Unsupervised Image Classification by Ideological Affiliation from User-Content Interaction Patterns

Xinyi Liu, Jinning Li, Dachun Sun, Ruijie Wang, Tarek Abdelzaher
Department of Computer Science, University of Illinois at Urbana Champaign
201 N. Goodwin Ave., Urbana, IL 61801

Matt Brown, Anthony Barricelli, Matthias Kirchner, Arslan Basharat
Kitware, Inc.
1712 Route 9, Clifton Park, NY 12065

Abstract

The proliferation of political memes in modern information campaigns calls for efficient solutions for image classification by ideological affiliation. While significant advances have recently been made on *text* classification in modern natural language processing literature, understanding the political insinuation in *imagery* is less developed due to the hard nature of the problem. Unlike text, where meaning arises from juxtaposition of tokens (words) within some common linguistic structures, image semantics emerge from a much less constrained process of fusion of visual concepts. Thus, training a model to infer visual insinuation is possibly a more challenging problem. In this paper, we explore an alternative unsupervised approach that, instead, infers ideological affiliation from *image propagation patterns* on social media. The approach is shown to improve the F1-score by over 0.15 (nearly 25%) over previous unsupervised baselines, and then by another 0.05 (around 7%) in the presence of a small amount of supervision.

1 Introduction

The paper addresses the challenge of *unsupervised* classification of digital images (e.g., memes) by ideological affiliation. While it is common to describe ideology as a scale from liberal to conservative (or left to right), this is *not* the definition we adopt in this paper. Ideological divides can arise for many reasons including religious differences, disagreements on historical context, and incompatibilities in the ranking of moral values (e.g., fairness versus loyalty), among many others. The classification algorithm aims to distinguish visual content of two conflicting sides of an ideological divide without a prior understanding of the nature of the underlying divide. The work is motivated by the proliferation of memes and other visual aids in marketing (Levinson 2001), social movements (Mina 2019), and political campaigns (Martínez-Rolán, Piñeiro-Otero, and others 2016), thus generating interest in automating the analysis of ideological and semantic connotations of images (Kiela et al. 2020; Theisen et al. 2021). Automating machine interpretation of such visual content as memes, however, is arguably a harder challenge than automating scene understanding (Naseer, Khan, and Porikli 2018; Grant and Flynn 2017; Xue, Fang, and Zhang 2018) or text understanding (Brown et al. 2020) due to the much less structured nature of the underlying creative process behind meme generation. New memes often utilize unique and subtle juxtapositions of concepts, whose novel nature challenges self-supervised models trained on existing patterns. This work skirts the problem by exploring an *alternative* unsupervised approach to image interpretation. Namely, we use visual similarity and interaction patterns between users and (families of) images to classify images by ideological leaning.

The idea of exploiting user-content interactions for unsupervised content classification has previously been explored by the authors in the context of classifying text posts (Al Amin et al. 2017; Li et al. 2022; Yang et al. 2020b). It stems from the observation that users interact with content that matches their beliefs. Thus, in an ideological clash, characteristic of today's growing polarization (Lelkes 2016; Petri and Biedenkopf 2021), different groups of users interact with different content items depending on their ideological leaning. While the classification algorithm need not know the ideological leaning of users ahead of time, it can cluster content by the way it propagates, thereby separating content into ideologically aligned categories.

This paper explores the logical follow-up question: can the same approach be successfully applied to images? How well will it work, and what design parameters are relevant to improving its performance? The underlying reason why a classification approach that relies on observing user-content interactions is of interest lies in that social media (the main digital media where memes propagate in the first place) associate posts with sources. This association allows one to assess ideological similarity among posts by assessing similarity in the sets of users (i.e., sources) who propagate them. This similarity measure is entirely independent of user identity, which can in fact be kept anonymous to the classifier. The unsupervised algorithm, therefore, does not exploit user features (and passed the IRB approval process).[1]

More specifically, in this paper, we use a modified variational graph auto-encoder, originally proposed for improving feature disentanglement in the latent space (Shao et al.

[1]This work passed the ethical approval process by the institutional review board (the IRB) and was deemed exempt.

2020) and subsequently adapted for (ideological) belief representation learning (Li et al. 2022). The adapted version, called InfoVGAE, is applied to a bipartite graph of sources and *visual assertions*, each representing a group of very similar images or memes. The graph is mapped into a latent space where visual assertions of a similar ideological leaning are clustered together. We evaluate the approach based on image dataset we collected from online controversies over the Russia-Ukraine war. The evaluation shows that the approach is successful at separating two clusters of ideologically distinct images; one represents pro-Kremlin imagery and the other pro-Ukraine imagery.

The rest of the paper is organized as follows. Section 2 reviews some background and presents our InfoVGAE-based image embedding algorithm. Section 3 describes the data set used for evaluation and presents evaluation results. Section 4 discusses limitations of the current approach and proposes avenues for future work. Section 5 summarizes related work. The paper concludes with Section 6.

2 Image Embedding and Classification

This section takes the reader through the step-by-step process of (i) identifying visual themes in messages, called *visual assertions*, (ii) constructing the user-assertion interaction graph from social media data, and (iii) performing self-supervised embedding on the resulting graph into a lower-dimensional ideological space. Semi-supervised extensions are also described.

2.1 Identifying Visual Assertions

In order to identify meaningful user-content interactions, we first need an approach for recognizing (and grouping together) content items that have very similar semantics. Each group of nearly identical items represents essentially the same intended message. Identifying such groups makes it easier to learn how a community propagates similar items.

We call each semantically-similar group of items an *assertion*. In this case, we are referring specifically to images. Thus, for the purposes of this work, we define a *visual assertion* as a set of images that share a high degree of similarity. We further assume that a given image is associated with at most one visual assertion, although extensions to many-to-many mappings are possible. Different definitions of image similarity will lead to different interpretations of what a visual assertion represents. In this work, we focus on near-duplicate images (i. e., sets of images that were very likely derived from each other through operations like resizing, cropping, recompression, color adjustment, or adding text-overlays, amongst many others). Examples of two visual assertions are shown in Figure 1. Specifically, we use a keypoint-based approach to identify near-duplicate images, where we declare a pair of images as near-duplicates if there is a sufficient number of matching ORB (Rublee et al. 2011) keypoints across both images, and if the affine image-to-image mapping estimated with RANSAC (Fischler and Bolles 1981) from the set of detected keypoints is empirically sensible. We can then compile the set of visual assertions from the cluster graph obtained from all detected pairs of near-duplicates in the set of candidate images. We narrow down the set of candidate images prior to the near-duplicate detection by excluding pairs of images with a low cosine similarity in the CLIP (Radford et al. 2021) embedding space.[2]

2.2 Constructing the User-Image Interaction Graph

Given an algorithm for clustering similar images into visual assertions, described above, we construct a user-assertion interaction graph as follows:

- **Step 1:** Extract the user(s) who posted each individual image in the collected image dataset. This is typically a straightforward look-up of object metadata using the respective social network API.
- **Step 2:** Cluster the images based on the method proposed in Section 2.1. Represent each image cluster by a visual assertion node.
- **Step 3:** Represent each user by a user node. Link each user node to all visual assertion nodes to which the user contributed images.

After applying the above procedure, we can model the users and visual assertions by a *Bipartite Heterogeneous Information Network (BHIN)* (Sun and Han 2012) given by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the number of vertices, $|\mathcal{V}| = N$, is the sum of user and assertion vertices, and the number of edges, $|\mathcal{E}| = M$, is the number of user-assertion links, as is shown in Figure 2. There are two vertex types in the graph, users and visual assertions. In general, the number of edge types could be R, representing different operation types in the social network, such as posting, commenting, replying, etc. In our implementation, we use only one edge type that represents posting/reposting.

2.3 Unsupervised Embedding

We embed the user-assertion interaction graph, described above, into a lower-dimensional latent space using a version of *variational graph auto-encoders*, called InfoVGAE (Li et al. 2022). Each dimension in the latent space represents a different ideological leaning. Thus, in the case of a two-sided conflict, we embed the user-assertion interaction graph into a two-dimensional space. The loss function of the embedding algorithm (i.e., the criterion optimized by the placement of nodes in the latent space) encourages (i) placing pairs of nodes connected by an edge onto the same latent axis, and (ii) placing pairs of nodes with no common edge onto geometrically orthogonal axes. Thus, content propagated by largely different sets of users ends up mapped to different axes, offering the basis for separating different ideological leanings. Below, we review the basic mathematical background on vanilla VGAEs then describe the used InfoVGAE.

[2]CLIP itself is not suitable for near-duplicate detection as it is generally more broadly indicative of semantic similarity.

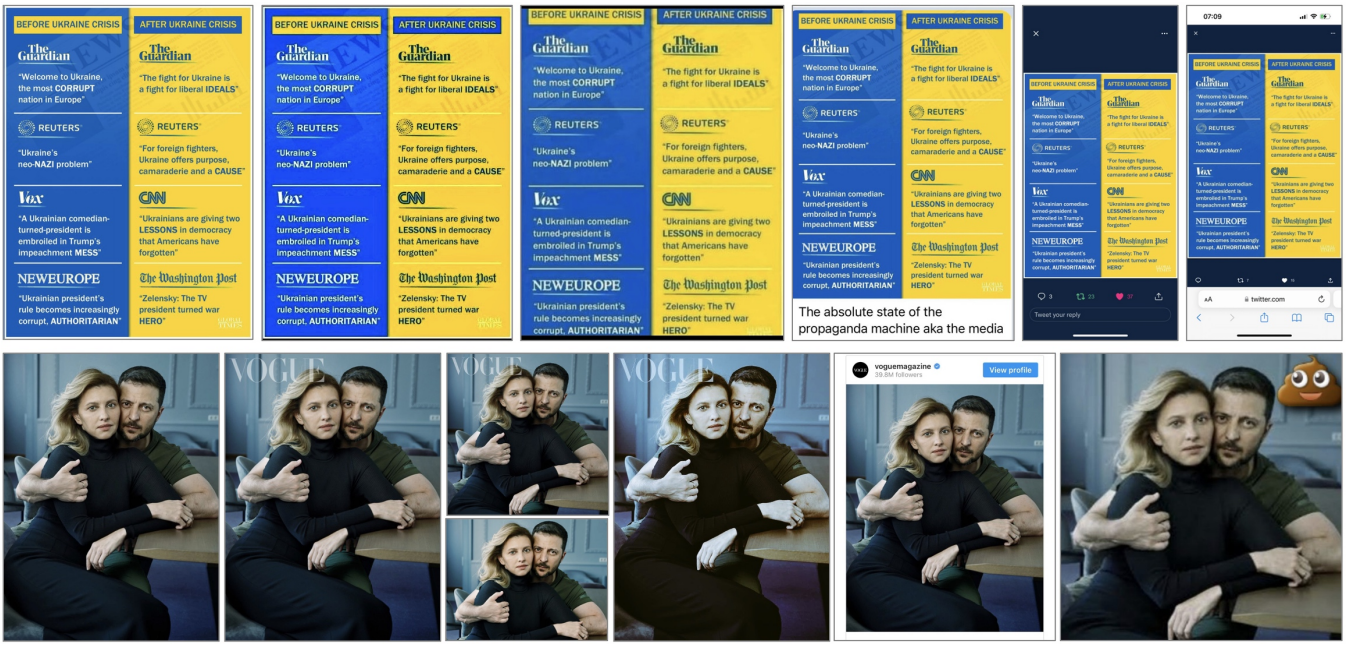


Figure 1: Representative examples of two identified near-duplicate visual assertions in the Russia-Ukraine image dataset.

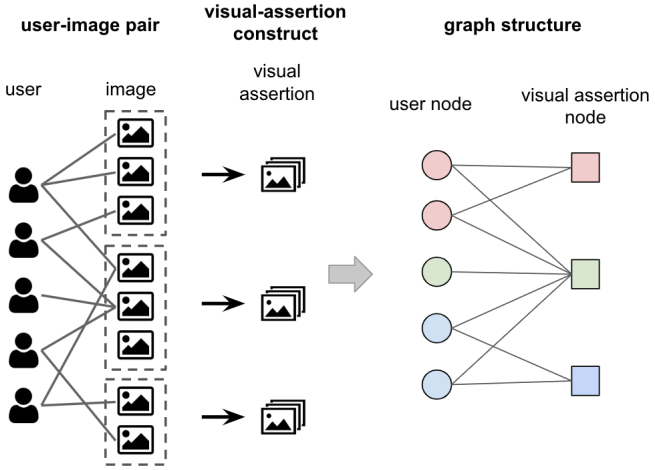


Figure 2: Three steps of converting user-image pairs to a Bipartite Heterogeneous Information Network (BHIN) user-image interaction graph.

VGAE Preliminaries: A variational graph auto-encoder (VGAE) (Kipf and Welling 2016) is an unsupervised algorithm designed to embed graph-structured data into a lower-dimensional latent space. Our VGAE accepts as input the undirected, unweighted graph, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, constructed in Section 2.2. The edges of $\mathcal{G}$ are represented by an adjacency matrix (with self-loops), denoted by $\boldsymbol{A}$. The nodes are represented by matrix $\boldsymbol{X} \in \mathbb{R}^{N \times F}$, called the input feature matrix, where F is the length of each node's feature vector. The typical VGAE consists of two parts, namely, the inference model (Encoder) and the generative model (Decoder), described below.

A. Inference Model (Encoder): The encoder exploits an L-layer *graph convolutional network* (GCN) as the neural network architecture. The hidden state of nodes is denoted by $\boldsymbol{E}^{(l)} \in \mathbb{R}^{N \times d_l}$, where d_l is the dimension of hidden state in the l^{th} layer. $\boldsymbol{E}^{(0)} = \boldsymbol{X}$ is the input feature. The per-layer computation within the GCN can be represented as $\boldsymbol{E}^{(l)} = GCN^{(l)}(\tilde{\boldsymbol{A}}, \boldsymbol{E}^{(l-1)})$, which receives the hidden state of the previous layer $\boldsymbol{E}^{(l-1)}$ and the normalized adjacency matrix $\tilde{\boldsymbol{A}}$ as input. More specifically, the GCN layer is formulated as

$$\boldsymbol{E}^{(l)} = \gamma\left(\tilde{\boldsymbol{A}}\boldsymbol{E}^{(l-1)}\boldsymbol{W}^{(l-1)}\right), \ (2 \leq l \leq L-1) \qquad (1)$$

To formulate the output of the encoder as variational latent space, the output of the last layer of the GCN represents its parameters as the mean and standard deviation vectors $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$. We have $\boldsymbol{\mu} = \tilde{\boldsymbol{A}}\boldsymbol{G}^{(L-1)}\boldsymbol{W}_{\boldsymbol{\mu}}^{(L-1)}$ and $\log \boldsymbol{\sigma} = \tilde{\boldsymbol{A}}\boldsymbol{G}^{(L-1)}\boldsymbol{W}_{\boldsymbol{\sigma}}^{(L-1)}$. The latent representation matrix $\boldsymbol{Z} \in \mathbb{R}^{N \times T}$ is then derived with the re-parameterization trick (Kingma and Welling 2013), with T as the dimension of the target latent space. We denote $\boldsymbol{z}_i$ as the latent vector of the i^{th} node. The inference model is defined as:

$$q(\boldsymbol{Z}|\boldsymbol{A}, \boldsymbol{X}) = \prod_{i=1}^{N} q(\boldsymbol{z}_i|\boldsymbol{A}, \boldsymbol{X}), \qquad (2)$$

where $q(\boldsymbol{z}_i|\boldsymbol{A}, \boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{z}_i|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ with $\mathcal{N}$ as the Gaussian Distribution.

B. Generative Model (Decoder): The generative model is given by the inner product between variables. It is formu-

lated as:

$$p(\boldsymbol{A}|\boldsymbol{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(\boldsymbol{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{z}_j), \tag{3}$$

where $p(\boldsymbol{A}_{i,j}|\boldsymbol{z}_i, \boldsymbol{z}_j) = \sigma(\boldsymbol{z}_i^T \boldsymbol{z}_j)$. $\boldsymbol{A_{i,j}}$ are the elements of $\boldsymbol{A}$ and $\sigma(\cdot)$ is the logistic sigmoid function. Finally, the VGAE model is trained to maximize the variational lower bound,

$$\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{X})}\left[\log p(\boldsymbol{A}|\boldsymbol{Z})\right] - D_{KL}\left[q(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{X})\|p(\boldsymbol{Z})\right] \tag{4}$$

InfoVGAE: While the vanilla VGAE model, described above, can be applied to learn the node representations from a general graph, we find it helpful to provide certain additional constraints on the learned representation to better separate content by propagation patterns. Specifically, we use a VGAE variant, called the Information-Theoretic Variational Graph Auto-Encoder (InfoVGAE) (Li et al. 2022). An InfoVGAE creates a non-negative latent space by replacing the standard Gaussian Distribution with a rectified Gaussian Distribution (Socci, Lee, and Sebastian Seung 1998). Thus, it models the posterior probability in Equation 2 as $q(\boldsymbol{z}_i|\boldsymbol{A},\boldsymbol{X}) \sim \mathcal{N}_+(\boldsymbol{z}_i|\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ where $\mathcal{N}_+$ is the rectified Gaussian distribution. This small modification has the profound effect that the only geometrically orthogonal latent representation becomes one where node embeddings are directly aligned with the axes. Thus, a loss function that favors orthogonality among non-interacting sets of items forces them to lie approximately on the axes, leading to an interpretable representation, where each axis maps differently-propagating content (and, hence, a different ideological leaning).

The InfoVGAE introduces two additional modifications to further improve the embedding by decreasing disentanglement between different dimensions in the latent space (so that each dimension can represent a different ideology).

First, it penalizes the correlation between different embedding axes, by jointly training a discriminator Φ to minimize the total correlation loss:

$$\mathcal{L}_r(\boldsymbol{Z}) = \mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z})}[log(\Phi(\boldsymbol{Z})) - log(1-\Phi(\boldsymbol{Z}))]. \tag{5}$$

Second, InfoVGAE explicitly controls the KL divergence with a PI-controller that manipulates a new parameter $\beta(t)$ (Shao et al. 2020). The final objective of the InfoVGAE is formulated as one of maximizing:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{Z}\sim q(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{X})}\left[\log p(\boldsymbol{A}|\boldsymbol{Z})\right] \\ &\quad - \beta(t) D_{KL}\left[q(\boldsymbol{Z}|\boldsymbol{A},\boldsymbol{X})\|p(\boldsymbol{Z})\right] - \lambda\mathcal{L}_r(\boldsymbol{Z})\end{aligned} \tag{6}$$

As we show in the evaluation section, the produced embedding directly separates the visual content into (two) clusters aligned with the different axes in the latent space. Each cluster corresponds to a different ideological leaning.

While the InfoVGAE is fully self-supervised, in the evaluation section, we additionally explore accuracy gains attained when the it is used in a semi-supervised manner, to leverage situations where some assertions are possible to label (by ideological leaning) ahead of time. In this scenario, we manually label some of the least popular visual assertions upfront. In the InfoVGAE training stage, since each latent axis corresponds to a different leaning, we augment the loss function with a regularization term that penalizes non-zero values of coordinates of the opposite (i.e., wrong) leaning for each labeled assertion, essentially forcing it to lie on a specific axis, as is shown in Figure 3. By doing so, we show that the embedding of other assertions is also improved.

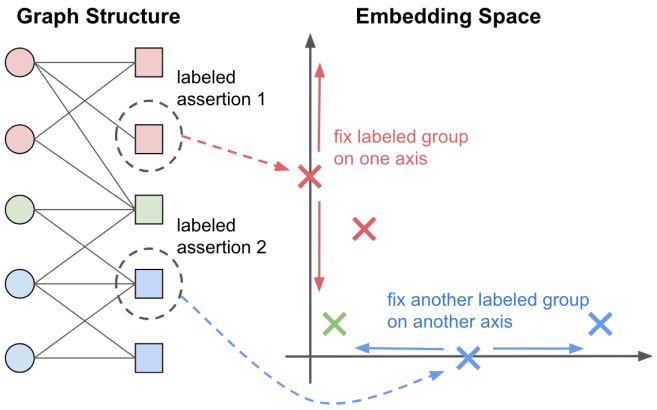


Figure 3: The left part of the figure is the constructed user-image BHIN, where several visual assertion nodes in each ideology group are human-labeled. The embedding of the labeled visual assertions in different ideology groups are fixed on different axis as is shown on the right.

2.4 Putting it Together

Figure 4 summarizes the methodology described in this paper. Specifically, we first identify visual assertions based on the similarity among images as discussed in Section 2.1. We then model user-assertion interactions by a bipartite graph, as described in Section 2.2. Finally, we feed the graph to the InfoVGAE for embedding, as described in Section 2.3 and perform ideological separation based on the embedding result.

3 Evaluation

In this section, we evaluate the performance of our proposed image ideology classification algorithm. Below, we describe the data set used for evaluation, the compared baselines, and the key performance results, respectively.

3.1 The Dataset

While the goal of this paper is to classify *images*, for ease of data collection and ground truth estimation, we exploited Twitter as a means to download media objects. No tweet text was used by the classification algorithm, however. More specifically, we collected media objects (images) posted on Twitter about the Russia-Ukraine war from 2022-05-01 to 2022-11-02 using a keyword API prompted separately once with pro-Kremlin then once with pro-Ukraine keywords. For the "pro-Ukraine" side, keywords were chosen that characterize the war as an act of aggression. For the "pro-Kremlin"

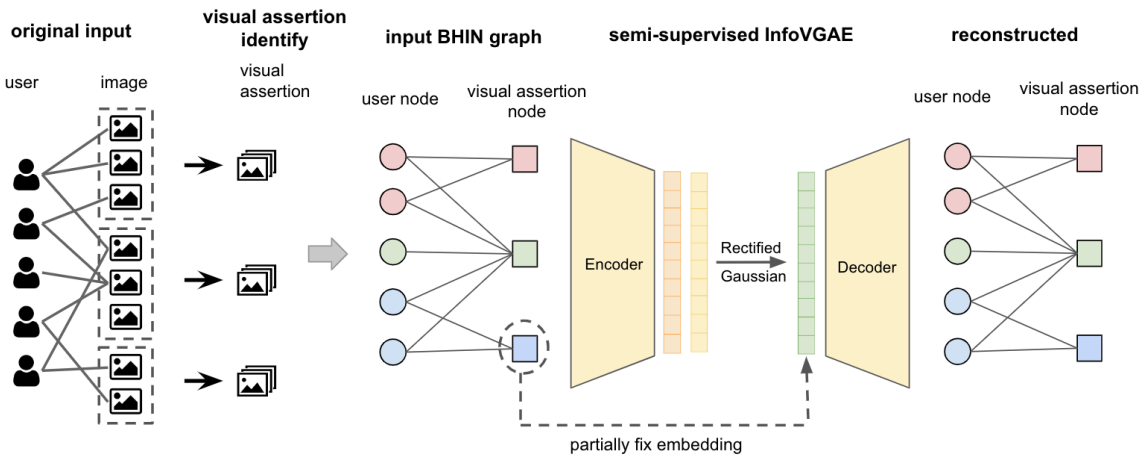


Figure 4: This is a figure showing the overall pipeline of our image ideology classification algorithm. Starting from the original user-image pair input, we firstly identify the visual assertion and convert the input data to a BHIN graph. Then we feed the BHIN graph to the semi-supervised InfoVGAE model and use the embedding result for classification.

side, we used keywords associated with Kremlin-sourced messaging, including slogans (e.g., #IstandwithPutin), allegations of Nazi influence in Ukraine, accusations that the West violated the Minsk Accords (often presented by the Russian side as a justification for the war), personal attacks on the Ukrainian president, representations of the war as self-defense against NATO, and claims of superiority of the BRICS bloc (in contrast to the G7 bloc) as a wider and more diverse representation of the World's population. All media content was auto-labeled as pro-Kremlin or pro-Ukraine, accordingly. After grouping similar images into visual assertions, a total of 21008 users and 21713 visual assertions were identified. To focus the results on "important" content only, we filtered the resulting graph to retain only those nodes that have more than 10 edges. After filtering, 713 users and 3097 assertions were included in the final evaluation.

3.2 Baselines

In comparing the InfoVGAE-based ideological separation to the state of the art, the following baselines were used for the unsupervised and the semi-supervised scenarios, respectively.

Unsupervised Models: We compare our unsupervised InfoVGAE model to two baseline models under the unsupervised setting:

- **Non-Negative Matrix Factorization (NMF)** (Al Amin et al. 2017): This unsupervised method detects the polarization in social network by factorizing the user-assertion matrix. Unlike VGAE, where the encoder is nonlinear and the decoder is linear, the encoder and decoder in NMF are both linear.
- **Belief Structured Matrix Factorization (BSMF)** (Yang et al. 2020a): This method enhances the NMF algorithm and can work well in the situations when community beliefs are partially overlapped.

Semi-supervised Models: We compare our semi-supervised InfoVGAE model with *semi-supervised TIMME* (Yang et al. 2020a), a semi-supervised multi-task and multi-relational embedding model. TIMME models the social networks as a heterogeneous graph first and then trains link prediction tasks and a latent representation task. The original TIMME model works on getting the latent representations of users, while in our task, we apply the same methodology that TIMME applies to the visual assertion network graph, and get the latent embedding for visual assertions.

3.3 Key Performance Results

Figures 5a, 5b, and 5c (embeddings on the first row of Figure 5) show the embedding results of the unsupervised approaches. The color-coding is based on ground-truth labels. Observe that the unsupervised InfoVGAE algorithm we proposed not only separates the visual assertions by ideology into largely non-overlapping clusters in the latent space but also aligns them with different axes; the x-axis represents a pro-Kremlin leaning, whereas the y-axis represents a pro-Ukraine leaning. The unsupervised algorithm does not label the axes itself, of course, but the alignment increases axis interpretability.

Figure 5d and Figure 5e shows the embedding result of semi-supervised models. In both cases, 50 pro-Ukraine and 50 pro-Kremlin assertions were manually labeled (of 3097). The figures show the embedding of *unlabeled* assertions only. As can be seen from Figure 5d, TIMME can not separate the visual assertions by ideology well when only given such a small amount of labeled data, while the semi-supervised InfoVGAE can not only separate visual assertions into different ideology groups but also ensure that the embedding of different groups aligns well with the respective axes. Note that, in this case, the user knows a priori that the x-axis is pro-Kremlin and the y-axis is pro-Ukraine. Points off the axes may be closer to neutral content, in that it is propagated to different degrees by both sides.

From the comparison of Figure 5c and Figure 5e, we can see two improvements in the embedding quality in the semi-supervised case: 1) The pro-Kremlin and pro-Ukraine visual assertions are more clearly separated. 2) The visual assertion embedding aligns better with the respective axes.

Table 1 summarizes the quantitative results, showing precision, recall, F1-score, and purity for clusters produced by the aforementioned models. The table confirms that the unsupervised InfoVGAE model we proposed achieves the best performance over other baseline models in each class (best unsupervised and best semi-supervised) on all metrics. In fact, our unsupervised algorithm also beats the semi-supervised TIMME. Besides, after labeling a small amount of data, the semi-supervised InfoVGAE model has a clear improvement over the unsupervised InfoVGAE model, increasing the F1 score by **5.59%** and improving the cluster purity by **9.60%**.

Finally, we show in Table 2 examples of the most popular visual assertions deemed by InfoVGAE as pro-Kremlin and those deemed as pro-Ukraine. The table also includes brief explanations of each image. We do not claim that the beliefs expressed in these images are necessarily espoused by all members of the group in question, but simply observe that these beliefs are expressed in messaging of the corresponding side (such as messaging from the Kremlin versus messaging from the UK Ministry of Defence).

The table illustrates the advantages of the approach used in this paper. As can be seen, the images are quite diverse; while some have overt text that reveals the stance, others are harder to interpret without proper context. By observing propagation patterns, we circumvent having to interpret the content and thus avoid reliance on complex context understanding.

4 Discussion

The paper presented a novel unsupervised approach for classifying images by ideological alignment. While the approach shows promising results, it has several limitations that need to be highlighted. Some are logistical restrictions of the current implementation. Others are research challenges that inspire future work. Below, we discuss the key limitations.

Applicability restrictions: The approach works best in situations involving social polarization and echo-chambers, where different groups disagree enough to interact with noticeably distinct subsets of content. There are other reasons (besides polarization) why different groups might interact with different content. For example, a group interested in cats and a group interested in race cars will access distinct content quite independently of their ideological leaning. For the solution described in this paper to work, the manner in which the original data set is curated thus becomes quite consequential. Ideally, the data set would be collected using keywords on a specific polarizing issue. Given a divisive umbrella topic, differences in content propagation patterns will likely be attributed to ideological disagreement. In the context of political debates, such polarization is (unfortunately) observed increasingly frequently. Hence, the limitation (arguably) does not pose a significant loss of applicability when it comes to classifying political imagery.

Visibility requirements: The approach depends on having adequate visibility into who posts what on the social medium. Social network access APIs change frequently. For example, Twitter recently restricted its access API, making it significantly more expensive to collect data at scale. Other platforms, such as Reddit and Mastodon (a new platform that has recently seen a surge of popularity in the wake of Twitter challenges) are more open. A discussion of the exact means needed for collecting the social media data is outside the scope of this work, as such a discussion is not specific to this paper. Rather, it is common to many other research directions that rely on collecting data for analysis from social media.

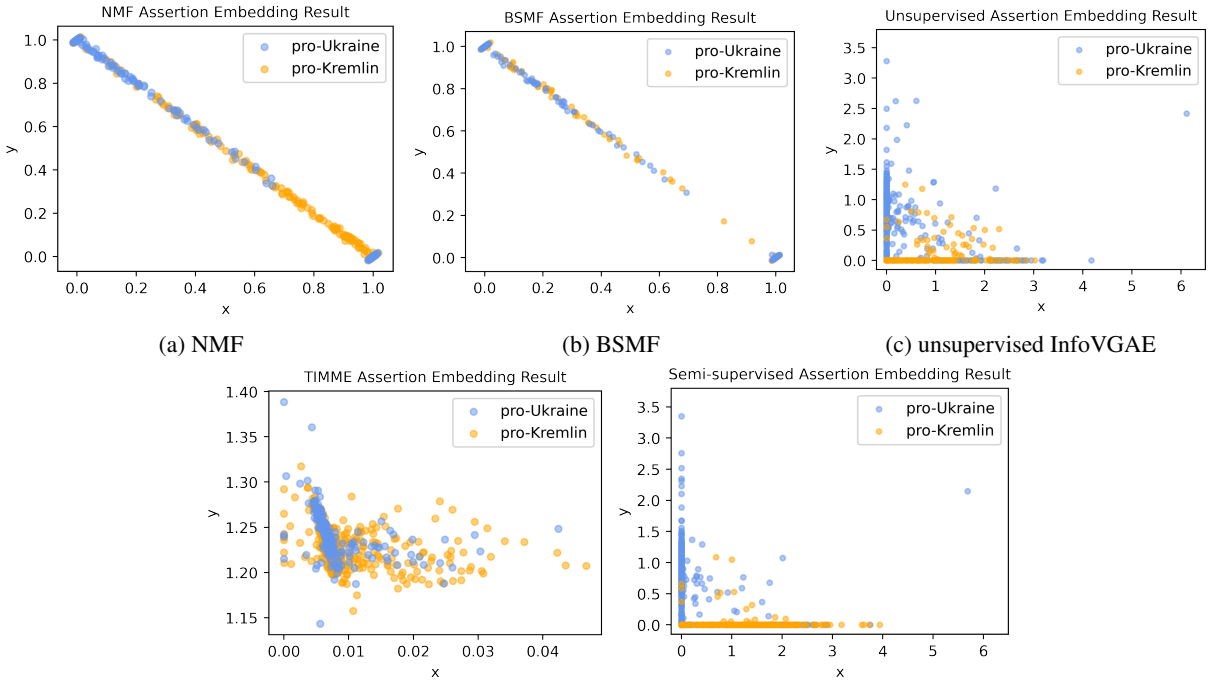
Neutral content prevalence: An important challenge to this approach is the handling of neutral content. In principle, even in a polarized debate, some imagery may be neutral and, as such, propagated by both camps. While the approach is fairly robust to having a certain amount of neutral content (InfoVGAE simply embeds it closer to the origin or near the diagonal that splits the space between two axes), some investigation is needed into the tipping point (in terms of the volume of such neutral content), after which the embedding fails to adequately separate different ideological leanings.

Semi-supervised and multimodal extensions: As alluded to in the paper, in many cases, while the ideological alignment of some images is hard to interpret automatically, other images are more direct. For example, a meme that explicitly labels Putin as a war criminal is easy to associate with anti-Russia sentiments using content analysis. It is therefore interesting to investigate optimal labeling strategies that maximize the efficacy of the semi-supervised solution. Such an approach is a topic for future work. Other extensions could leverage multimodal information, such as images and their captions, to improve classification accuracy.

The work is a step towards understanding the role and use of images in online political discourse, from misinformation and disinformation to political advertising and mobilization. A key to such understanding is scalable analysis. An initial step is to separate messaging of different sides of political discourse, at scale. Below, we review some related work and outline further opportunities for extension and synergies with alternative approaches.

5 Related Work

The key idea of exploiting content propagation patterns on social media for purposes of understanding the content itself was proposed in prior work under the name of *social sensing* (Wang et al. 2019). Early examples included fact-finding (Wang et al. 2012; 2014; 2013), event detection (Wang et al. 2017; Gao et al. 2018; Li et al. 2019; Atefeh and Khreich 2015), polarization detection (Al Amin et al. 2017; Yang et al. 2020b), and belief representation learning (Li et al. 2022). However, previous incarnations of this idea were generally applied to the understanding of



(a) NMF (b) BSMF (c) unsupervised InfoVGAE

(d) semi-supervised TIMME (e) semi-supervised InfoVGAE

Figure 5: Comparison of assertion embedding based on unsupervised model and semi-supervised model (only the unlabeled data are visualized).

Table 1: Statistical evaluation of assertion embedding performance. The evaluation compares InfoVGAE with unsupervised baseline models including NMF and BSMF, and semi-supervised baseline models including TIMME. We also compare the performance of the unsupervised and semi-supervised version of InfoVGAE.

Unsupervised Models	Precision	Recall	F1	Purity
NMF(Al Amin et al. 2017)	0.6663	0.6192	0.6311	0.6192
BSMF(Yang et al. 2020a)	0.5720	0.5583	0.5644	0.6377
unsupervised InfoVGAE (Ours)	**0.7171**	**0.8735**	**0.7876**	**0.6811**
Semi-supervised Models	Precision	Recall	F1	Purity
semi-supervised TIMME	0.6147	0.5449	0.5594	0.5449
semi-supervised InfoVGAE (Ours)	**0.8039**	**0.8872**	**0.8435**	**0.7771**

users and texts. The efficacy of the approach to classifying memes has not been systematically explored.

The work described in this paper is an instance of social polarization and stance detection tasks that focus on the automatic separation of users and their posts on social media according to their polarity (Tucker et al. 2018; ALDayel and Magdy 2021). Most existing work exploits the interaction data as well as the user profile data to separate users. For example, some work applies matrix factorization on the feature matrix to successfully separate the users and their assertions (Al Amin et al. 2017; Yang et al. 2020b). The authors in (Darwish et al. 2020) also explore applying the dimensional deduction methods and clustering algorithms to infer the polarity of users. In (Li et al. 2022; Xiao et al. 2020), the authors model the social interaction data as a graph and apply graph representation learning to understand the polarization in the beliefs of users and their posts. However, this work has not been applied to images.

In general, computer vision techniques can be applied to solve image classification tasks. Convolutional neural networks (CNNs) (Simonyan and Zisserman 2014; He et al. 2016; Krizhevsky, Sutskever, and Hinton 2017) and more modern architectures such as vision transformers have been widely successful in a broad variety of classification tasks but would require topic-specific annotations for fine-tuning for the problem discussed in this paper. While existing work applied CNNs to image sentiment detection (You et al. 2015), object identification (Zhao et al. 2019), and segmentation (Minaee et al. 2021), it does not directly allow ideological classification of images due to the need for understanding additional political context. In (Xi et al. 2020), the author combines the party affiliation classification of politicians' photos and facial sentiment detection techniques to classify the ideology of images. However, the performance of this model is limited because it is very difficult to generalize to photos of unseen politicians as well as other non-

Table 2: A case study of classified visual assertions into pro-Kremlin and pro-Ukraine groups.

Pro-Kremlin		Pro-Ukraine	
Visual Assertion	Explanation	Visual Assertion	Explanation
	An image in support of pro-Kremlin messaging that associates Ukraine with Nazi influence (depicted in the image by Nazi symbols) and satirizes Western support for Ukraine.		An image that compares aerial photographs from June 2021 and June 2022, showing the devastating effects of Russian attacks on Ukrainian targets.
	An image with a derogatory stance towards Western leaders (France, the U.S., and the U.K.) accusing them of promoting local interests at the expense of other nations.		An image of an update by a UK intelligence agency with a pro-Ukraine stance.
	A pro-Kremlin image whose stance is clearly evidenced by the anti-Ukraine writings on the poster.		An image of the damage that Russian attacks have brought about in Ukraine.
	A fake image that mocks the president of Ukraine, implying an anti-Ukraine stance. The original (from Vogue) was shown in Figure 1.		A map of Russian attacks and troop locations by Defence Intelligence, a UK Department of Defence agency supporting Ukraine.
	Another image depicting an alleged neo-Nazi influence in Ukraine, a hallmark of Kremlin messaging, in an attempt to justify military activities.		This image presumably shows Azov soldiers in Russian custody. It is possibly intended to convey mistreatment, although the stance of this image is unclear.

politician people, and therefore suffers from the over-fitting problem. Recent large vision-language models such as CLIP (Radford et al. 2021) are one avenue to inject more implicit contextual knowledge into (zero-shot) image classification, but ultimately always hinge on their ability to instill meaning into the content of images. In contrast, the proposed approach derives meaning solely from the interaction data between social entities and the images shared among them.

Graph representation techniques are proven to be effective in detecting ideology. Most existing graph-based techniques construct a graph using the historical interaction data on social networks and learn representations from it. InfoVGAE (Li et al. 2022) constructs a heterogeneous bipartite graph and learns the unsupervised belief representations of users and tweets with the property of ideological separation. In (Gu et al. 2017), the authors model the ideology of users with a message propagation formula on heterogeneous types of social links in an unsupervised manner. In (Akoglu 2014), the author constructs a signed bipartite network and formulates the polarity analysis as an unsupervised link classification task. However, most existing graph-based models only consider the users and posts as nodes. They have not been tested on multi-media memes. In this paper, we propose to classify the multi-media memes with a graph containing users and multi-media memes. We also explore enhancing the performance of classification with some minimal supervision (or multi-media meme annotations).

6 Conclusions

The work is a step towards improving the understanding of visual memes used in political discourse. The approach can be thought of as implicitly relying on a form of crowdsourcing, where individuals use their cognitive skills to interpret messages then act accordingly (e.g., like, up-vote, forward, or ignore the meme). Their collective actions result in a propagation pattern unique to the meme's content. This propagation pattern is then used to help distinguish different ideological leaning in memes. Evaluation has shown that the general idea of leveraging propagation patterns of memes in the population is a promising approach that deserves more systematic investigation. The discussion section identified limitations and further avenues for future work. The authors will pursue such opportunities in future publications.

Acknowledgments

Research reported in this paper was sponsored in part by DARPA awards HR001121C0165 and HR00112290105, the DoD Basic Research Office award HQ00342110002, and the Army Research Laboratory under Cooperative Agreement W911NF17-20196.

References

Akoglu, L. 2014. Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Al Amin, M. T.; Aggarwal, C.; Yao, S.; Abdelzaher, T.; and Kaplan, L. 2017. Unveiling polarization in social networks: A matrix factorization approach. In *INFOCOM*, 1–9. IEEE.

ALDayel, A., and Magdy, W. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58(4):102597.

Atefeh, F., and Khreich, W. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence* 31(1):132–164.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 141–152.

Fischler, M. A., and Bolles, R. C. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24(6):381–395.

Gao, T.; Bao, W.; Li, J.; Gao, X.; Kong, B.; Tang, Y.; Chen, G.; and Li, X. 2018. Dancinglines: an analytical scheme to depict cross-platform event popularity. In *Database and Expert Systems Applications: 29th International Conference, DEXA 2018, Regensburg, Germany, September 3–6, 2018, Proceedings, Part I*, 283–299. Springer.

Grant, J. M., and Flynn, P. J. 2017. Crowd scene understanding from video: a survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13(2):1–23.

Gu, Y.; Chen, T.; Sun, Y.; and Wang, B. 2017. Ideology detection for twitter users via link analysis. In *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*, 262–268. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems* 33:2611–2624.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N., and Welling, M. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM* 60(6):84–90.

Lelkes, Y. 2016. Mass polarization: Manifestations and measurements. *Public Opinion Quarterly* 80(S1):392–410.

Levinson, J. C. 2001. *Guerrilla Creativity: Make Your Message Irresistible with the Power of Memes*. Houghton Mifflin Harcourt.

Li, J.; Gao, Y.; Gao, X.; Shi, Y.; and Chen, G. 2019. Senti2pop: sentiment-aware topic popularity prediction on social media. In *2019 IEEE International conference on data mining (ICDM)*, 1174–1179. IEEE.

Li, J.; Shao, H.; Sun, D.; Wang, R.; Yan, Y.; Li, J.; Liu, S.; Tong, H.; and Abdelzaher, T. 2022. Unsupervised belief representation learning with information-theoretic variational graph auto-encoders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1728–1738.

Martínez-Rolán, X.; Piñeiro-Otero, T.; et al. 2016. The use of memes in the discourse of political parties on twitter: analysing the 2015 state of the nation debate. *Communication & Society* 29(1):145–160.

Mina, A. X. 2019. *Memes to movements: How the world's most viral media is changing social protest and power*. Beacon Press.

Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; and Terzopoulos, D. 2021. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence* 44(7):3523–3542.

Naseer, M.; Khan, S.; and Porikli, F. 2018. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE access* 7:1859–1887.

Petri, F., and Biedenkopf, K. 2021. Weathering growing polarization? the european parliament and eu foreign climate policy ambitions. *Journal of European public policy* 28(7):1057–1075.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 8748–8763.

Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *International Conference on Computer Vision (ICCV)*, 2564–2571.

Shao, H.; Yao, S.; Sun, D.; Zhang, A.; Liu, S.; Liu, D.; Wang, J.; and Abdelzaher, T. 2020. Controlvae: Controllable variational autoencoder. In *ICML*, 8655–8664. PMLR.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Socci, N. D.; Lee, D. D.; and Sebastian Seung, H. 1998. The rectified gaussian distribution. *NIPS* 350–356.

Sun, Y., and Han, J. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* 3(2):1–159.

Theisen, W.; Brogan, J.; Thomas, P. B.; Moreira, D.; Phoa, P.; Weninger, T.; and Scheirer, W. 2021. Automatic discovery of political meme genres with diverse appearances. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 714–726.

Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Wang, D.; Kaplan, L.; Le, H.; and Abdelzaher, T. 2012. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, 233–244.

Wang, D.; Abdelzaher, T.; Kaplan, L.; and Aggarwal, C. C. 2013. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *2013 IEEE 33rd international conference on distributed computing systems*, 530–539. IEEE.

Wang, D.; Amin, M. T.; Li, S.; Abdelzaher, T.; Kaplan, L.; Gu, S.; Pan, C.; Liu, H.; Aggarwal, C. C.; Ganti, R.; et al. 2014. Using humans as sensors: an estimation-theoretic perspective. In *IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks*, 35–46. IEEE.

Wang, S.; Giridhar, P.; Wang, H.; Kaplan, L.; Pham, T.; Yener, A.; and Abdelzaher, T. 2017. Storyline: Unsupervised geo-event demultiplexing in social spaces without location information. In *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation*, 83–93.

Wang, D.; Szymanski, B. K.; Abdelzaher, T.; Ji, H.; and Kaplan, L. 2019. The age of social sensing. *Computer* 52(1):36–45.

Xi, N.; Ma, D.; Liou, M.; Steinert-Threlkeld, Z. C.; Anastasopoulos, J.; and Joo, J. 2020. Understanding the political ideology of legislators from social media images. In *Proceedings of the international aaai conference on web and social media*, volume 14, 726–737.

Xiao, Z.; Song, W.; Xu, H.; Ren, Z.; and Sun, Y. 2020. Timme: Twitter ideology-detection via multi-task multi-relational embedding. In *KDD*, 2258–2268.

Xue, J.-R.; Fang, J.-W.; and Zhang, P. 2018. A survey of scene understanding by event reasoning in autonomous driving. *International Journal of Automation and Computing* 15(3):249–266.

Yang, C.; Li, J.; Wang, R.; Yao, S.; Shao, H.; Liu, D.; Liu, S.; Wang, T.; and Abdelzaher, T. F. 2020a. Disentangling overlapping beliefs by structured matrix factorization. *arXiv e-prints* arXiv–2002.

Yang, C.; Li, J.; Wang, R.; Yao, S.; Shao, H.; Liu, D.; Liu, S.; Wang, T.; and Abdelzaher, T. F. 2020b. Hierarchical overlapping belief estimation by structured matrix factorization. In *ASONAM*.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 29.

Zhao, Z.-Q.; Zheng, P.; Xu, S.-t.; and Wu, X. 2019. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems* 30(11):3212–3232.