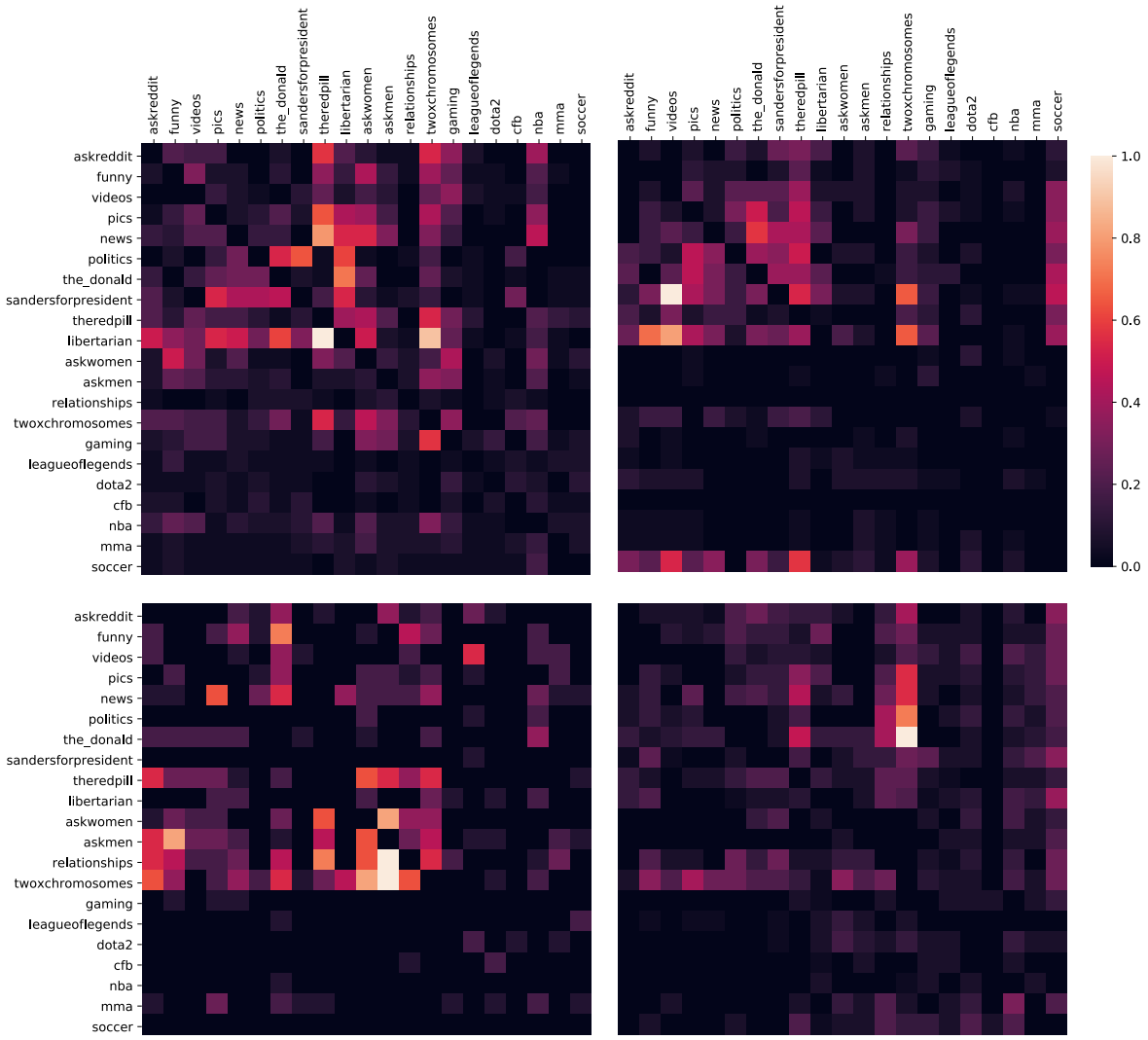

Figure 5: Community misalignment with respect to Government cluster (top left), Conflict cluster (top right), sex cluster (bottom left), religion cluster (bottom right).