Uncovering Hidden Factions through Text-Network Representations: Unsupervised Public Opinion Mapping of Iran on Twitter in the 2022 Unrest

Sahar Omidi Shayegan[†‡] **Jean-Francois Godbout**[‡§] **Reihaneh Rabbany**[†‡]
[†] McGill University
[‡] Mila - Quebec AI Institute
[§] University of Montreal
sahar.omidishayegan@mail.mcgill.ca

Abstract

Ideological mapping on social media is typically framed as a supervised classification task that depends on stable party systems and abundant annotated data. These assumptions fail in contexts with weak political institutionalization, such as Iran. We recast ideology detection as a fully unsupervised mapping problem and introduce a text-network representation system, uncovering latent ideological factions on Persian Twitter during the 2022 Mahsa Amini protests. Using hundreds of millions of Persian tweets, we learn joint text–network embeddings by fine-tuning ParsBERT with a combined masked-language-modeling and contrastive objective and by passing the embeddings through a Graph Attention Network trained for link prediction on time-batched subgraphs. The pipeline integrates semantic and structural signals without observing labels. Density-based clustering reveals eight ideological blocs whose spatial relations mirror known political alliances. Alignment with 883 expert-labeled accounts yields 53% accuracy. This label-free framework scales to label-scarce contexts, offering new leverage for studying political debates online.

1 Introduction

Political ideology detection on Twitter (now X) has become a central task in computational social science, especially within the context of English-speaking democracies. Numerous studies have leveraged social media data to classify users' political leanings and examine partisan dynamics online (e.g., Barberá (2015); Pennacchiotti & Popescu (2011); Pelrine et al. (2023); Yu et al. (2023); Törnberg (2023)). Most of the work done studies this topic in Western contexts or English language countries, where ideologies map clearly onto formal party affiliations (Rodríguez-García et al., 2022; Chen et al., 2017; Jiang et al., 2023).
In this work, we focus on the Iranian Twittersphere during the "Mahsa Amini" protests of 2022—a period marked by intense social and political upheaval (Khorramrouz et al., 2023). Unlike democracies with clear-cut partisan alignments, Iran's political divisions are largely ideological and fluid, lacking institutionalized party structures. This introduces unique challenges for computational approaches to analyze partisan debates, with ambiguous group boundaries and subtle self-identification.

To address this gap, we propose a method for capturing the ideological landscape in politically complex settings, integrating two key signals from social media users: (1) textual content derived from tweets, retweets, and user biographies, and (2) structural features from their retweeting network. These signals are jointly modeled using unsupervised representation learning, where we refine text embeddings through a Graph Attention Network (GAT) using the structure of the retweet graph (Veličković et al., 2018).
In addition to identifying ideological groupings, our method provides insights into their relative positioning in the representation space. Our results show that the spatial layout of user embeddings reflects ideological proximity, with overlapping or opposing groups mapping accordingly into a meaningful representation of Iran's main political divisions.

Although we report classification metrics, the learning and discovery in certain settings are entirely unsupervised meaning the graph-refined text embeddings are produced without ever observing labels; clustering the embedding space unveils ideological groupings. To validate these clusters, we rely on a small, high-precision reference set of 883 users whose affiliations were manually verified by domain experts and exhibit unmistakable partisan cues (e.g., explicit slogans, known public figures). We map clusters to ideological labels via maximum-matching against this reference set and fit a multilayer perceptron for post-hoc evaluation. This provides a quantitative check on how faithfully the unsupervised clusters align with expert knowledge. As shown in Figure 1, our method embeds users in a representation space that closely reflects the relative positioning of ideological classes and estimates their distribution within the ideological landscape. This design keeps discovery label-free while still offering several familiar metrics for comparison.

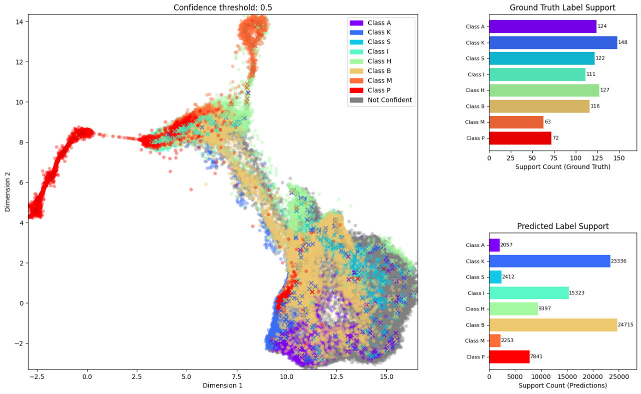


Figure 1: Left: 2D projection of the embedding space showing predicted labels (O) and ground-truth labels (X), filtered by a confidence threshold of 0.5 (see Section 6.2). Top right: distribution of ground-truth hyperpartisan labels. Bottom right: distribution of predicted labels above the confidence threshold.

The main contributions of this work are as follows.

- **Task Formulation:** The first unsupervised framework for ideology detection in settings without well-defined political parties or formal partisan labels.
- **Methodology:** We combine contrastive and masked language modeling to learn text embeddings, and refine them using GNNs based on retweet interactions, which leads to a better graphical representation of political divisions.
- **Application Context:** We apply our method to a large-scale dataset of Persian Twitter activity during the Mahsa Amini protests, enabling a data-driven exploration of ideological conflicts in Iran.

2 Related Work

Ideology Detection via Text and Social Network Analysis. The task of political ideology detection on social media has traditionally focused on Western liberal democracies, where users' partisan identities often align with established political parties. In these settings, researchers have used both textual content and network structure to infer the ideological leanings of online actors. One study demonstrated that language in tweets can predict users' political ideology along a fine-grained spectrum (Preoţiuc-Pietro et al., 2017), while another leveraged ideological phrase indicators to detect how politicians frame issues (Johnson et al., 2017). More recently, Yu et al. (2023) showed that models like RoBERTa outperform even large LLMs on ideology detection by distinguishing between 'explicit' and 'implicit' cues. Moving beyond textual analysis, network-based approaches leverage homophily (i.e., the tendency for like-minded users to connect within social networks (Barberá, 2015; Enjolras & Salway, 2022)) to infer users' political ideologies. In this context, retweet and follow networks have been used to place users on a latent ideological space (Barberá, 2015) and to identify ideological clustering in polarized events (Enjolras & Salway, 2022). These graph structures often reveal ideological affinity even in the absence of textual data, and can uncover community-level polarization as well as echo chambers.

Importantly, combined approaches have been shown to significantly enhance performance. TIMME (Xiao et al., 2020) and Retweet-BERT (Jiang et al., 2023) jointly model user content (tweets or bios) and network structure (retweets, mentions), showing that user embeddings informed by both modalities are more robust than using either one of these approaches

Class	Label	Description Based on Bio Content
A	Anti-IR	Explicit opposition to the Islamic Republic; support for revolutionary movements or criticism of leadership.
K	Kurdish Identity	Identifies as Kurd or highlights challenges faced by Kurd people; may use Kurdish language.
S	Monarchist	Supports monarchy or constitutional monarchy; references Pahlavi family or phrases like "long live the Shah."
I	Israeli Affiliation	Mentions Israel; may identify as Israeli or be linked to Israeli media or politics.
H	Human Rights	Mentions "Human Rights" in bio; signals advocacy for related issues.
B	Baluch Identity	Identifies as Baluch or highlights concerns of the Baluch people.
M	MEK	Refers to or affiliates with the People's Mojahedin Organization of Iran (MEK).
P	Pro-IR	Supports the Islamic Republic or its leadership.

Table 1: Observed user classes (A–P) confidently detectable based on user bios.

on their own. For instance, Retweet-BERT achieved macro-F1 scores above 97% on U.S. political datasets by embedding users based on retweet diffusion patterns and profile text. He et al. (2024) further demonstrated that injecting graph structure into transformer-based models enables better detection of mixed-ideology communities, uncovering ideological gradients that text-only models are not able to detect.

Contrastive Learning and Text Representation in NLP. Within NLP, contrastive learning has emerged as a powerful strategy for learning sentence or user embeddings, especially in unsupervised or semi-supervised contexts. DeCLUTR (Giorgi et al., 2021) and CLEAR (Wu et al., 2020) combined MLM with a contrastive loss to encourage semantically similar inputs (e.g., paraphrases or augmentations) to have nearby embeddings. These methods significantly improved downstream classification and clustering performance. Our work builds on this line of research by implementing a joint *MLM+Contrastive* objective to fine-tune Persian language models for user embedding, followed by graph-based refinement using a GAT (Veličković et al., 2018) framework. This approach allows us to encode textual ideology signals and retweet relational structure in a unified representation space, without relying on extensive labeled data. To our knowledge, this represents the first attempt to apply this modeling strategy to the Iranian Twittersphere, during a major political unrest.

Domain Background. While most prior work in computational ideology detection has focused on Western settings, a growing body of research is examining political discourse on Persian Twitter. Kermani & Rasouli (2022) mapped the retweet network during the 2017 Iranian presidential election, identifying three main ideological factions (reformists, conservatives, and the diaspora) despite the lack of formal party structures. Azadi & Mesgaran (2021) quantified the reach of regime-supporting vs. dissident accounts, showing that dissidents were rapidly gaining influence despite the state's disproportionate presence. Rahmati et al. (2022) introduced a dataset of tweets from pro- and anti-regime figures and showed that deep learning models could infer ideological stance from tweets. However, these studies either relied on manual annotation, supervised classification, or isolated analysis of content or network. In contrast, our work develops an unsupervised framework tailored to the fluid, non-partisan, complex political environment of Iran. By jointly modeling language and network structure, we capture overt and covert ideological signals, offering a scalable approach to mapping ideological landscapes in low-label, high-uncertainty settings.

3 Dataset

In response to the political unrest in Iran, a surge of ideologically charged hashtags began appearing on Persian Twitter starting in September 2022. In this work, we leverage the dataset proposed in earlier works (Omidi Shayegan et al., 2024), which utilized the Twitter Research API to collect live tweets from October 18, 2022, to January 11, 2023. The tweet-gathering process was guided by 26 initial hashtags—both in Persian and English—carefully selected with regard to the Iranian political landscape. In total, the dataset comprises 231 million tweets from approximately 3.9 million unique users.

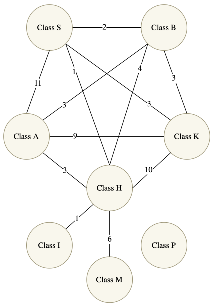


Figure 2: The overlaps of classes (true positive multi-classes). Weight of each edge indicates the number of users with both labels.

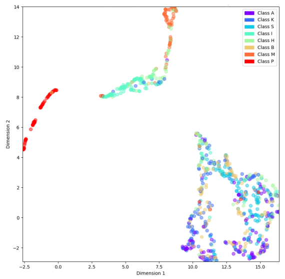


Figure 3: 2D node representation of ground truth labeled users in the node space.

Graph Building. We first utilize the tweets and retweets in the dataset to: (1) construct a retweet graph where users are nodes and directed edges (*retweetee* → *retweeter*, encoding the flow of information) indicate retweet actions, and (2) create a text corpus for each user as their node features.

The initial raw graph consisted of 3.1M nodes and 196.8M directed edges. We leverage temporal information, as the dataset provides timestamps for all tweets and retweets, enabling us to assign a precise timestamp to each edge.

When this directed multi-graph was converted into an undirected simple graph (by collapsing directionality and removing multiple edges between node pairs) the resulting graph retained only 57.9M unique edges. This implies that approximately 70% of the original edges represented repeated interactions between the same node pairs.
After applying the pruning strategy described in detail in Section 4, the resulting graph was substantially smaller, consisting of 148K nodes and 33.8M edges.

Annotation of Validation Set. We annotate a set of users to evaluate the proposed unsupervised method. This annotation was done based on self-identified information in users' profile bios. We started by a label-agnostic approach, not knowing what specific categories we would encounter in this dataset. The first step was to identify the ideological leanings present within the data. We focused on the top users (i.e., those with the highest number of retweets) and examined whether these users self-identified with any ideologies.

We filtered the users based on their retweet counts ($\geq$45K retweets, N=597) and manually reviewed the bios of these users to check for any direct mentions of ideology. Approximately 56% of them explicitly stated their ideology or referenced specific groups in their bios. After manually extracting the self-identified ideologies of the top users, we kept only the ideologies that were mentioned by a notable number of users and created corresponding classes for them, which corresponds to eight categories. In this context, "ideologies" mainly refer to their stance in a political event. For example, some users were particularly concerned with the human rights situation and cared enough to mention this stance in their profile bio. We grouped them into the "Human Rights" class (Class H). Although "Human Rights" is not strictly a political ideology, it represents a stance, point of view, or political concern.

Table 1 summarizes the observed categories, which we were able to confidently label based on the information expressed in user bios. [1] For each class, by going through the corresponding bios, we gathered keywords and identifiers what users commonly used to indicate their ideologies. These identifiers could be hashtags, phrases, or words related to their ideology. This allowed us to extend the labeled set by looking for the identifiers in users' bios. After the users with identifiers were automatically selected, we filtered users with at least 100 retweets and labeled 150 top retweeted users per ideology, if available.

[1] We acknowledge that there might be subgroups within these categories, specifically within the Class A and P. However, our annotation strategy is not able to uncover the nuanced differences among those subgroups. This categorization is not exhaustive and is identified merely based on observed and confidently detectable groups to evaluate our computational method. Therefore, it is not valid for any other purposes.

For seven major ideological groups (excluding the "human rights" category , Class H , since their identifiers were very strong already), we manually validated these automatic labels, retaining only the correct ones. There were 52 cases out of 915, where more than two labels applied to the user. For those cases, we devised a classification hierarchy prioritizing minority and less-represented groups. For users with multiple labels, only the highest-priority label was retained. Finally, 32 of the labeled users were removed during the pruning process of the graph, as they no longer existed in the final graph after the preprocessing stage. We obtained a labeled dataset consisting of 883 users, each assigned a single label, representing the respective ideological groups mentioned earlier.

4 Method

Our method consists of several steps: (1) preprocessing the textual data to create a corpus for each user; (2) fine-tuning ParsBERT using MLM+CL objective; (3) embedding each user's corpus with the fine-tuned model; (4) constructing and preprocessing the retweet graph; (5) assigning text embeddings as node features; and (6) training a GAT with a Link Prediction (LP) objective. The following sections detail each step.

4.1 User Text Embedding

Each user is represented as a node in the graph. In GAT, nodes require initial feature representations, which are propagated through edges to learn contextualized embeddings. To provide these features, we assign text embeddings to each user node, derived from their associated textual data. We make the corpuses by combining five of their randomly selected original tweets, the user's five most popular retweets, and their biography information.

Prior work has shown that ParsBERT (Farahani et al., 2021) demonstrates strong performance in capturing the political landscape of Iranian Twitter (Omidi Shayegan et al., 2024). We fine-tune ParsBERT by combining Masked Language Modeling (MLM) and Contrastive Loss (CL), referred to as MLM+CL in this study. The MLM is trained on a corpus comprising tweets and retweets from 500 random users. Additionally, for the CL, we construct pairs of tweets labeled based on the ideology of their authors: pairs are similar if the authors share the same ideology, and dissimilar otherwise. We combine MLM and CL using a weighted sum. The contrastive loss is:

$$L_{CL} = \sum_{k=1}^{B} \left[(1 - S_k)\frac{1}{2}d_k^2 + S_k\frac{1}{2}\max(0, m - d_k)^2 \right] \tag{1}$$

Here, S_k is the binary similarity label for pair k, where $S_k = 0$ if the tweets share the same label and $S_k = 1$ otherwise; d_k denotes the embedding distance between the tweets in pair k; m is a margin parameter tuned through hyperparameter optimization; and B is the batch size. The contrastive loss L_{CL} penalizes dissimilar pairs whose embedding distance is smaller than the margin m, and encourages similar pairs to have minimal distance.

The total loss for the training step is calculated by combining the contrastive loss L_{CL} and the MLM loss L_{MLM} through a weighted average. The weight K is a hyperparameter that is optimized during the hyperparameter tuning phase. The total loss for this training step is defined as: $L_{MLM+CL} = (K)L_{CL} + (1 - K)L_{MLM}$.

4.2 Graph Preprocessing

Once the graph was created as described in Section 3, we preprocessed the graph. This stage involves several steps. We first remove self-loops and discard isolated nodes. Next, we prune the edge list by removing approximately 90% of the nodes with the lowest out-degree. Pre-computed user embeddings are then attached to corresponding nodes, and for users without embeddings, we assign the average embedding of all available users.

Batching. Given the large size of the graph, processing it in its entirety would demand an immense amount of memory, making it impractical to handle all at once. Therefore, the graph must be split into batches before being passed to the GAT. Since the retweets have timestamps, the temporal information can be used to split the graph into smaller timeframes, using each timeframe as a batch. The size of each batch is fixed while progressing in time.

In our LP task (detailed in Section 4.3), positive links are actual edges present in the batch and negative links are generated by sampling from node pairs not forming edges in the batch. The ratio of existing to non-existing edges in our graph is 1:330, so a 1:1 sampling ratio is unsuitable for LP. We experimented with different ratios within the limits of our computational resources and adopted a 1:20 ratio for training. Finally, for each batch, the edge sets are partitioned into mutually exclusive training, validation, and test subsets.

4.3 Graph Neural Network Link Prediction

Once the graph, its features, and the edge splits are prepared, they are passed to the neural network. We use GAT to learn node representations (Veličković et al., 2018), using text embeddings as input features propagated through the graph via message passing. We train the model using a link prediction (LP) objective, where node labels are not required. Instead, edges serve as binary supervision: existing edges are treated as positive samples, and non-existent edges as negative samples. The model is optimized to maximize cosine similarity between connected node pairs and minimize it for unconnected pairs. Our architecture consists of four GAT layers followed by a fully connected MLP as the final layer.[2]

5 Experiments

To understand the structure and properties of the learned node representation space, we conduct classification, clustering, and 2D visualization of the embeddings. As baselines, we include "Random", which assigns each user a randomly initialized vector with the same size as the text embeddings and "Fixed", which assigns all users the same feature vector, initialized as the average of all user text embeddings.

5.1 Reducing Dimensionality

To enable 2D visualization of the learned representations, we computed low-dimensional projections and evaluated their effectiveness as well. We initially experimented with both t-SNE and UMAP, but found that t-SNE failed to preserve the structural properties of the data. As a result, we used UMAP for all dimensionality reduction experiments. The impact of this reduction is reported in Table 2 under the "Reduced" and "Not Reduced" settings.

5.2 Classification for Evaluation

We use an MLP classifier to evaluate the representations with labeled data. Since the graph training procedure is unsupervised and based on LP, we do evaluation through performance in a downstream classification task. To ensure robust results, each model (i.e., each GAT and its attached text embedding combination) is trained five times and evaluated independently.

For each instance (i.e., every time the model is trained) we pass the node representations through MLP classifiers with different hidden layers of sizes $[2, 4, ..., 128]$, training a classifier five times for each size. Resulting in 35 accuracy measurements. Then, the hidden size with the best average validation performance is used to evaluate the instance ten times on the test set. The average of these ten scores is reported as the instance-level accuracy. Finally, we compute the mean and standard deviation across the five instances to report the model's overall performance and stability.

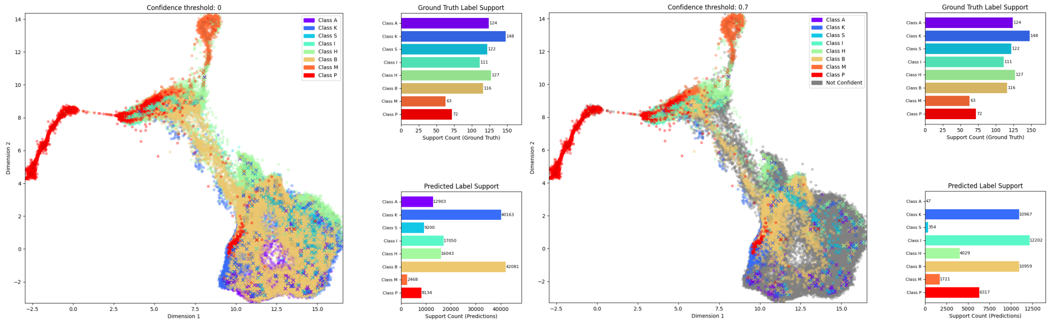
5.3 Clustering User Representations

Once we obtained the user representations, we aimed to determine whether they exhibited meaningful clusters in our test set. We anticipated that these clusters would align with the classes identified earlier during the labeling process, as the most influential users (i.e., those with the greatest influence in the graph) had self-identified with these groups, and a significant number of users in the network were retweeting them. Consequently, we attempted to cluster the user representations into eight clusters, using KMeans algorithm. To evaluate the alignment between these clusters and the classes, we applied three methods:

[2] The hyperparameters are *gnn_classes=32*, *num_heads=3*, *mlp_classes=8*, and *gnn_out_channels=16*. For the optimization we use *learning_rate*≃*0.050*, *warmup_steps*=20, step-wise learning rate decay with *gamma=0.999* and early stopping.

Method	Classification acc %		Clustering acc %		Clustering ARI	
	Reduced	Not Reduced	Reduced	Not Reduced	Reduced	Not Reduced
MLM+CL	44.9±4.6	**52.5±2.5**	**45.1±2.1**	42.4±2.4	0.177±0.010	0.195±0.023
MLM	45.2±1.0	50.4±3.0	44.8±4.2	41.3±1.6	0.197±0.026	0.162±0.007
Random	39.9±1.5	42.7±1.2	38.3±2.1	37.2±3.0	0.134±0.015	0.130±0.011
Fixed	30.5±2.4	13.8±0.9	33.5±2.5	32.7±1.7	0.085±0.020	0.083±0.012

Table 2: Classification and clustering results for graph representations, with and without dimensionality reduction.



(a) Confidence Threshold = 0

(b) Confidence Threshold = 0.7

Figure 4: Comparison of predicted and ground-truth labels in the 2D embedding space at different confidence thresholds. Each subfigure shows predicted labels (O) and ground-truth labels (X), with corresponding label distributions.

1. **Cluster-to-Class Alignment:** To evaluate clustering performance, we use an assignment algorithm (Crouse, 2016) to find the optimal one-to-one mapping between discovered clusters and ground-truth classes. This mapping allows us to interpret each cluster as representing a specific class, enabling the computation of clustering accuracy with the ground truth class labels.
2. **Cluster Centers as Class Anchors:** We analyzed the cluster centers by identifying the nearest users in the embedding space and inspecting their ground-truth labels. This analysis helps reveal if certain classes are concentrated around specific cluster centers, offering insight into the correspondence between clusters and true class distributions.
3. **Adjusted Random Index (ARI)**: We report the ARI (Chacón & Rastrojo, 2023) for the clusters by comparing them with the ground truth labels. ARI is a similarity score between -0.5 and 1.0. ARI = 1 means the cluster and the ground truth labels are the same.

6 Results and Discussions

To assess the quality of the user representations, we evaluate them in supervised classification and unsupervised clustering tasks mentioned in Section 4. Table 2 presents the results for models trained with MLM and MLM+CL, with and without dimensionality reduction.

6.1 Classification Accuracy

The highest classification accuracy is achieved by the MLM+CL model without dimensionality reduction, reaching 52.5%, indicating that contrastive learning effectively structures the embedding space for supervised tasks. The baseline MLM model also performs well (50.4%), showing that MLM captures meaningful semantic features even without supervised data. As a reference, GAT trained with randomly initialized node features achieves 42.7% accuracy—substantially above the 12.5% chance level for eight classes.

Furthermore, Table 2 shows that the graph achieves relatively good performance even when initialized with random node features (42.7% ± 1.2 %), suggesting that the graph structure alone contains meaningful information for ideology detection. This table also shows that the Random features perform better than initializing all of the nodes with the same vector.

6.2 Confidence of Predictions

To assess the confidence of our MLP model's predictions, we applied varying thresholds to the softmax output, retaining only those predictions with the highest probability. This filtering allows us to examine how prediction distributions shift as confidence increases.

Model									
MLM+CL	Cluster	0	1	2	3	4	5	6	7
	Class	M	S	P	A&K	I	P	S&A	H
	Overlaps	-	-	-	9	-	-	11	-
MLM	Cluster	0	1	2	3	4	5	6	7
	Class	H&M	A&S	I	K	H	H	P	S&B
	Overlaps	6	11	-	-	-	-	-	2

Table 3: Comparison of cluster centers as class anchors between MLM+CL (limited supervision) and MLM (fully unsupervised) models. In the case of multiple classes, the number of overlaps of the classes is provided.

As shown in Figures 1 and 4, raising the confidence threshold alters the distribution of predicted labels, revealing that some classes are predicted with higher certainty than others. In particular, Figure 4b shows that labels such as Class A and Class S are less confidently predicted. This is likely due to substantial overlap between these classes, making them more difficult for the model to distinguish.

6.3 Clustering Accuracy and ARI

We assessed clustering alignment using the linear sum assignment algorithm. The MLM+CL model outperforms by achieving a clustering accuracy of 45.1% ± 1.1% compared to 44.8% ± 4.2% for the MLM model. This suggests that even without label supervision on the evaluation layer, the clusters obtained with KMeans can help us group the ideologies.

Interestingly, dimensionality reduction improves clustering performance across both models, indicating that projecting the embeddings to a lower-dimensional space may help disentangle ideological groupings in the absence of labels. Notably, the best "fully-unsupervised" result is achieved by the reduced MLM model, reaching 44.8% clustering accuracy—despite having no access to labels during training or evaluation. This performance is specifically important to us because it highlights the utility of our method for exploring political spaces where ground-truth labels are limited or unavailable.

The highest ARI is achieved by the MLM model with dimensionality reduction (0.197 ± 0.026), suggesting that while reduction may slightly harm classification accuracy, it can improve clustering alignment by filtering out noisy dimensions. This pattern is consistent with the clustering accuracy results. The MLM+CL model without reduction follows closely (0.195 ± 0.023), and given the overlapping standard errors, the difference is not statistically significant. Notably, the MLM+CL model performs better without reduction (0.195 ± 0.023 vs. 0.177 ± 0.010), indicating that contrastive learning yields more structured embeddings without the need for further compression.

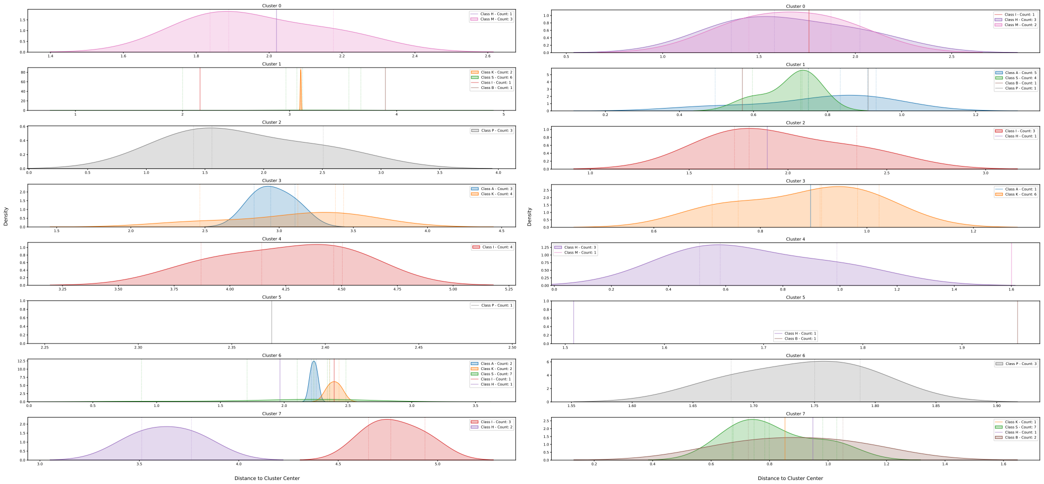
6.4 Effect of Dimensionality Reduction

Dimensionality reduction shows mixed effects. While it generally lowers classification performance, likely due to loss of fine-grained discriminative features, it improves clustering accuracy across all models and increases ARI in the MLM setting. This suggests that reduction may help by smoothing the embedding space and emphasizing broader ideological groupings. In contrast, the MLM+CL model achieves its highest ARI without reduction, showing that contrastive supervision benefits from preserving high-dimensional structure. The classification results further support this, as the models perform best without reduction—likely because contrastive signals are more effectively captured by the MLP in the original embedding space.

6.5 Spatial Distribution of Ideologies in Representation Space

To better understand the ideological landscape, we reduced the node representations to 2D. Figure 3 shows the labeled data, based on the best-performing model. Classes with high overlap, such as Classes A, K, and S (see Figure 2), appear closely positioned, making them harder to distinguish. This proximity likely contributes to model's lower confidence when labeling Classes A and S, which might be predicted interchangeably.

In contrast, Class P forms a well-separated cluster, making it easier for the model to classify. This is supported by Figures 1 and 4, where its support remains stable even at higher confidence thresholds. We also observe that Classes I, H, and M form a distinct cluster, separate from the denser region composed of Classes A, S, K, and B. Upon closer inspection,



(a) Using MLM+CL; Limited supervision.

(b) Using MLM; Fully unsupervised.

Figure 5: Distance distribution of the closest labeled users to each cluster center. Dotted lines indicate labeled users, with density estimates overlaid.

we find that the core cluster mainly consists of anonymous users, likely representing the broader public, while the isolated cluster is composed mostly of known political figures or activists. The instance used for the visualizations achieved a classification accuracy of 60.9% while not reduced and 49.6% on the 2D reduced representations.

6.6 Cluster Centers as Class Anchors

To evaluate the alignment the discovered clusters with ground truth classes, we examined if each class could be mapped to a distinct cluster. Figure 5a shows the distance distribution of the 5% closest labeled users to each cluster center. Based on this proximity analysis, we derived the cluster-to-class mapping presented in Table 3. For the MLM+CL model, 7 out of 8 classes align with a distinct cluster. In the case of the MLM model, all 8 classes are represented, though with lower confidence due to a higher degree of label overlap within clusters. We can also see that the classes which have more overlaps in Figure 2 tend to be harder to distinguish for the model.

7 Conclusion and Future Work

In this work, we presented an unsupervised framework for mapping political ideologies in low-label environments with less well-defined ideologies, by combining contrastive language modeling with graph-based representation learning. Applied to Persian Twitter data during the 2022 Mahsa Amini protests, our method revealed eight latent ideological factions and demonstrated strong alignment with expert-verified ground truth labels. Notably, the spatial layout of the discovered clusters mirrored known political cleavages, offering a unique lens into Iran's political landscape—one that lacks well-defined party structures.

Our results show that combining textual content and retweet network structure can provide valuable insights into user alignment without relying on predefined labels. This is especially relevant where party systems are informal , and large-scale labeled data is unavailable.

Nonetheless, the approach has certain limitations. Access to user-level interaction data is becoming more restricted due to changes in platform APIs, making future data collection less straightforward. In addition, our method emphasizes prominent users and overt ideological cues, which may limit its ability to detect more subtle or emerging perspectives.

Future research could extend this framework to other linguistic and political settings, adapting it to different sources of user interactions and addressing current constraints on data accessibility. Incorporating temporal dynamics or additional modalities (such as images or hyperlinks) may also improve the resolution of detected ideological patterns. More broadly, unsupervised approaches like this can support scalable, data-driven analysis of political discourse in varied online environments.

References

P. Azadi and M. B. Mesgaran. The Clash of Ideologies on Persian Twitter, June 2021. URL https://iranian-studies.stanford.edu/iran-2040-project/publications/clash-ideologies-persian-twitter.

Pablo Barberá. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76–91, January 2015. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpu011. URL https://doi.org/10.1093/pan/mpu011. Publisher: Cambridge University Press.

José E. Chacón and Ana I. Rastrojo. Minimum adjusted Rand index for two clusterings of a given size. *Advances in Data Analysis and Classification*, 17(1):125–133, March 2023. ISSN 1862-5355. doi: 10.1007/s11634-022-00491-w. URL https://doi.org/10.1007/s11634-022-00491-w.

Wei Chen, Xiao Zhang, Tengjiao Wang, Bishan Yang, and Yi Li. Opinion-aware Knowledge Graph for Political Ideology Detection. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 3647–3653, Melbourne, Australia, August 2017. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/510. URL https://www.ijcai.org/proceedings/2017/510.

David F. Crouse. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, August 2016. ISSN 0018-9251. doi: 10.1109/taes.2016.140952. URL http://ieeexplore.ieee.org/document/7738348/. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

Bernard Enjolras and Andrew Salway. Homophily and polarization on political twitter during the 2017 Norwegian election. *Social Network Analysis and Mining*, 13(1): 10, December 2022. ISSN 1869-5469. doi: 10.1007/s13278-022-01018-z. URL https://doi.org/10.1007/s13278-022-01018-z.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. ParsBERT: Transformer-based Model for Persian Language Understanding. *Neural Processing Letters*, 53(6):3831–3847, December 2021. ISSN 1573-773X. doi: 10.1007/s11063-021-10528-4. URL https://doi.org/10.1007/s11063-021-10528-4.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations, May 2021. URL http://arxiv.org/abs/2006.03659. arXiv:2006.03659 [cs].

Zihao He, Ashwin Rao, Siyi Guo, Negar Mokhberian, and Kristina Lerman. Reading Between the Tweets: Deciphering Ideological Stances of Interconnected Mixed-Ideology Communities. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1523–1536, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.104/.

Julie Jiang, Xiang Ren, and Emilio Ferrara. Retweet-BERT: Political Leaning Detection Using Language Features and Information Diffusion on Social Networks, April 2023. URL http://arxiv.org/abs/2207.08349. arXiv:2207.08349 [physics].

Kristen Johnson, I-Ta Lee, and Dan Goldwasser. Ideological Phrase Indicators for Classification of Political Discourse Framing on Twitter. In Dirk Hovy, Svitlana Volkova, David Bamman, David Jurgens, Brendan O'Connor, Oren Tsur, and A. Seza Doğruöz (eds.), *Proceedings of the Second Workshop on NLP and Computational Social Science*, pp. 90–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2913. URL https://aclanthology.org/W17-2913/.

Hossein Kermani and Fatemeh Rasouli. Protesting is Not Everything: Analyzing Twitter Use During Electoral Events in Non-democratic Contexts. *Journal of Digital Social Research*, 4(4):22–51, November 2022. ISSN 2003-1998. doi: 10.33621/jdsr.v4i4.116. URL https://jdsr.se/ojs/index.php/jdsr/article/view/116. Number: 4.

Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. For Women, Life, Freedom: A Participatory AI-Based Social Web Analysis of a Watershed Moment in Iran's Gender Struggles, July 2023. URL http://arxiv.org/abs/2307.03764. arXiv:2307.03764 [cs].

Sahar Omidi Shayegan, Isar Nejadgholi, Kellin Pelrine, Hao Yu, Sacha Levy, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. An Evaluation of Language Models for Hyperpartisan Ideology Detection in Persian Twitter. In Atul Kr. Ojha, Sina Ahmadi, Silvie Cinková, Theodorus Fransen, Chao-Hong Liu, and John P. McCrae (eds.), *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pp. 51–62, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.eurali-1.8/.

Kellin Pelrine, Anne Imouza, Zachary Yang, Jacob-Junqi Tian, Sacha Lévy, Gabrielle Desrosiers-Brisebois, Aarash Feizi, Cécile Amadoro, André Blais, Jean-François Godbout, and Reihaneh Rabbany. Party Prediction for Twitter, August 2023. URL http://arxiv.org/abs/2308.13699. arXiv:2308.13699 [cs].

Marco Pennacchiotti and Ana-Maria Popescu. A Machine Learning Approach to Twitter User Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):281–288, 2011. ISSN 2334-0770. doi: 10.1609/icwsm.v5i1.14139. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14139. Number: 1.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 729–740, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1068. URL https://aclanthology.org/P17-1068.

Ali Rahmati, Ehsan Tavan, and Mohammad Ali Keyvanrad. Predicting Content-based Political Inclinations of Iranian Twitter Users Using BERT and Deep Learning. *AUT Journal of Mathematics and Computing*, (Online First), December 2022. doi: 10.22060/ajmc.2022.21895.1120. URL https://doi.org/10.22060/ajmc.2022.21895.1120.

Miguel Ángel Rodríguez-García, Soto Montalvo Herranz, and Raquel Martínez Unanue. URJC-Team at PoliticEs 2022: Political Ideology Prediction using Linear Classifiers. 3202, September 2022. URL https://ceur-ws.org/Vol-3202/politices-paper11.pdf.

Petter Törnberg. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning, April 2023. URL http://arxiv.org/abs/2304.06588. arXiv:2304.06588 [cs].

P Veličković, A Casanova, P Liò, G Cucurull, A Romero, and Y Bengio. Graph attention networks. 2018. doi: 10.17863/CAM.48429. URL https://www.repository.cam.ac.uk/handle/1810/301348. Publisher: OpenReview.net.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. CLEAR: Contrastive Learning for Sentence Representation, December 2020. URL http://arxiv.org/abs/2012.15466. arXiv:2012.15466 [cs].

Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. TIMME: Twitter Ideology-detection via Multi-task Multi-relational Embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2258–2268, August 2020. doi: 10.1145/3394486.3403275. URL http://arxiv.org/abs/2006.01321. arXiv:2006.01321 [cs, stat].

Hao Yu, Zachary Yang, Kellin Pelrine, Jean Francois Godbout, and Reihaneh Rabbany. Open, Closed, or Small Language Models for Text Classification?, August 2023. URL http://arxiv.org/abs/2308.10092. arXiv:2308.10092 [cs].