PhoMemes 2022 Model Description

Summary: I featurized users by embedding their images and then taking distribution statistics and cluster-based features on each of the users' embedded images. I then used a Support Vector Machine model with pseudo-labeling to create a model to predict a user's status by their images

Problem: For this problem (Challenge 2), you are tasked with determining a social media user's status (part of a disinformation campaign or benign) depending on the images they post. This problem is further complicated by the additional task of determining which disinformation campaign a user comes from, if it is from a disinformation campaign. Each user account typically has at least 20 photos in order to build each user's features.

Method: I choose first to represent each user as a stochastic image-generating function; each user posts photos according to some unknown, stochastic function. Further, a user's photo-posting function is assumed to be influence by whether it is part of a disinformation campaign as well as which disinformation campaign it is part of, if it is indeed part of one. So, the images for each user represent samples from each user's stochastic photo-posting function, which in turn can be used to determine whether that user is part of a disinformation campaign or not. I futhur added to these features by including the number of images each user has in a cluster (determined by k-means on the image embeddings) based on https://arxiv.org/abs/2110.01183.

To place the photos into a usable format for follow on machine learning tasks, I choose to embed each of the photos into a numeric, latent space. Using a pre-trained EfficieNet B7 model (https://pytorch.org/vision/stable/models.html).

Having obtained embeddings for every photo in the dataset, I then used distributional statistics for each user's embedded photos to featurize each user. First, since there are only 120 labeled accounts in the dataset, I reduced the dimensionality of all of the photo embeddings to reduce noise-to-signal ratio, using UMAP(https://umap-learn.readthedocs.io/en/latest/) to have all of the embeddings be of 15 dimensions. Then, I took the distributional statistics of mean, standard deviation, skewness, kurtosis, and differential entropy of the reduced-dimension embeddings of all of a user's images. These statistics where then concatenated together to form a feature vector for each user. I combined these with image cluster counts for each user to create the feature vectors for the users.

Then, I used some simple supervised learning models to predict a user's label, given their features. Since the training data set is small, I used simpler models with regularization to avoid overfitting. I also trained each model as a multi-class classification model with the campaign labels as the labels. Based on this, I found a Support Vector Machine with a polynomial kernel to give the best results in terms of F1 scores and balanced accuracy scoring. Finally, since there is a large amount of unlabeled, test data available I did pseudo-labeling to further improve the model.