Detecting Polarization with Minimal Supervision: Spectral Clustering on Stance-Augmented Embeddings

Ling Lei (ll499), Qinyang Yu (qy50), Yushan Shi (ys468)

ECE 684 Final Project

Abstract

Detecting ideological polarization requires identifying whether a corpus contains an underlying two-camp structure, yet such a structure is often difficult to observe in standard semantic embeddings without substantial stance annotation. This study proposes a minimal-supervision approach that makes such a structure geometrically detectable for large-scale data. Using a small subset of stance-labeled examples, we extract a stance-relevant direction and append it as an additional dimension to standard embeddings, forming a stance-augmented embedding space for the remaining unlabeled corpus. Spectral clustering on this space reveals a clearer two-camp separation than in raw embeddings. We verify that this separation is not a visualization artifact: the spectral bipartition aligns almost perfectly with the first principal component across supervision ratios. Across global train–test splits, even 10–20% supervision ratio yields consistent gains in polarization recoverability, with further improvements as more labels are added. These intrinsic gains translate into higher clustering accuracy on true stance labels, demonstrating extrinsic utility with minimal annotation cost. These results show that minimal supervision can reveal or strengthen latent ideological structure directly from text, offering a scalable, interpretable, and label-efficient tool for polarization detection in large unlabeled corpora.

1 Introduction

Ideological polarization–the division of public discourse into opposing camps–has become a growing challenge for both social science and natural language processing (NLP) research.

Detecting such polarization from text is essential for understanding online discourse, tracking shifts in public opinion, and mitigating misinformation.

Most computational efforts to date have focused on *stance detection*, classifying individual texts as favoring or opposing a specific target (Conforti et al., 2020; Mohammad et al., 2016; Roy and Goldwasser, 2020). These methods operate at the instance level and typically require substantial annotated data. In contrast, *polarization detection* concerns whether a corpus as a whole exhibits a coherent two-camp division, which is a broader and more structural phenomenon. While some recent studies have examined polarization using social network information such as retweet or follower graphs (Darwish, 2019; Garimella et al., 2018; Lyu and Luo, 2022), such data are often unavailable or incomplete, and their analysis depends heavily on access to network metadata. As a result, relatively little work has explored how polarization can be detected directly from text representations alone, without relying on large-scale stance annotation or real-world network structures.

Recent advances in text representation learning have provided new tools for studying ideological polarization in language. Early work used word embeddings to estimate ideological placement in political corpora, mapping lexical choices to continuous ideology scales (Rheault and Cochrane, 2020). Subsequent research developed methods to uncover bias directions in contextualized embeddings without supervision (Hofmann et al., 2022), and proposed partisanship-aware topic modeling or cross-community alignment frameworks to capture ideological differences between groups (He et al., 2021; Milbauer et al., 2021). While these studies have deepened understanding of how ideology is encoded in text, they primarily focus on identifying ideological dimensions or bias directions rather than determining whether a corpus itself contains a recoverable two-camp division–the hallmark of polarization. Moreover, some approaches rely on explicit partisan supervision (He et al., 2021) or complex alignment architectures (Milbauer et al., 2021), which limits their scalability and interpretability in large unlabeled corpora. Even unsupervised methods such as Hofmann et al. (2022) reveal ideological axes but do not make polarization a structural phenomenon directly observable in the embedding space.

This study addresses that gap by proposing a simple, label-efficient framework that directly detects polarization as an emergent, intrinsic two-camp structure in textual embed-

dings using spectral clustering, rather than as a boundary learned by a purely supervised classifier. We first establish an unsupervised baseline and show that standard semantic embeddings, such as SBERT, fail to reveal a clear two-camp structure, which is consistent with findings that topical representations capture semantics but not ideology in stance detection research (Dong et al., 2017). In contrast, our minimally supervised variant uses a small subset of stance-labeled examples to extract a stance-relevant direction, making latent polarization detectable and improving recoverability across multiple evaluation metrics.

Our contributions are threefold. First, social network graphs, often crucial for analysis of community structure, are rarely available due to privacy, access, or platform restrictions. We introduce a weakly supervised framework for polarization detection that operates entirely on text, without relying on network data. Second, most text corpora lack stance labels, making supervised ideological classification infeasible at scale. We empirically show that, while unsupervised semantic embeddings fail to reveal polarization structure, introducing minimal stance supervision makes the two-camp organization clearly recoverable, demonstrating both the effectiveness and label efficiency of our approach. Third, digital platforms and research institutions increasingly require scalable, low-cost, near–real-time tools to monitor polarization in large unlabeled and rapidly evolving datasets. As a real-world implication, our proposed method offers a low-cost and easily deployable solution for detecting emerging ideological divisions. Together, these results show that weak stance supervision can substantially enhance the separation and quantification of ideological polarization, providing a practical and interpretable tool for large-scale social media analysis and intervention.

The remainder of this paper is organized as follows. Section 2 describes the dataset and preprocessing steps. Section 3 presents the proposed embedding and spectral clustering framework. Section 4 reports the main results and evaluations. Section 5 concludes.

2 Data

2.1 Descriptive Statistics

The dataset comes from SemEval (International Workshop on Semantic Evaluation) (Mohammad et al., 2017) and focuses on a single target (Donald Trump). In total, we have 707 observations. For each observation, the data contains the following six features: *ID*, *Target*, *Tweet*, *Stance*, *Opinion towards*, *Sentiment*. Table 1 shows the detailed value distribution for each feature.

Feature	Value	Count	Total Count
Target	Donald Trump	707	707
Stance	Against	299	707
	None	260	
	Favor	148	
Opinion towards	Other	356	707
	Target	324	
	No One	27	
Sentiment	Negative	481	707
	Positive	193	
	Neither	33	

Table 1: Value Distribution for Each Feature.

Table 2 reports the cross-tabulation for *Sentiment* and *Stance* with respect to *Opinion towards*. *Sentiment* captures the overall emotional expression of the text (positive, negative, or neutral), while *Stance* is defined relative to a pre-specified target (here, Donald Trump) and reflects whether the person supports, opposes, or does not take a position on that target. Thus, a tweet can express negative sentiment (e.g., sadness, anger) while still taking a pro-Trump stance, or positive sentiment while taking an anti-Trump stance. Table A provides some examples.

2.2 Data Preprocessing

All tweet text was converted to lowercase and cleaned of noise, such as URLs, @mentions, the # notation, and whitespace. Since the research target is detecting polarization toward

Opinion To.	No One	Other	Target	All
Stance				
Against	1	122	176	299
Favor	0	3	145	148
None	26	231	3	260
Sentiment				
Negative	3	296	182	481
Neither	13	12	8	33
Positive	11	48	134	193
All	27	356	324	707

Table 2: Cross-table of *Stance* or *Sentiment* w.r.t *Opinion towards*.

the target "Donald Trump," we exclude the ambiguous category of "None" in *Stance*, and only focus on two categories: "Favor" and "Against." After data cleaning, *Stance* has 299 observations against target "Donald Trump," and 148 favors observations. For cleaned Tweet contexts, Table 3 shows examples of the cleaned tweet texts.

3 Method

Our objective is to determine whether a corpus contains an underlying two-camp ideological structure and to make such structure—if it exists—geometrically detectable with no or minimal stance supervision.

We work with a collection of tweets $\{x_i\}_{i=1}^{N}$, each of which carries a stance label

$$y_i \in \{0 \text{ (Against)}, 1 \text{ (Favor)}\}.$$

These labels are available for the entire dataset but are deliberately used in two different ways. First, they are completely ignored when constructing a fully unsupervised baseline, in which we ask whether polarization can be recovered from semantic information alone. Second, in the minimal-supervision setting, only a randomly selected subset of labels is used to extract a stance-relevant direction; the remaining labels are withheld and used solely for evaluation.

To study label efficiency in the minimal-supervision setting, we fix a supervision ratio

ID	Tweet_Cleand
20004	stupid is as stupid does! showedhis true colors; seems that he ignores that us was invaded, & plundered,not discovered semst
20006	donald trump isn't afraid to roast everyone. semst
20007	donald trump for president? i can dig it. semst
20011	100% support trump.. semst
20013	considering the fact that bush was a president of this country, i don't see it a joke that trump is running ! election2016 semst
20017	i like mexicans who come to us legally. istandwithtrump semst
20018	we need obama out and in the white house asap semst
20020	dear : you are an idiot. america politics sticktoyourhair semst
20022	donald trump inhaled so much air spray & started to believe he is the title holder for the united states. semst
20024	. should've kept his mouth shut & not run for president. he is making the biggest fool out of himself. he's fired semst
20026	so i guess univision is fair & balanced. these are the people that r helping shape usa! semst
20027	presidentialelection2016 make plans to help your future now, so that later you don't regret it, again! vote semst
20029	you know that middle class is latino's, right? you just shitted all over that working class you claim to need. semst
20032	rt : let's not forget lost money running a casino. semst
20034	did not apply to immigrants one of the trade basis, win to win. ignorance can not be excuse semst

Table 3: Cleaned Tweet Texts

$r \in (0, 1)$ and sample rN labeled examples as a weakly supervised training set D_{train}. The remaining $(1 - r)N$ tweets are treated as unlabeled test data D_{test} for all representation learning and clustering steps. For each r, we (i) learn a stance-aligned direction from D_{train}, (ii) construct a stance-augmented embedding space for D_{test}, (iii) perform unsupervised spectral clustering on the test set, and (iv) evaluate the recovered two-camp structure using polarization metrics. The unsupervised baseline corresponds to the special case where no stance labels are used at any stage and clustering is performed directly on raw semantic embeddings.

3.1 Embedding Space

We begin by encoding each tweet into a semantic sentence embedding using Sentence-BERT:

$$z_i = f_{\text{SBERT}}(x_i) \in \mathbb{R}^d,$$

where we use `all-MiniLM-L6-v2` with embedding dimension $d = 384$.

3.1.1 Baseline Model

In the fully unsupervised setting, the representation space is simply $\{z_i\}$. We construct a graph and run spectral clustering directly on these semantic embeddings. This pipeline serves as a reference point: if polarization is already visible in standard embeddings, then clustering on $\{z_i\}$ should recover a two-camp partition; if not, we expect poor separation and low agreement with stance labels. Since these embeddings primarily capture topical and semantic similarity between tweets, we hypothesize that they do not yield a geometry in which ideological camps form clearly separable clusters.

3.1.2 Stance-augmented Model

In the minimal-supervision setting, we inject a weak stance signal into the embedding space using only D_{train}. We fit a logistic regression model on the labeled training embeddings:

$$p(y = 1 \mid z) = \sigma(w^\top z + b), \text{ where } \sigma(t) = \frac{1}{1 + e^{-t}}.$$

The role of this model is not to perform classification on the test set, but to extract a stance-aligned direction w in the semantic space along which the labeled examples are most separable. Applying the trained model to the unlabeled embeddings in D_{test} yields a continuous stance score

$$p_i = p(y_i = 1 \mid z_i),$$

for each test tweet.

To make these scores comparable across supervision levels, we standardize them over the test set and obtain

$$\tilde{p}_i = \frac{p_i - \mu_p}{\sigma_p + 10^{-8}},$$

where μ_p and σ_p are the mean and standard deviation of $\{p_i\}_{i \in D_{\text{test}}}$. We add a small constant

10^{-8} to the denominator to ensure numerical stability, preventing division by zero in cases where the stance scores have near-zero variance. We then rescale the standardized scores by a fixed factor $\lambda = 4.0$ to obtain a stance coordinate

$$s_i = \lambda \tilde{p}_i.$$

This single scalar encodes where each tweet lies along the learned stance axis.

Finally, we augment each test embedding with this stance coordinate and obtain

$$\tilde{z}_i = [\, z_i \parallel s_i \,] \in \mathbb{R}^{d+1}.$$

The resulting stance-augmented embeddings preserve the original semantic information while adding one explicitly stance-informed dimension. Because we only add a single coordinate derived from a weak classifier, the overall geometry is minimally perturbed: if no latent polarization exists, spectral clustering on $\{\tilde{z}_i\}$ will not artificially create a two-camp structure; if such structure is present but obscured in $\{z_i\}$, we hypothesize that the added dimension should make it more geometrically visible.

3.2 Graph Construction and Spectral Clustering

To study community structure in the chosen embedding space, we construct a similarity graph over the test set. In the unsupervised baseline, the graph is built from the raw semantic embeddings $\{z_i\}$; in the minimal-supervision model, it is built from the stance-augmented embeddings $\{\tilde{z}_i\}$. In both cases, we define pairwise affinities using a Radial Basis Function (RBF) kernel:

$$A_{ij} = \exp\!\big(-\gamma \|u_i - u_j\|^2\big), \quad \gamma = \frac{1}{d+1},$$

where $u_i = z_i$ for the baseline and $u_i = \tilde{z}_i$ for the stance-augmented model. The hyperparameter γ is scaled by the dimensionality to keep distances numerically well-behaved across spaces. The resulting affinity matrix A defines a weighted graph $G = (D_{\text{test}}, A)$ whose connectivity reflects local similarity in the chosen representation.

We detect two-camp structure in this graph using spectral clustering, which seeks a partition that cuts as few high-weight edges as possible. Let D be the diagonal degree matrix with $D_{ii} = \sum_j A_{ij}$. We compute the normalized Laplacian

$$L = I - D^{-1/2} A D^{-1/2}.$$

Spectral theory implies that the smallest eigenvalue of L is zero and that its second-smallest eigenvector–the Fiedler vector–captures the most natural bipartition of the graph. We solve

$$Lu_1 = \lambda_1 u_1, \quad Lu_2 = \lambda_2 u_2,$$

with eigenvalues ordered as

$$0 = \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots .$$

We then apply k-means with $k = 2$ to the entries of u_2, obtaining cluster assignments $c_i \in \{0, 1\}$ for each test tweet. The procedure is identical in the baseline and stance-augmented settings; only the underlying embedding space (and hence the corresponding A and L) differ.

3.3 Polarization Metrics

To evaluate polarization detection, our primary metric is the Silhouette score, which measures how well-separated the two clusters are in the embedding space:

$$\text{Silhouette} = \frac{1}{M} \sum_i \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the mean distance from point i to other points in its assigned cluster and $b(i)$ is the mean distance to the nearest opposing cluster, with $M = |D_{\text{test}}|$. Higher Silhouette values indicate tighter within-cluster cohesion and larger between-cluster separation in the semantic geometry, whereas values near zero imply heavy overlap between groups. Because the true stance labels encode the underlying two-camp ideological division in this dataset, the degree of geometric separation directly reflects how well this latent polarization is recovered

and expressed in the embedding space.

We also measure structural polarization on the affinity graph. The Fiedler eigenvalue λ_2 (the second-smallest eigenvalue of the normalized Laplacian L) quantifies how close the graph is to an ideal two-block structure. Smaller values of λ_2 indicate weaker connectivity between the two clusters, which indicates a stronger two-block structure in the high-dimensional semantic space and therefore a sharper bipartition.

Finally, we report the modularity Q of the two-way partition obtained from spectral clustering, which measures the strength of division of a network into modules. Modularity compares the observed within-cluster edge weight to the expected weight under a degree-preserving null model:

$$Q \;=\; \frac{1}{2m}\sum_{i,j}\Big(A_{ij} - \frac{k_i k_j}{2m}\Big)\mathbf{1}\{c_i = c_j\},$$

where A is the affinity matrix on D_{test}, k_i is the weighted degree of node i, $m = \frac{1}{2}\sum_{i,j} A_{ij}$, and c_i is the cluster assignment. Higher values of Q indicate that nodes within the same cluster are more densely connected than expected by chance and nodes in different clusters have sparser between-group connections, reflecting stronger community-level polarization.

4 Results

4.1 Baseline and Stance-Augmented Results

Table 4 reports the baseline and stance-augmented model results—the model performance of polarization detection, measured by three metrics described in Section 3.3—under varying supervision ratios (0.1–0.9).

The baseline results are uniformly weak but stable. Silhouette scores remain between 0.03 and 0.04, the Fiedler eigenvalue λ_2 stays consistently near 1 across all splits, and modularity Q remains stable at 0.0001. All three metrics indicate that the raw semantic space behaves like an almost uniform cloud, with no detectable geometric separation. In other words, the pure BERT embedding shows no intrinsic two-block structure indicative of polarization, aligned with our hypothesis. To provide an illustrative visualization, we applied Principal Component Analysis (PCA) to project the high-dimensional semantic embeddings onto a

	Baseline Model			Stance-Augmented Model		
r	Silhouette	λ_2	Q	Silhouette	λ_2	Q
0.10	0.0348	1.0023	0.0001	0.5274	0.9249	0.0259
0.20	0.0354	1.0026	0.0001	0.5426	0.9244	0.0266
0.30	0.0357	1.0030	0.0001	0.5369	0.9250	0.0251
0.40	0.0379	1.0035	0.0001	0.5500	0.9252	0.0259
0.50	0.0393	1.0043	0.0001	0.5395	0.9253	0.0265
0.60	0.0411	1.0054	0.0001	0.5584	0.9265	0.0265
0.70	0.0419	1.0073	0.0001	0.5558	0.9286	0.0269
0.80	0.0434	1.0110	0.0001	0.5472	0.9331	0.0255
0.90	0.0605	1.0225	0.0001	0.5030	0.9475	0.0242

Table 4: Baseline and Stance-Augmented Model Results

two-dimensional plane, as shown in the top panels of Figure 1 and Figures A–H in the appendix. Across all supervision ratios, take Figure 1 ($r = 0.1$) as an example: the PCA visualizations of the baseline model (top-left panel) consistently exhibit a uniform cloud structure with no apparent separation. Consequently, the spectral clustering assignments shown in the baseline panels (top-right subgraph) fail to recover any meaningful polarization structure.

By contrast, the stance-augmented model exhibits clear structural changes: Silhouette scores increase to 0.50–0.56, the Fiedler value drops toward 0.92–0.95, and modularity Q rises to 0.024–0.027. All three metrics shift in the desired direction, confirming the improvement by introducing stance information. The geometry becomes substantially more separable in the new semantic space, which is also intuitively visualized in the bottom panels of Figure 1 and Figures A–H. Taking Figure 1 ($r = 0.1$) as an example, the stance-augmented model exhibits a clear separation in the PCA plot (bottom-left panel). Correspondingly, the spectral clustering results (bottom-right panel) now more accurately recover the polarization structure, with cluster assignments more closely matching the true stance labels in this augmented embedding space.

Remarkably, these improvements remain stable across different train-test splits, which indicates that the stance-augmented model remains robust in its effectiveness. Supervision on only 10% of the data is sufficient to reconfigure the semantic space. Therefore, even a

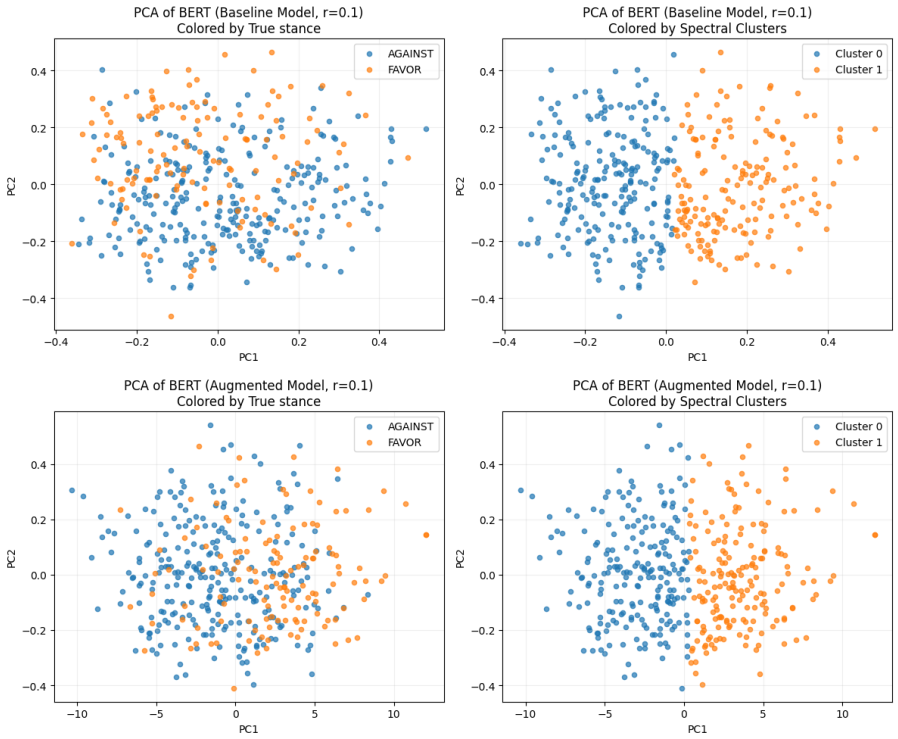

Figure 1: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.1$

limited, small subset of artificial stance labels can substantially alter the geometry, resulting in clear cluster separation. This suggests that the underlying BERT space is malleable and highly responsive to small amounts of task-specific labeled information. We formally verify this robustness across different supervision ratios in Section 4.2.

To verify that the observed alignment between PCA and spectral clustering is not accidental, we further compute, for each supervision ratio $r \in \{0.1, \ldots, 0.9\}$, the absolute Pearson correlation between the projection scores on the first principal component (PC1) of the stance-augmented embeddings and the entries of the Fiedler vector of the normalized Laplacian. Figure 2 shows that the correlation remains extremely high ($|\text{corr}| > 0.998$) across all values of r, confirming that the spectral bipartition essentially coincides with the dominant variance direction introduced by the stance augmentation. This validates that the PCA separation is an expected structural outcome of the augmented embedding space, rather than a visualization artifact.

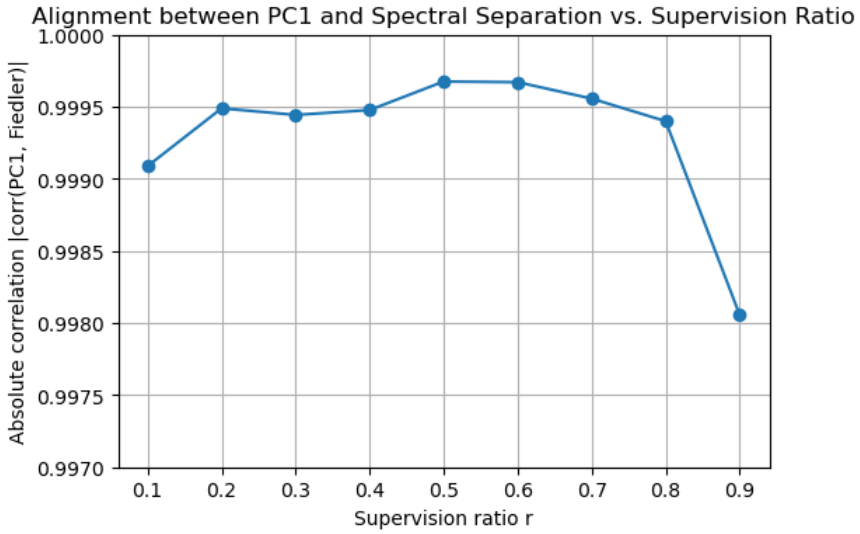

Figure 2: Alignment between PC1 and spectral separation vs. supervision ratio.

4.2 Supervision Ratio Efficiency Evaluation

To more precisely examine how minimal stance supervision affects the recoverability of polarization, we vary the supervision ratios for the stance-augmented model

$$r \in \{0.01, 0.02, \ldots, 0.99\}.$$

For each r, we resample D_{train}, reconstruct the stance-augmented embeddings, rebuild the graph, and rerun spectral clustering. We repeat the entire pipeline for $K = 10$ independent runs and report, for each metric, the mean and 95% confidence interval across runs. Finally, we draw a smooth line graph for each metric across different values of r, which provides a detailed view of how sensitive the efficiency of polarization detection is to the proportion of stance supervision.

Figure 3, 4, and 5 show that moderate supervision can lead to significant recoverability improvements. Across the range of 10%–80% supervision, the Silhouette score reaches a level of 0.50 at the beginning already and shows a fairly steady growth with a higher supervision ratio, as depicted in Figure 3. Figure 4 exhibits that the Fiedler value λ_2 remains stably low (around 0.93), indicating a robust two-block structure. Figure 5 shows a stable, strong community structure. All three figures suggest that even modest amounts of stance labels

are already sufficient to reconfigure the semantic space into a clear polarized geometry, and the efficiency remains robust with a higher supervision ratio.

As supervision exceeds 80%, we observe misbehavior patterns in all three metrics. Silhouette continues to improve (reaching ∼0.66) but the confidence interval is strikingly wider. The Fiedler value increases sharply toward 1.17, signaling a weakening of the two-camp structure. Modularity Q shows a drastic increase and greater variance in this high-supervision regime. The misbehavior is most likely caused by the small dataset size and overfitting to fine-grained stance distinctions. With excessive supervision, the model may fragment the semantic space into multiple sub-communities rather than maintaining the coarse two-block polarization. In other words, too much labeled information causes the geometry to capture idiosyncratic stance patterns at the expense of the macro-level left-right divide. Additionally, a smaller test set at high supervision ratios contributes to larger variance across runs.

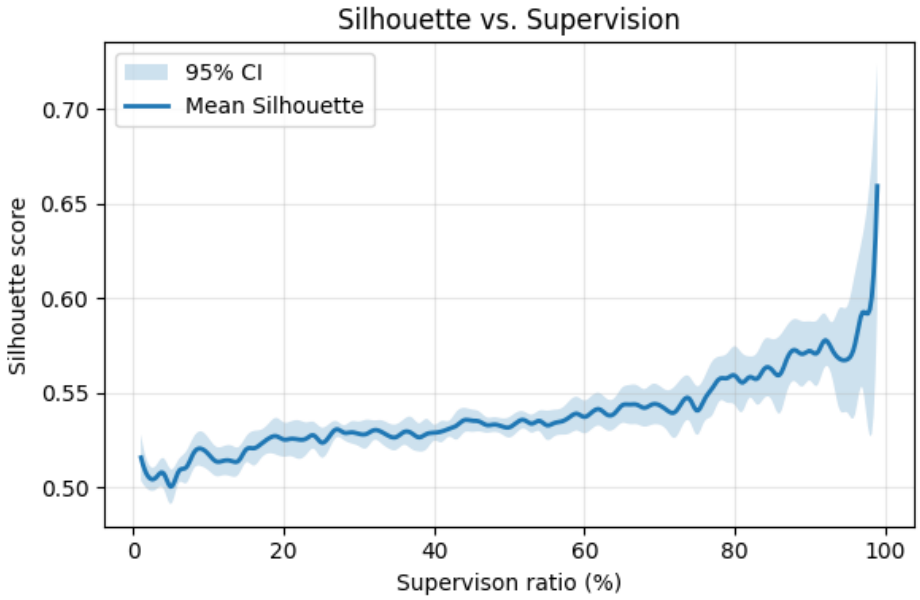

Figure 3: Silhouette vs. supervision ratio; higher scores indicate clearer two-camp separation.

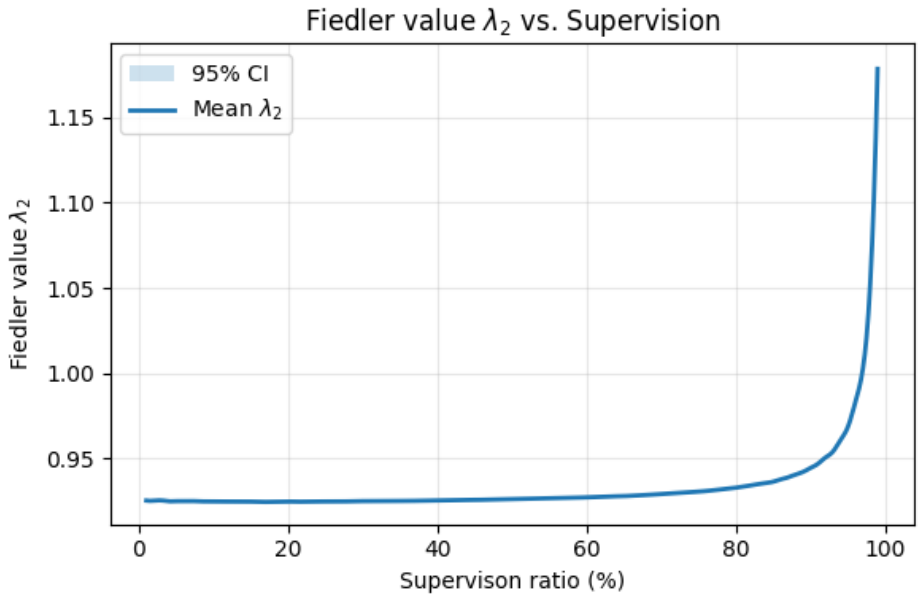

Figure 4: Fiedler value λ_2 vs. supervision ratio; lower values indicate clearer two-camp separation.

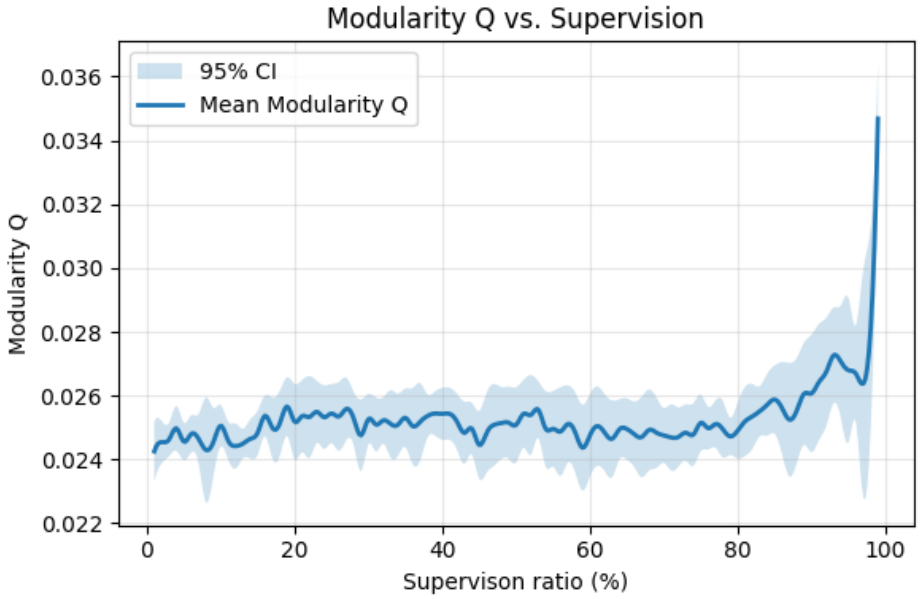

Figure 5: Modularity Q vs. supervision ratio; higher values indicate stronger community structure.

4.3 Extrinsic Utility Evaluation

The previous sections establish that the stance-augmented embedding produces a clear two-block geometry, and that this structure is robust to different supervision ratios. However, intrinsic separability alone does not guarantee practical utility. A meaningful validation of the method requires showing that the induced geometry improves performance on a downstream task. In our setting, the natural downstream objective is *polarization detection*: given only the learned two-way partition from spectral clustering, can we recover the underlying stance labels of the samples in the test set?

To operationalize this evaluation, we use the spectral cluster assignments as predicted stance labels, and compute the clustering accuracy against the ground-truth binary stance in the test set. A higher accuracy indicates a more interpretable and operationally useful embedding, since the recovered clusters more faithfully correspond to the true ideological camps. We report accuracy over supervision ratios $r \in \{0.01, 0.02, \ldots, 0.99\}$, and repeat the entire pipeline $K = 10$ times for each r to estimate the mean accuracy and its 95% confidence interval.

Figure 6 presents the results. The baseline model exhibits stable but uniformly low accuracy, fluctuating around 0.66 across all supervision ratios. This confirms that the raw BERT embedding lacks a recoverable polarization structure: spectral clustering on the pure semantic space does not yield meaningful stance labels, and increasing the amount of labeled data does not improve performance in this regime.

In contrast, the stance-augmented model demonstrates a substantial and consistent improvement. With only 10% supervision, accuracy rises from $\sim$0.66 to $\sim$0.70. As r increases to the range of 20%–60%, accuracy climbs steadily and stabilizes around 0.75 with narrow confidence intervals. This result shows that a small number of stance labels are sufficient to reconfigure the semantic space into a geometry that directly aligns with the underlying ideological camps. Even partial supervision therefore yields a meaningful polarization signal that can be exploited by an unsupervised clustering method.

At very high supervision ratios ($r > 80\%$), the variance increases and the mean accuracy exhibits instability. We attribute this behavior to the reduced test set size and the potential

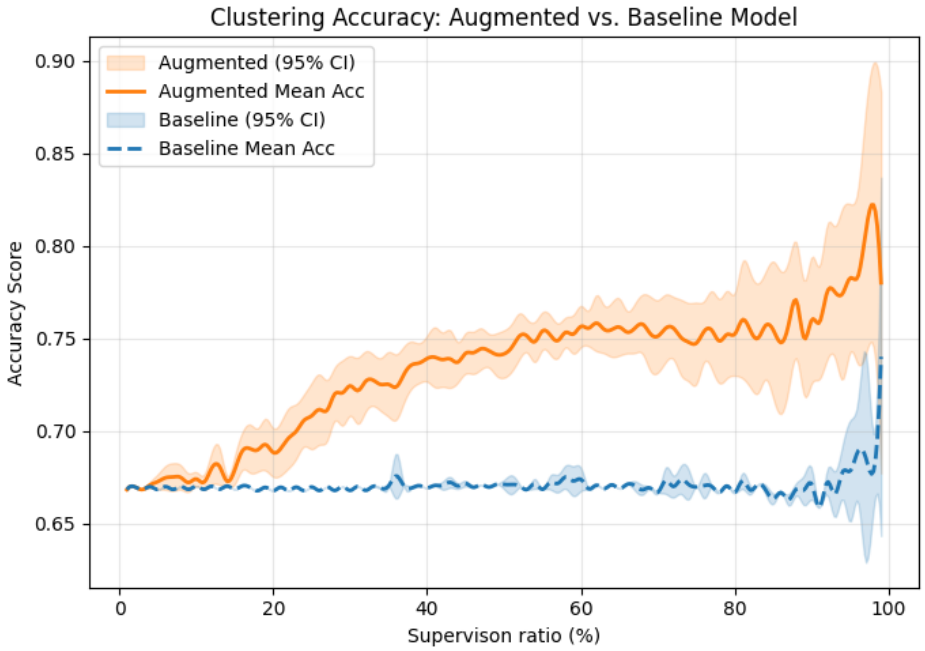

Figure 6: Clustering accuracy for polarization detection vs. supervision ratio.

over-fragmentation of the semantic space when the stance dimension dominates the representation. Nevertheless, the key observation remains: in the practically relevant regime of limited annotated data, the stance-augmented embedding reliably enhances polarization detection accuracy.

Overall, these results confirm the extrinsic utility of stance augmentation. The geometric separation induced by minimal supervision is not merely a visualization effect, but translates into a measurable improvement on a real downstream task. The method enables effective recovery of polarization structure from text with only a small fraction of manual stance annotations.

5 Conclusion

This study demonstrates that minimal stance supervision can make latent ideological polarization directly observable: whether two coherent, well-separated communities naturally emerge in a high-dimensional embedding space. Unlike classification-based approaches, our framework treats polarization as a structural property of the representation space by intro-

ducing a stance-aligned direction into standard semantic embeddings and performing spectral clustering. The observed geometric separation is not a visualization artifact: we show that the first principal component of the augmented space is consistently aligned with the Fiedler vector of the graph Laplacian across all supervision ratios.

Moreover, even with as little as 10–20% labeled data, the model consistently enhances the visibility of two-camp structure across multiple polarization metrics, confirming that weak stance guidance is sufficient to uncover meaningful ideological division. Importantly, this intrinsic structure also translates into extrinsic utility, as clustering accuracy improves by 8–10 percentage points in the augmented model, demonstrating that the recovered split meaningfully reflects true ideological camps.

This label-efficient approach offers a scalable and interpretable tool for analyzing polarization in large text corpora that lack stance labels or network structure information, with potential applications in real-time monitoring of public discourse and information diffusion.

Despite its effectiveness, the approach has several limitations. First, the current analysis excludes the "None" stance category, potentially omitting neutral or ambivalent voices that could play a stabilizing role in polarization dynamics. Future work could incorporate multi-class stance signals or continuous ideology spectra to capture more nuanced political alignments. Second, the dataset used in this study is relatively small and limited to a single topic and target (Donald Trump), which may constrain generalizability. Applying the framework to larger and more diverse corpora, such as multi-topic political debates, news media, or cross-platform social data, would better test its robustness. Additionally, the current method relies on a linear, uni-dimensional stance direction. Exploring nonlinear, multi-dimensional mappings or adaptive stance manifolds could further improve the representation of complex ideological spaces, thereby increasing the recoverability of polarization structure in the embedding space.

References

Conforti, C., Berndt, J., Pilehvar, M. T., Giannitsarou, C., Toxvaerd, F., and Collier, N. (2020). Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

Darwish, K. (2019). Quantifying polarization on twitter: the kavanaugh nomination. In *International conference on social informatics*, pages 188–201. Springer.

Dong, R., Sun, Y., Wang, L., Gu, Y., and Zhong, Y. (2017). Weakly-guided user stance prediction via joint modeling of content and social interaction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1249–1258.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922.

He, Z., Mokhberian, N., Câmara, A., Abeliuk, A., and Lerman, K. (2021). Detecting polarized topics using partisanship-aware contextualized topic embeddings. *arXiv preprint arXiv:2104.07814*.

Hofmann, V., Pierrehumbert, J. B., and Schütze, H. (2022). Unsupervised detection of contextualized embedding bias with application to ideology. *arXiv preprint arXiv:2212.07547*.

Lyu, H. and Luo, J. (2022). Understanding political polarization via jointly modeling users, connections and multimodal contents on heterogeneous graphs. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4072–4082.

Milbauer, J., Mathew, A., and Evans, J. (2021). Aligning multidimensional worldviews and discovering ideological differences. In *Proceedings of the 2021 conference on empirical methods in Natural Language Processing*.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, 17(3).

Rheault, L. and Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political analysis*, 28(1):112–133.

Roy, S. and Goldwasser, D. (2020). Weakly supervised learning of nuanced frames for analyzing polarization in news media. *arXiv preprint arXiv:2009.09609*.

Appendix: Tables

Index	Tweet	Stance	Sentiment
12	Considering the fact that Bush was a president of this country, I don't see it a joke that Trump is running ! #Election2016 #SemST	Favor	Negative
55	@realDonaldTrump . . . Hillary Clinton #SemST	Favor	Negative
114	A vote against Trump is a vote against America #SemST	Favor	Negative
169	I suddenly find myself liking @nbc a whole lot more than ever before. forPresidentoftheClubofIdiots #SemST	Against	Positive
208	@NBCUniversal maybe you could have a reality show trying to find a new host of celebrity apprentice #YoureFired #SemST	Against	Positive
228	Hutch_USA Follow ===> sarah_brannick <== if you want to #FreeAmir . . . #SemST	Against	Positive

Table A: Examples where *Stance* and *Sentiment* disagree.

Appendix: Figures

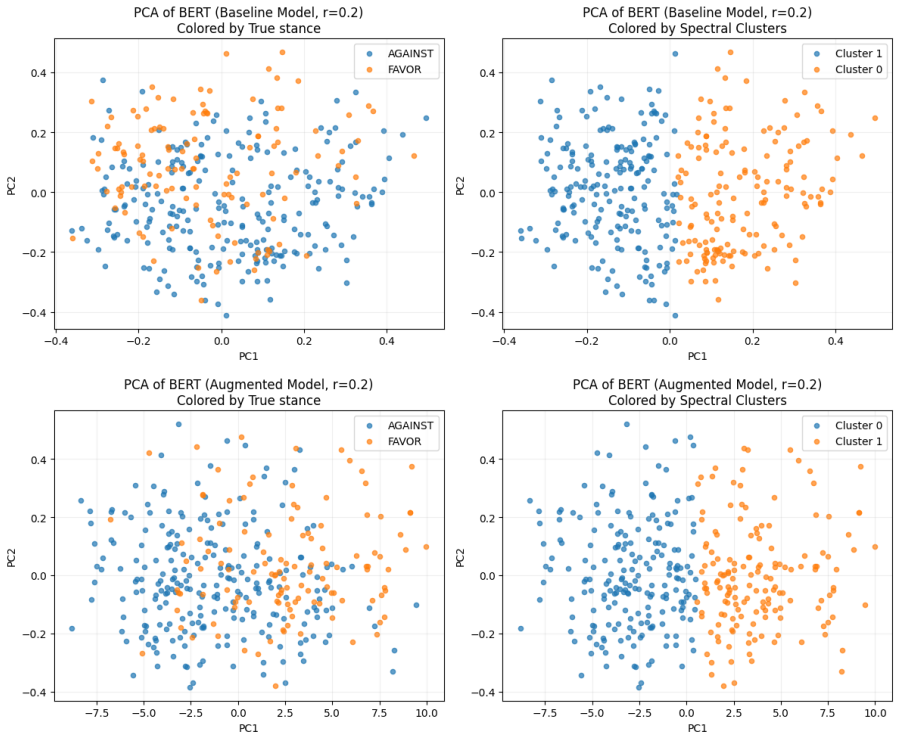

Figure A: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.2$

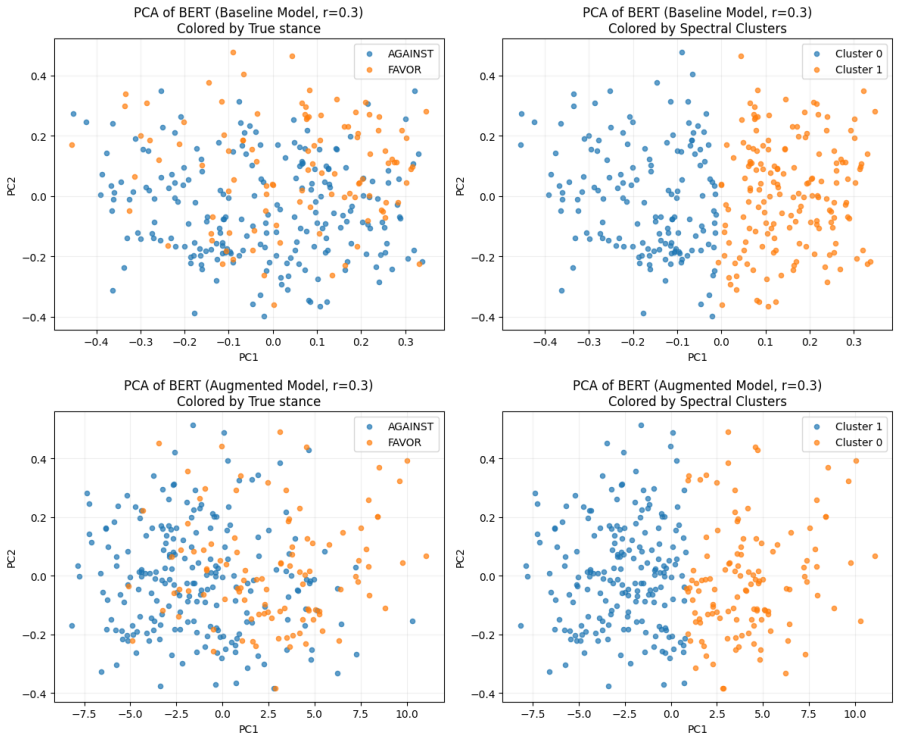

Figure B: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.3$

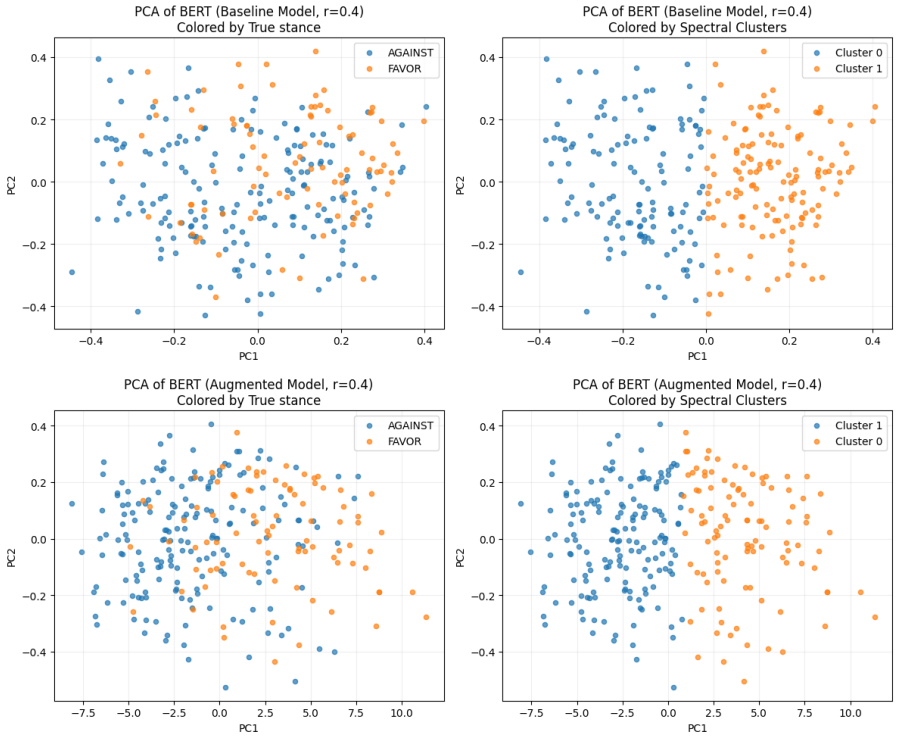

Figure C: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.4$

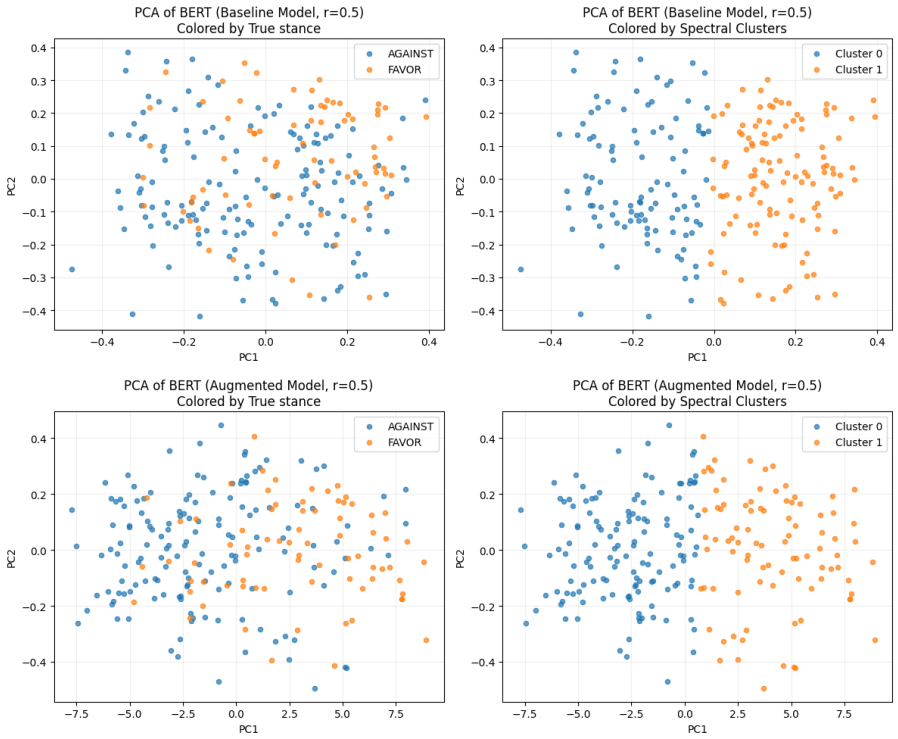

Figure D: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.5$

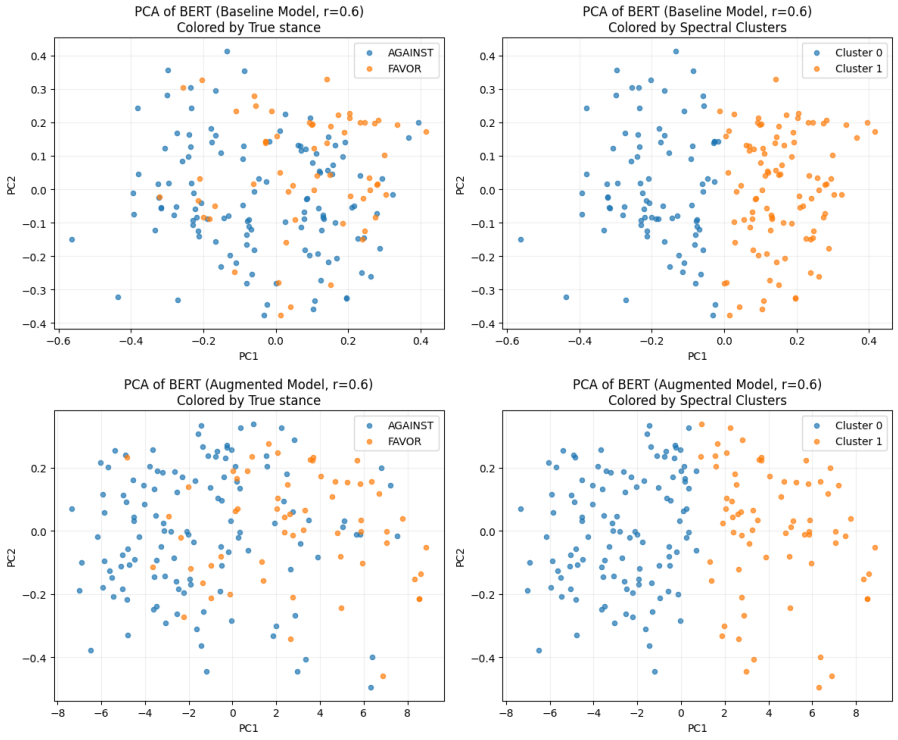

Figure E: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.6$

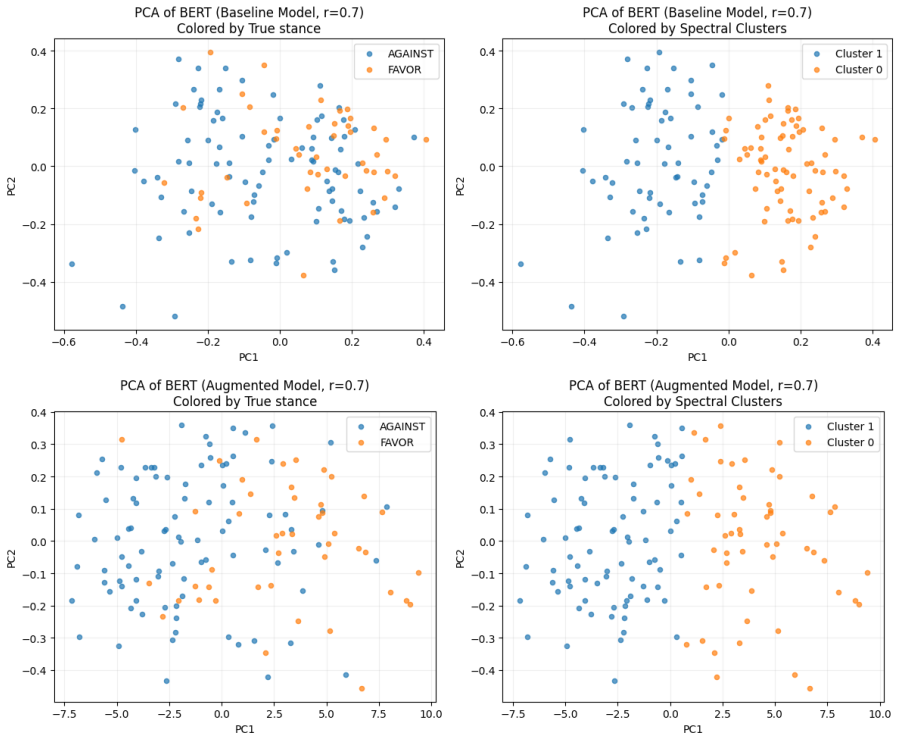

Figure F: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.7$

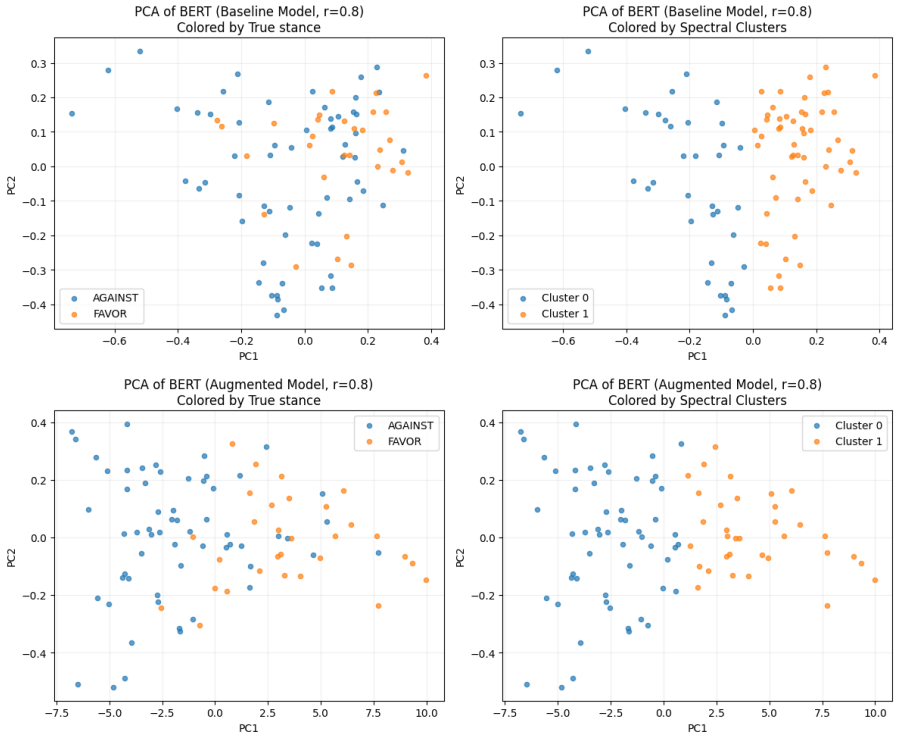

Figure G: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.8$

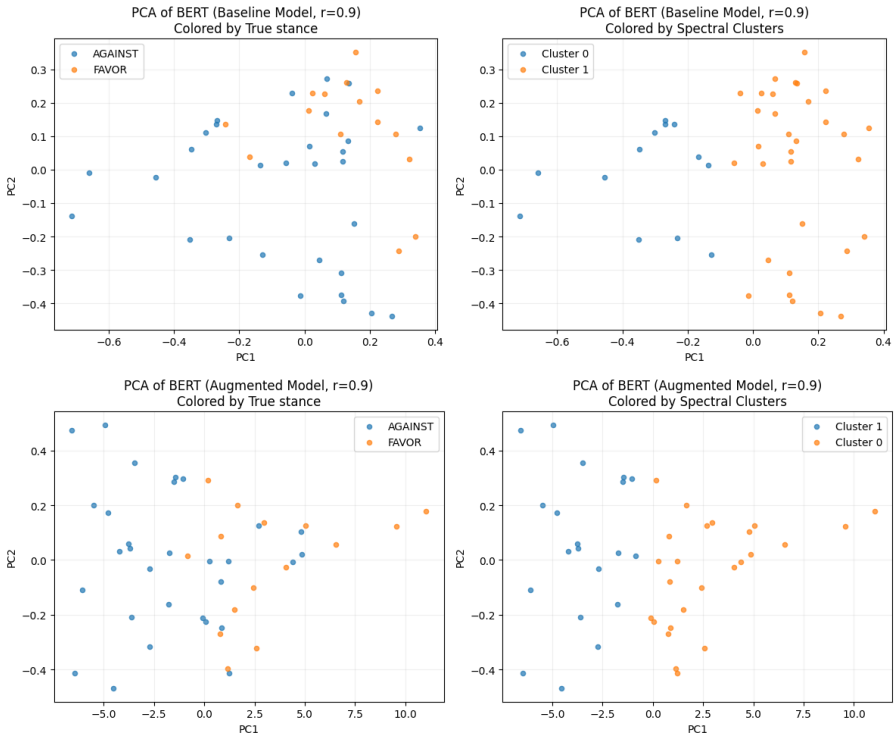

Figure H: PCA of baseline (top) and stance-augmented (bottom) embeddings at $r = 0.9$