Mapping Visual Themes among Authentic and Coordinated Memes

Keng-Chi Chang

Department of Political Science
University of California, San Diego
kechang@ucsd.edu, kengchichang.com

Abstract

What distinguishes authentic memes from those created by state actors? I utilize a self-supervised vision model, DeepCluster, to learn low-dimensional visual embeddings of memes and apply K-means to jointly cluster authentic and coordinated memes without additional inputs. I find that authentic and coordinated memes share a large fraction of visual themes but with varying degrees. Coordinated memes from Russian IRA accounts promote more themes around celebrities, quotes, screenshots, military, and gender. Authentic Reddit memes include more themes with comics and movie characters. A simple logistic regression on the low-dimensional embeddings can discern IRA memes from Reddit memes with an out-sample testing accuracy of 0.84.

Introduction

Visual memes (broadly defined as images-with-text) are everywhere on social media; a large fraction is political. According to a panel of 490K Twitter users with voter registration, 19% of their tweets are classified as memes, and 30% of the memes are politically relevant (Du, Masood, and Joseph 2020). Another study on political misinformation in Indian WhatsApp groups finds that 30% of the visual misinformation are memes (Garimella and Eckles 2020). There are legitimate concerns around state-linked online information operations affecting political behavior, but most studies to date do not leverage the wealth of data in images.

This project aims to document what kinds of visual frames are commonly promoted by state actors compared to generic, authentic memes promoted by regular non-state users. I use a large sample of data from Russian IRA accounts released by Twitter and collect a large sample of authentic memes from `r/memes` on Reddit. I feed a balanced sample of both coordinated state-linked memes and authentic memes (memes promoted by regular users) into a self-supervised vision model (Caron et al. 2019) to learn the lower-dimensional representations for each meme. I then apply the standard K-means clustering algorithm to the representations to find the clusters of memes. I find that coordinated and authentic memes differ in the visual themes and that a simple logistic regression on the lower-dimensional representations can achieve reasonable accuracy in predicting coordinated vs. authentic memes (on the test set, AUC 0.91, accuracy 0.84, F_1-score 0.84).

Compared with similar methods relying on multimodal neural networks (Beskow, Kumar, and Carley 2020; Du, Masood, and Joseph 2020), Bag-of-Visual-Words (BOVW), or Perceptual hashing (pHash) (Zannettou et al. 2018, 2019), this transfer learning framework does not rely on extensive tagging (cf. multimodal models), does not only learn on local visual features (cf. BOVW), and does not require memes to be nearly identical (cf. pHash).

Prior Works and Limitations

Twitter released a ground truth dataset of state-linked operations, including the 1.8M images posted by the accounts controlled by the Russian Internet Research Agency (IRA) during the 2016 US Presidential election. Qualitative studies pointed out that IRA employees are assessed for meme-making capabilities (DiResta, Grossman, and Siegel 2021). Other studies used *textual data* from IRA found asymmetric flooding of *entertainment*, not necesarrily *politics*, as a strategy (Cirone and Hobbs 2022), and that textual content is a reasonable predictor for state-linked campaigns (Alizadeh et al. 2020). Previous work, in a similar effort, also documented the spread of IRA memes online using Perceptual hashing (pHash) of images (Zannettou et al. 2019). However, there is a lack of systematic understanding of the amplification of visual themes by state-linked actors compared to organic ones.

Methodology

This project has the following steps, which will be explained in subsections.

1. Collect state-linked images, organic memes, and non-meme image-with-text data (as negative samples).
2. Classify state-linked images into memes vs. non-memes.
3. Extracting embeddings of visual feature *jointly* for both coordinated and authentic memes.
4. Cluster memes based on the vectors of embeddings.
5. Label the clusters and compare the difference in proportions between coordinated and authentic memes.
6. Train a simple baseline model based on visual embeddings to distinguish coordinated and authentic memes.

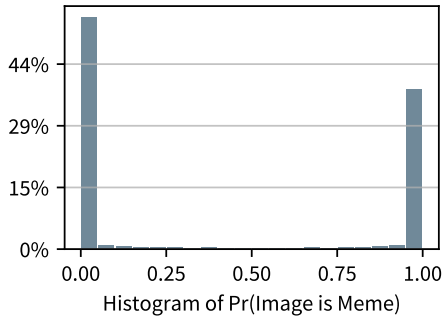

Figure 1: Distribution of predicted probability of memes for IRA images

Data collection

The primary coordinated data are the images shared by IRA on Twitter. Other than the dataset from Twitter, I collected 26K generic memes collected from the `r/meme` subreddit and 15K non-meme image-with-text data (COCO-text, (Veit et al. 2016)) as negative samples for training meme vs. non-meme classifier. The reason for choosing non-meme image-with-text data as negative samples for training is that we would not want models simply picking up textual features in images and classifying images into with vs. without text. We contend there there can still be coordinated memes in the `r/meme` subreddit, but the percentage should be low and is ouside of the scope of this project.

Classify images into memes

I trained a state-of-the-art deep learning classifier based on ResNet-50 to classify the images shared by IRA into memes vs. non-memes. The accuracy is greater than 0.97 on the test set. Most predicted probabilities are either 0 or 1. Based on this, I find that around 40% of the images shared by the IRA accounts can be classified as memes (see Figure 1 for a histogram of predicted probability).

Figure 8 in Appendix shows a sample of predicted meme vs. non-meme IRA images. There is room for improvements in accuracy. But since there is no universally accepted definition to serve as ground truth for labeling memes, this might be a simpler procedure without human labeling. For better comparisons, the later analysis will only use the IRA subsample that I classified as memes (predicted probability >0.9) to compare with authentic memes on Reddit.

Extracting representations of visual features

Most traditional image classification tasks are based on supervised learning, but it is hard to scale, especially for memes. Another approach, recently picked-up by social scientists (eg, Torres (2018)), is to extract keypoints via scale-invariant feature transform (SIFT) and find Bag-of-Visual-Words (BOVW) feature representation (Sivic and Zisserman 2003) by building patches around the neighbor of keypoints and finding clusters of patches. However, this method tends to only focus on local features around the patches and can be not meaningful enough for interpretation.

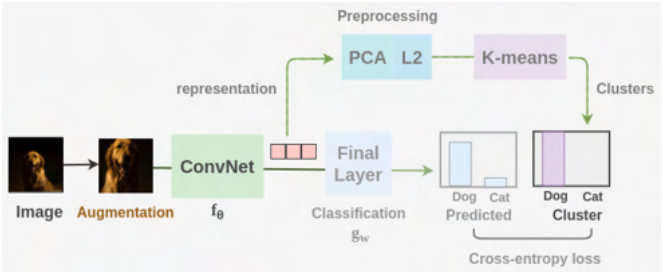

Figure 2: DeepCluster Pipeline (from Chaudhary (2020))

This paper leverages DeepCluster (Caron et al. 2019, also introduced to the social scientists by Zhang and Peng (2021) that found successes in social science applications), a recent self-supervised method for clustering images. See Figure 2 for the pipeline for DeepCluster. Specifically, DeepCluster learns *pseudo-labels* iteratively by grouping features into clusters and uses the subsequent assignments as supervision to update the weights of the convolutional neural network (ConvNet).

I feed a balanced sample of coordinated IRA memes and authentic Reddit memes (each of size 26K) into DeepCluster *jointly*. Notice that the IRA/Reddit labels are *not* inputs of the model since this is an unsupervised algorithm, and we also don't want the model to memorize the labels at this stage. For faster training, I use pre-trained weights released by the authors (based on the VGG-16 model trained on the ImageNet dataset). ImageNet contains 1.28M images

Figure 3: **Example memes from Clusters 5 (top) and 46 (bottom).** For each, the first row contains the 5 memes nearest to the center of the cluster; the 2-4 rows contain 15 random memes from that cluster. Red border indicates the meme is from IRA; blue border indicates the meme is from Reddit.

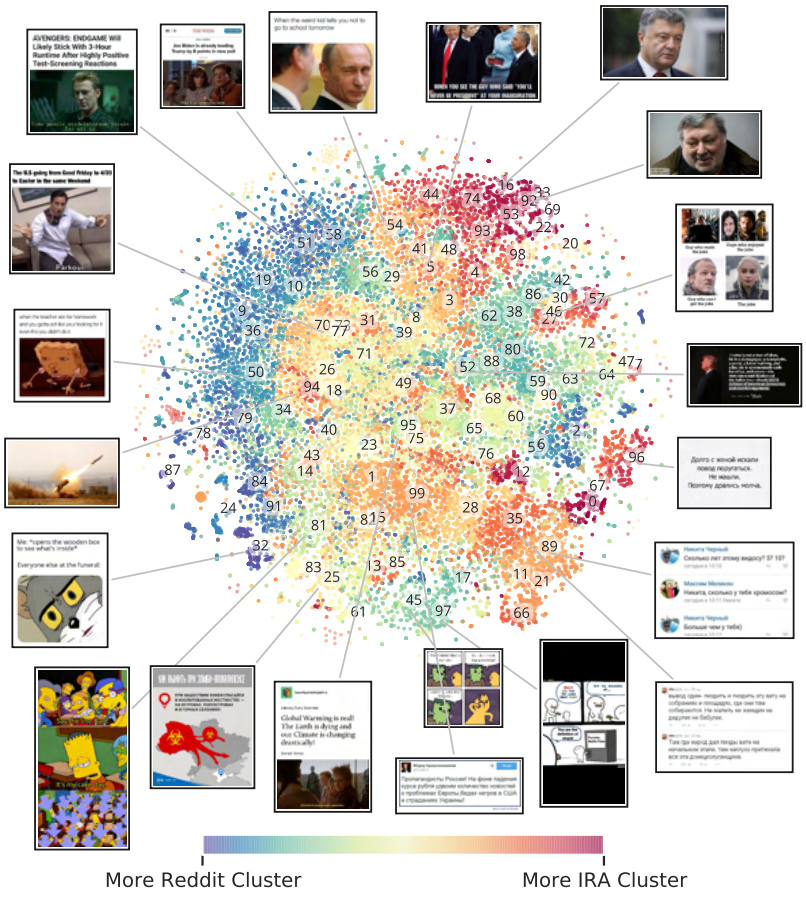

Figure 4: **t-SNE projection of IRA and Reddit memes.** Each point represents a meme from IRA or Reddit. Each color indicates a cluster (K-means with K=100 and Euclidean distance) on the embedding space learned from DeepCluster (Caron et al. 2019). Each number indicates the index of the cluster, labeled at the centroid of the cluster. Red indicates that the cluster has the highest percentage of memes from IRA; blue indicates that the cluster has the highest percentage of memes from Reddit. See Figure 5 for a complete list of the 100 clusters.

of 1000 categories such as scenes, places, and objects. Thus, the learned embeddings should be helpful in finding general themes in images, not just localized features (e.g., textures) like those in BOVW. The code and pre-trained weights are publicly available on the GitHub repository of Facebook Research. I extract the final layer of the ConvNet before classification (a 4096-dimensional vector) for each meme as a representation of the visual features.

Cluster the memes and label the clusters

After getting representations for each meme, I train the standard K-means algorithm with K=100 and Euclidean distance on the 4096-dimensional embedding space. The choice of K is still arbitrary at this stage. The idea is to choose a large enough K and combine similar clusters at the later stage.

After clustering, I see the images within each cluster to label the clusters. Specifically, I sample 5 representative memes (memes that are nearest to the center of the clusters)

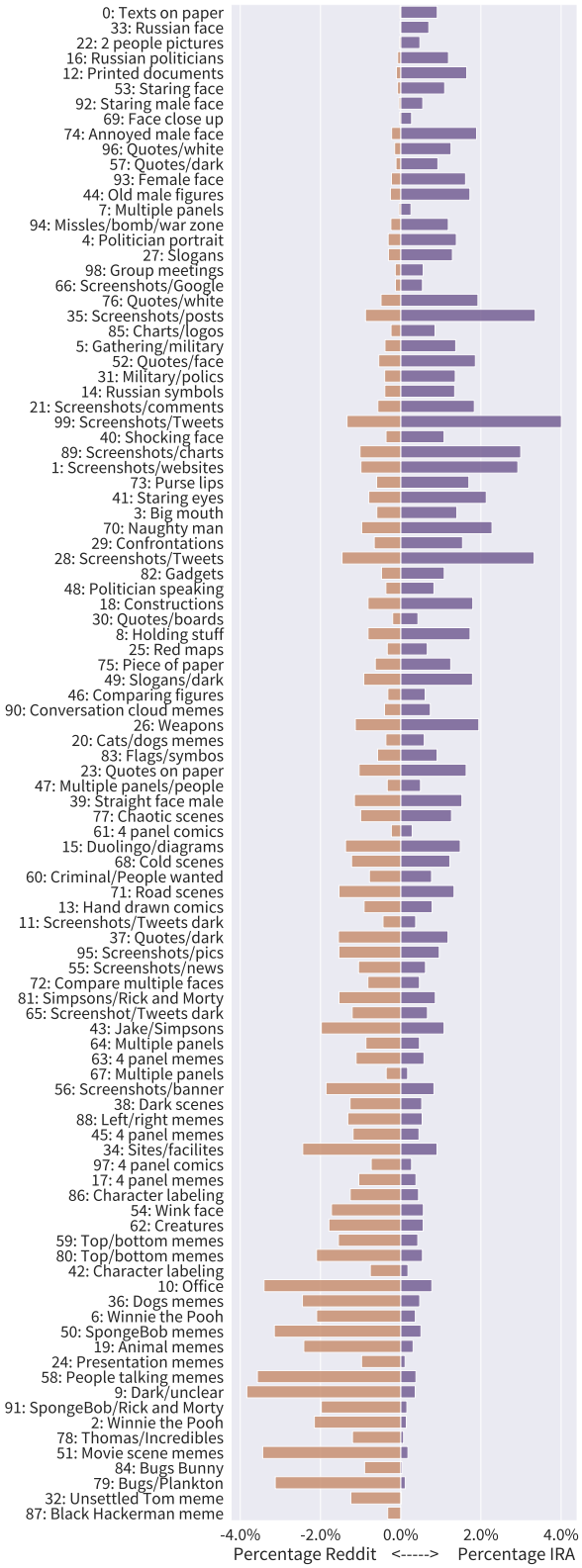

Figure 5: **Cluster labels and shares of IRA/Reddit memes** Percentages are calculated by the number of memes from IRA (Reddit) in a cluster out of total number of memes from IRA (Reddit), respectively. The clusters are ordered by the percentages.

and 15 random memes within that clusters (to ensure the robustness of the distance measure). See Figure 3 for examples from Clusters 5 and 46. Some more examples are in Figure 7 in the Appendix.

Preliminary Findings

Figure 4 plots the t-SNE projection of the learned visual embeddings for each meme. Each point represents a meme from IRA or Reddit; each color indicates a clustering result from K-means. Each number indicates the index of the cluster, labeled at the centroid of the cluster. For each cluster, we also calculate the percentage of memes in that cluster (for IRA and Reddit memes separately). Red indicates that the cluster has the highest percentage of memes from IRA; blue indicates that the cluster has the highest percentage of memes from Reddit.

One can see that clusters located near the top of Figure 4 involve mostly pictures of public figures (politicians, celebrities). Clusters located near the bottom involve mostly screenshots (Twitter/Facebook posts, quotes/slogans, news websites/headlines, messages, etc.). Clusters located near the top left consist of the common "memes": pictures surrounded by text. Clusters located near the bottom left consist of comics, maps, charts, etc. Most clusters are complex mixing of images, text, pictures, screenshots, and comics.

Figure 5 presents a complete list of clusters, and labels, along with the percentage count within IRA/Reddit memes. For example, cluster 99 (one of the clusters involving Screenshots/Tweets) accounts for 4% of the IRA memes. The top row indicates that the cluster has the highest relative percentage of IRA memes (around 1% of IRA memes and no Reddit memes); the bottom row indicates that the cluster has the lowest relative percentage of IRA memes (0.5% of Reddit memes and no IRA memes). One can see that, towards the top of the list (more common in IRA memes), there are more themes around pictures of public figures, quotes, slogans, screenshots, and scenes related to military or gender. In comparison, towards the bottom of the list (more common in Reddit memes), there are more comics, cats/dogs, superheroes, and movie scenes. Noticeably, Reddit memes usually evolve around fixated "frames" where free online meme-creating tools can help you create memes with the same frame without editing the whole meme yourself. Although these tools are wildly available in the West, it seems like the IRA accounts are not utilizing these tools to generate memes with popular frames.

Can machine learning discern IRA memes from Reddit memes simply by using the 4096-dimensional visual representations? I train a simple logistic regression using only the visual representations learned by DeepCluster with a 70/30 train/test split. This simple baseline using visual representations alone achieves training accuracy 0.90, AUC 0.91, testing accuracy 0.84, precision 0.84, recall 0.84, F_1-score 0.84. See Figure 6 for the confusion matrix for this logistic regression.

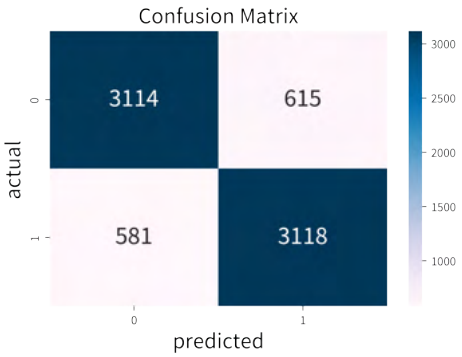

Figure 6: **Confusion matrix for logistic regression predicting IRA memes using only visual representations**

Discussions and Future Steps

In these preliminary experiments, I find that coordinated IRA memes and authentic Reddit memes share a large set of visual themes but with varying degrees. IRA memes promote more pictures of celebrities, quotes, screenshots, and images related to military and gender. Reddit memes involve more comics and movie characters. I also find that using a simple logistic regression on the learned visual representations can reasonably discern coordinated memes from authentic ones.

The proposed method, based on DeepCluster (Caron et al. 2019), does not rely on labels and can learn broader themes of images. In contrast, BOVW only learns about local visual features within patches, and pHash requires that images be nearly identical. They can be less useful in identifying the visual themes of memes.

I plan to extend this framework to find better representations of visual themes:

- With the successes of multimodal transformer models (such as VisualBERT, ViLBERT, and VL-BERT) in Facebook's Hateful Memes Challenge, we can extract texts and entity/race tags and learn a more flexible model to get richer embeddings not only based on vision but also interacts with texts and other augmented information.
- It is possible to better preprocess memes to strip off structures that are less related to themes (such as a number of panels within a meme) so that meme structures would not dominate during clustering.
- It is also possible to utilize more flexible clustering models so that each meme does not only belong to one cluster but a distribution of clusters (similar to the Latent Dirichlet Allocation, Blei, Ng, and Jordan (2003)) or even to include covariates such as source, time, or other metadata for building clusters (similar to the Structural Topic Model, Roberts et al. (2013)).

Appendix

- Figure 7: examples of representative memes from selected clusters.

- Figure 8: examples of IRA images predicted as memes and non-memes.

References

Alizadeh, M.; Shapiro, J. N.; Buntain, C.; and Tucker, J. A. 2020. Content-Based Features Predict Social Media Influence Operations. *Science Advances*, 6(30): eabb5824.

Beskow, D. M.; Kumar, S.; and Carley, K. M. 2020. The Evolution of Political Memes: Detecting and Characterizing Internet Memes with Multi-Modal Deep Learning. *Information Processing & Management*, 57(2): 102170.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3(null): 993–1022.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2019. Deep Clustering for Unsupervised Learning of Visual Features. *arXiv:1807.05520 [cs]*.

Chaudhary, A. 2020. A Visual Exploration of DeepCluster.

Cirone, A.; and Hobbs, W. 2022. Asymmetric Flooding as a Tool for Foreign Influence on Social Media. *Political Science Research and Methods*, 1–12.

DiResta, R.; Grossman, S.; and Siegel, A. 2021. In-House Vs. Outsourced Trolls: How Digital Mercenaries Shape State Influence Strategies. *Political Communication*, 0(0): 1–32.

Du, Y.; Masood, M. A.; and Joseph, K. 2020. Understanding Visual Memes: An Empirical Analysis of Text Superimposed on Memes Shared on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 153–164.

Garimella, K.; and Eckles, D. 2020. Images and Misinformation in Political Groups: Evidence from WhatsApp in India. *Harvard Kennedy School Misinformation Review*.

Roberts, M. E.; Stewart, B. M.; Tingley, D.; and Airoldi, E. M. 2013. The Structural Topic Model and Applied Social Science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Sivic; and Zisserman. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, 1470–1477 vol.2.

Torres, M. 2018. Give Me the Full Picture: Using Computer Vision to Understand Visual Frames and Political Communication. *Working Paper*, 30.

Veit, A.; Matera, T.; Neumann, L.; Matas, J.; and Belongie, S. 2016. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. *arXiv:1601.07140 [cs]*.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the Origins of Memes by Means of Fringe Web Communities. *arXiv:1805.12512 [cs]*.

Zannettou, S.; Caulfield, T.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2019. Characterizing the Use of Images in State-Sponsored Information Warfare Operations by Russian Trolls on Twitter. *arXiv:1901.05997 [cs]*.

Zhang, H.; and Peng, Y. 2021. Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research.

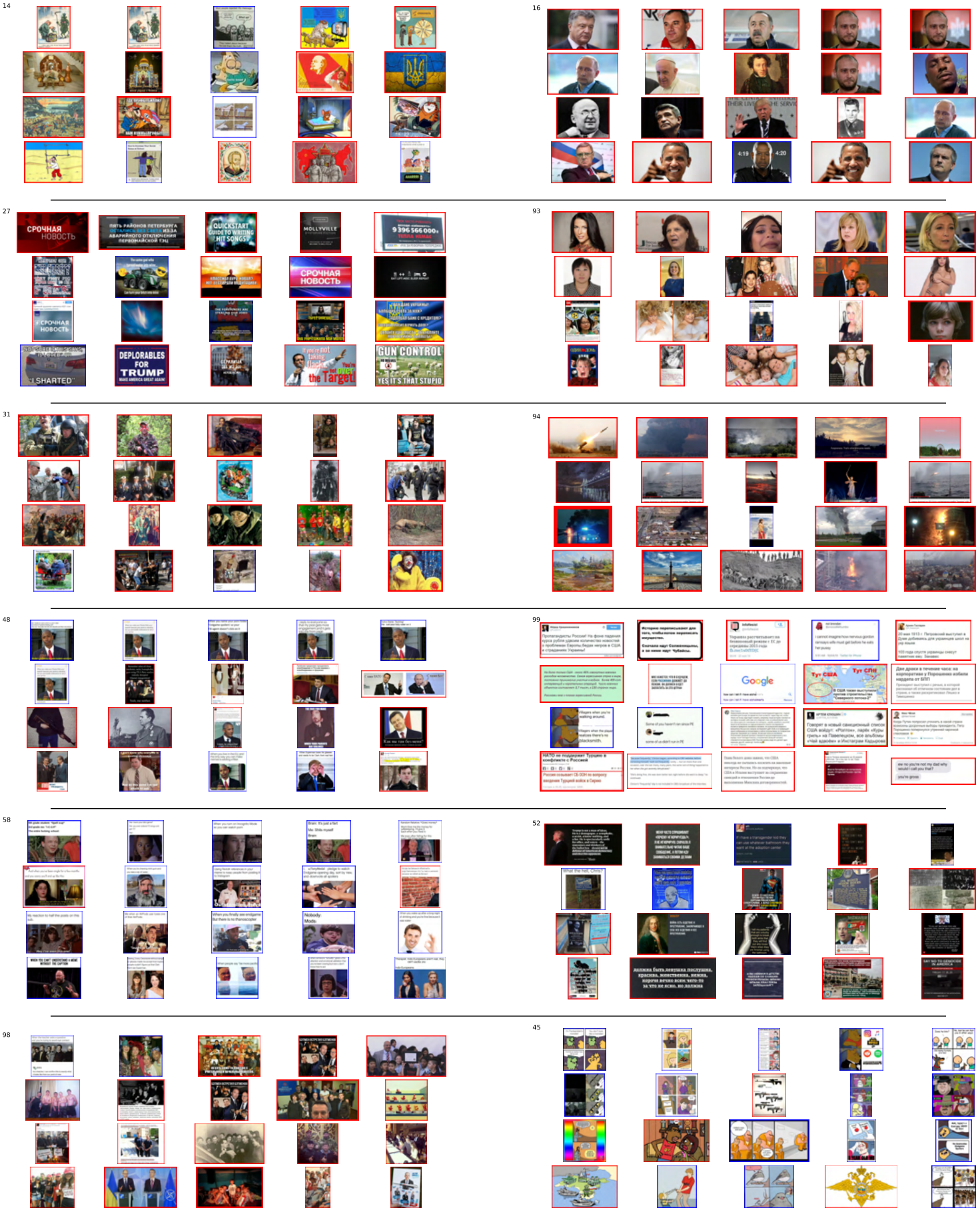

Figure 7: **Representative memes from selected clusters.** For each panel, the index of cluser is on the top left. The first row contains the 5 memes nearest to the center of the cluster; the 2-4 rows contain 15 random memes from that cluster. Red border indicates the meme is from IRA; blue border indicates the meme is from Reddit.

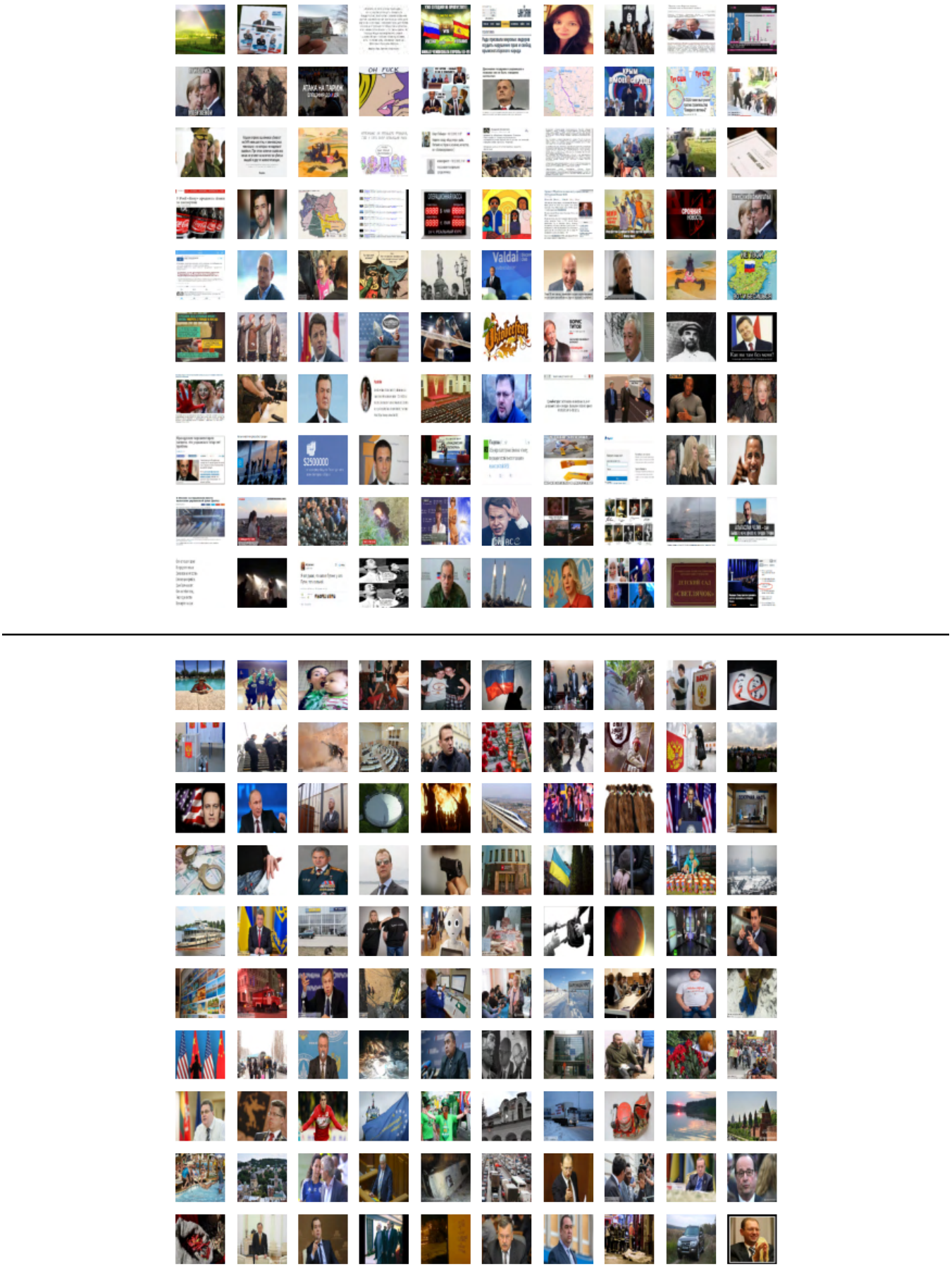

Figure 8: **Predicted memes (top) and non-memes (bottom) of IRA images.** See Methodology for details.