IDENTIFYING COORDINATING ACCOUNTS ON TWITTER THROUGH TIME-BASED LATENT FACTORS

BY

MUKHILSHANKAR UMASHANKAR

Senior Thesis in Computer Science

University of Illinois Urbana-Champaign, [2023]

Adviser:

Professor Hari Sundaram

ABSTRACT

Lately, on Twitter, journalists across the globe are being attacked. These attacks have been evidenced to arrive in a flurry of hate comments the instant the journalist tweets, specifically women journalists. It appears to be coordinated as blatant patterns in the comments could be observed. These comments reduce the credibility of journalists and any personality unfairly. It is essential to identify these coordinating accounts. This thesis identifies the coordinating agents in a specific journalist's timeline, Rana Ayyub. She is one of the most brutally targeted women journalists on Twitter. For this task, the thesis runs two models: an unsupervised generative model from literature and an adapted version of the same model. The performance of the individual models is discussed, and a comparison between the two models is presented. The results of the two models are corroborated by descriptive statistics that suggest that the two models can distinguish between coordinated and uncoordinated accounts.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

The use of social media to distort public opinion through astroturfing is becoming a pressing issue (Sharma et al., 2019). People's perceptions of political candidates and social issues are being sculpted with deliberate disinformation campaigns. The large audience and dependence on social media for information have made it a convenient target. The Russian Internet Research Agency's (IRA) intervention in the US presidential election of 2016 is a well-documented disinformation campaign (Schoch et al., 2022). The campaign had disguised Russian agents as American citizens, who had introduced politically divisive narratives and spread disinformation (Al-Rawi & Rahman, 2020). These campaigns are undertaken across the globe. Studies suggest that during the South Korean 2012 presidential election campaign, the National Intelligence Service (NIS) participated in an astroturfing campaign to manipulate voters towards the eventual winner, Park Geun-hye (Keller et al., 2017).

While many of these campaigns directly influence political opinions, there have also been disinformation campaigns that divide people on social issues like social distancing policies during COVID (Sharma et al., 2021). These issues tend to become political in nature by exaggerating the divide.

Disinformation campaigns are being used to manipulate public opinion on a variety of themes, not limited to directly influencing voters. UNESCO's recent research discussion paper (UNESCO, 2021) sheds light on the disinformation efforts prevalent in invalidating critical journalism. Women journalists across the globe are targeted by disinformation purveyors to

reduce the public's trust in their journalism (UNESCO, 2021). These women have been discriminated against on a wide range of themes ranging from networked misogyny to religious bigotry and racism. They are also being attacked in person with some suffering devastating consequences (Sharma, 2022).

One such journalist whose voice has been repressed with online abuse is Rana Ayyub (Sharma, 2022), a prominent Indian journalist. She has received overwhelming hate messages due to her gender, religion, and other factors that have nothing to do with her journalistic capabilities (Sharma, 2022). The attacks against Rana Ayyub's tweets are characterized by high amounts of abusive replies within seconds of her initial post, suggesting a potential coordinated effort to discourage her from critical journalism.

Apart from affecting the journalist, these attacks also infringe upon the ideals of freedom of expression within that society. This presents a serious problem as the public could be steered away from specific journalists who attempt to present facts or an opposing viewpoint to a powerful group. It is essential to identify the accounts that are coordinating and manipulating public opinion in this manner.

To this end, this thesis explores the idea of identifying the coordinated accounts on Twitter that target women journalists, specifically Rana Ayyub.

Rana Ayyub's online abuse is characterized by high volumes of hateful content in a short period from when she tweets (Sharma, 2022). These accounts that reply in this manner are expected to

influence each other, carrying latent features in their online activities. This observation made apparent the need to adopt a model that can pick up on latent patterns in time among the coordinated accounts' activities. In 2021, a group of researchers developed a model that achieves a very similar purpose (Sharma et al., 2021). This model, Attentive Mixture Density Network with Hidden Account Group Estimation (AMDN-HAGE), dealt with identifying accounts that coordinated a COVID disinformation campaign by exploiting patterns in their online activity sequences. It is an unsupervised generative model that was adapted in this thesis to identify coordinated accounts on Rana Ayyub's Twitter timeline.

Data collection

Rana Ayyub's Twitter timeline data was collected by interfacing with twarc2. The command line tool's conversation feature obtains all data of a specified tweet. The conversation data retrieved includes the tweet features for each tweet underneath the conversation tree rooted on Rana Ayyub's original tweet. These comprise the tweet features of all reply tweets, direct replies to the original tweet, and nested replies that form a dialogue.

The conversation data obtained from Rana Ayyub's timeline was restricted to the tweets she posted between 2018 and 2022. During this period, she had started 4690 conversations, with a total of 190040 unique users participating, amounting to a total of 803101 tweets.

The raw data obtained was processed appropriately and then inputted into the original AMDN-HAGE model and an adapted version of the same model to identify the coordinated accounts. A

comparison of performance between the two versions of the model was done, providing insights into the model's abilities and the behavior of coordinated and uncoordinated groups.

CHAPTER 2: MODEL DESCRIPTION

The AMDN-HAGE model takes as input individual account activity traces ordered in time. It uses these observed account activities to identify collective behavior by jointly modelling the individual activity traces and latent account groups (Sharma et al., 2021).

The activity trace is a sequence of online activities ordered in time. It takes the form of an array $C_s = [(u_1, t_1), (u_2, t_2), (u_3, t_3) \ldots (u_n, t_n)]$ (Sharma et al., 2021). The pair (u_1, t_1) represents an activity by user u_1 at time t_1. The activity refers to a user's response to a Twitter conversation initiated by Rana Ayyub. This model was adapted to include the type of tweet, which could be either of original tweet, quoted tweet, or a reply to a tweet.

A single pair (u_1, t_1) is encoded using a 193-length embedding. It is a concatenation of User embedding, positional encoding, temporal embedding, and type of tweet encoding. The user, position, and temporal embeddings were each a 64-length embedding, while the type of tweet was a single value. As a batch of input to the model contains an activity trace of length 128, the input tensor is of shape (128,193) for the adapted model and of shape (128,192) for the original model. The model is split into two distinctive sub-models namely, AMDN and HAGE.

Attentive Mixture Density Network (AMDN)

The input tensor is passed through the first part of the model, AMDN, which models account activity traces. The tensor is passed into an attention mechanism that provides a representation of an event using attention on the events preceding (Sharma et al., 2021). The resulting event

representation is referred to as the context vector which is the same shape as the input tensor as each activity on the activity trace gets a context vector. Each of the 128 context vectors, c_i, encodes the history of events, H_{t_i}, up till time t_i (Sharma et al., 2021). The encoded vector is then passed through a decoder, a learnable conditional density function, which generates a probabilistic distribution of the succeeding event time, $\tau(u_i, t_i)$, conditioned on the events prior in the activity trace, $p(\tau \mid H_\tau)$ (Sharma et al., 2021).

A mixture of log-normal distributions is used to generate the probabilistic distribution. The conditional PDF, $p(\tau_i \mid H_{\tau_i})$, is defined as in equation (1) (Sharma et al., 2021, p. 1445),

$$p(\tau_i|w_i, \mu_i, s_i) = \sum_{k=1}^{K} w_i^k * \frac{1}{\tau s_i^k \sqrt{2\pi}} * \exp\left(-\frac{(\log \tau_i - \mu_i^k)^2}{2 * \left(s_i^k\right)^2}\right. \qquad (1)$$

where,

$$w_i = \sigma(V_w c_i + b_w), \quad s_i = \exp(V_s c_i + b_s), \quad \mu_i = V_\mu c_i + b_\mu$$

Context vector c_i and learnable parameters *V* and b are used to define the mixture weights w_i, μ_i, s_i. Maximum likelihood estimation is used to learn the account embeddings and the parameters for the encoder-decoder model described. The model is trained with gradient backpropagation, comparing with ground truth activity traces, to output the learned account embeddings. The training can be captured by equation (2) (Sharma et al., 2021, p. 1445),

$$\theta_a^*, E^* = argmax_{\theta_a, E} \log p(C_s \mid U;\ \theta_a, E) \qquad (2)$$

where θ_a is the model's parameters, E is the account embeddings, C_s is the activity trace, and U is the set of all accounts. This concludes the first part, AMDN, of the complete model.

Hidden Account Group Estimation (HAGE)

The learned account embeddings from activity traces are then passed into the second part of the model, HAGE. This part helps in identifying anomalous group behavior, thereby detecting coordination.

The HAGE model uses a Gaussian Mixture Model (GMM) to cluster the latent account embeddings (Sharma et al., 2021). The model contains 'i' social groups with each group modelled as the Gaussian distribution $\mathcal{N}(\mu_i, \Sigma_i)$ with μ_i being the cluster mean and Σ_i the covariance matrix. An account embedding, E_{u_j} for account u_j is drawn from the 'i' mixture of Gaussians and is distributed as $\Sigma_i p(i)\mathcal{N}(E_{u_j}; \mu_i, \Sigma_i)$. By training this model jointly with the AMDN model, the hidden groups can be captured just based on activity traces. The training of the HAGE model is another maximum likelihood problem, captured mathematically by equation (3) (Sharma et al., 2021, p. 1445),

$$\theta_g^* = argmax_{\theta_g} \log p\big(U;\ \theta_g, E^*\big) \qquad (3)$$

where θ_g is the model's parameters, E^* the input latent account embeddings, and U is the set of all accounts.

The joint learning problem is formulated as the bilevel optimization formulation in equations (4) and (5) (Sharma et al., 2021, p. 1445),

$$\theta_a^*, E^* = argmax_{\theta_a, E} \left[\log p(C_s \mid U;\ \theta_a, E) + \ max_{\theta_g}(\log p\big(U;\ \theta_g, E\big))\right] \qquad (4)$$

$$\theta_g^* = argmax_{\theta_g} \log p\big(U;\ \theta_g, E^*\big) \qquad (5)$$

The bilevel optimization is solved using iterative optimization according to the pseudocode in Image 1.1.

Algorithm 1 Training Algorithm for AMDN-HAGE

Require: Activity traces (C_s), Account set (U)
Ensure: Generative model (θ_a, θ_g and E)

1: $\theta_a^{(0)}, E^{(0)} \leftarrow \text{argmax}_{\theta_a,E} \log p(C_s|U;\theta_a,E)$
2: Set i as 1 {Iteration index}.
3: **while** not converged **do**
4: $\theta_g^{(i)} \leftarrow \text{argmax}_{\theta_g} \log p(U;E^{(i-1)},\theta_g)$ using EM algorithm
5: $\theta_a^{(i)}, E^{(i)} \leftarrow \text{argmax}_{\theta_a,E} \log p(C_s,U;\theta_g^{(i)},\theta_a,E)$ using SGD or its variants
6: $i \leftarrow i+1$.
7: **end while**

Image 1.1: Training Algorithm for AMDN-HAGE (Sharma et al., 2021, p. 1445)

The above algorithm was run on a single GPU with all of Rana Ayyub's timeline Twitter conversations between the years 2018- 2022.

Results

The results are a binary classification for each account, coordinated or uncoordinated. It is assumed that the binary class with a smaller number of members belongs to the coordinated group as they exhibit collectively anomalous behavior (Sharma et al., 2021). The original AMDN-HAGE model outputted a total of 2135 coordinated accounts out of a total of 190040 unique users, while the adapted AMDN-HAGE model outputted 3452 coordinated accounts for the same number of unique users. 513 accounts were classified as coordinating in both models.

For each account, the model also outputs a 64-length user embedding that encodes their coordinated behavior. Using Principal Component Analysis (PCA) the 64-length user embedding can be reduced to a 2-dimensional vector, which can be visualized for both the original and adapted model.

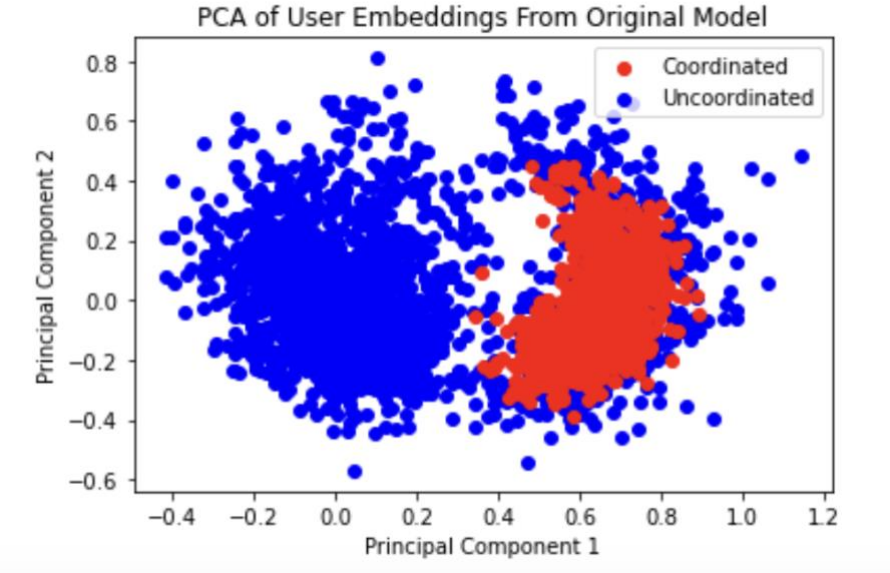

Image 1.2: PCA of User Embeddings From Original Model

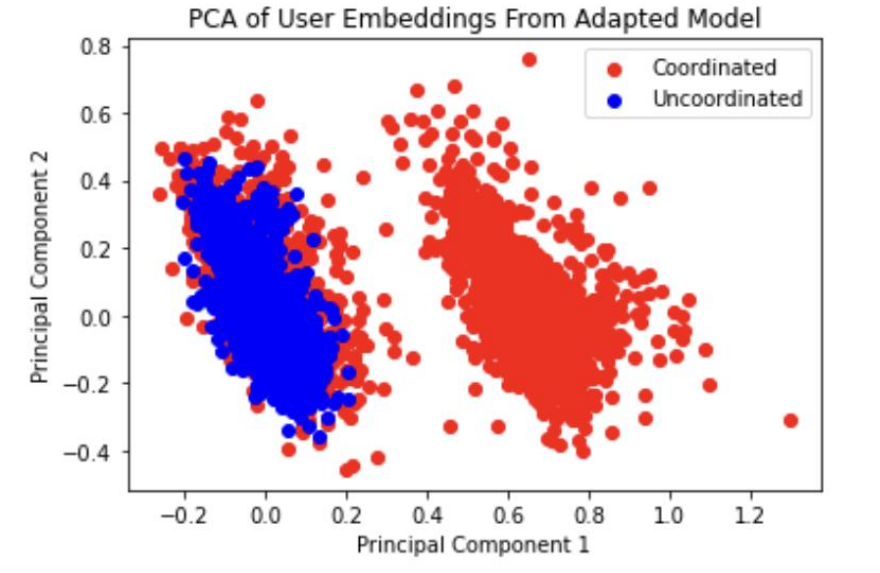

Image 1.3: PCA of User Embeddings From Adapted Model

The two Images 1.2 and 1.3 suggest that the user embedding between coordinated and uncoordinated accounts looks very different for most accounts. This can be inferred from the large blue cluster on the left in Image 1.2 and the large red cluster on the right in Image 1.3. The other cluster in both images has a higher probability of containing misclassifications as it appears that the coordinated and uncoordinated embeddings are similar after reducing them to 2-dimensions.

From the two images, it is evident that the adapted model's clusters are denser. This strongly indicates that the coordinated accounts, in the cluster on the right in Image 1.3 exhibit anomalous behavior.

PCA provides initial insights into the structure of the user embeddings. However, it is a linear algorithm that does not account for polynomial relationships between the features (Saurabh, 2022). Moreover, such linear dimensionality reduction algorithms emphasize on mapping dissimilar data points in the higher dimension to distant data points in the lower dimension. In doing so they do not place similar data points close to each other. Non-linear dimensionality reduction algorithms better preserve the structure in the lower dimension representation (Saurabh, 2022).

To this end, the user embeddings were passed through a non-linear dimensionality reduction algorithm: t-Distributed Stochastic Neighbor Embedding (t-SNE) (Tjøstheim et al., 2022). t-SNE converts the Euclidean distance between data points in the higher dimension into conditional

probabilities that capture the similarity of the points. Similarly, the conditional probability is computed for data points in the lower dimension.

The data points x_i and x_j in the higher dimension have corresponding data points y_i and y_j in the lower dimension. The conditional probabilities $p_{j|i}$ and $q_{j|i}$ capture the similarity of the data points in their respective dimensions. If $p_{j|i}$ and $q_{j|i}$ are equal for all pairings of data points, then the higher dimension data will be perfectly replicated into the lower dimension. Using this logic, the t-SNE algorithm seeks to minimize the sum of the difference between the conditional probabilities of all pairings of data points (Tjøstheim et al., 2022).

The t-SNE algorithm converts the 64-dimensional user embeddings for all the users into a 2-dimensional vector. The output is plotted in Images 1.4 and 1.5.

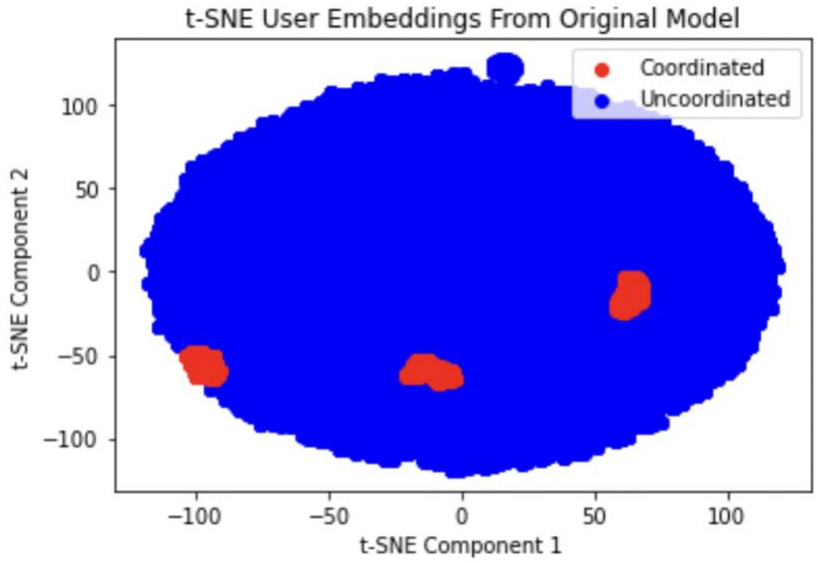

Image 1.4: t-SNE User Embeddings From Original Model

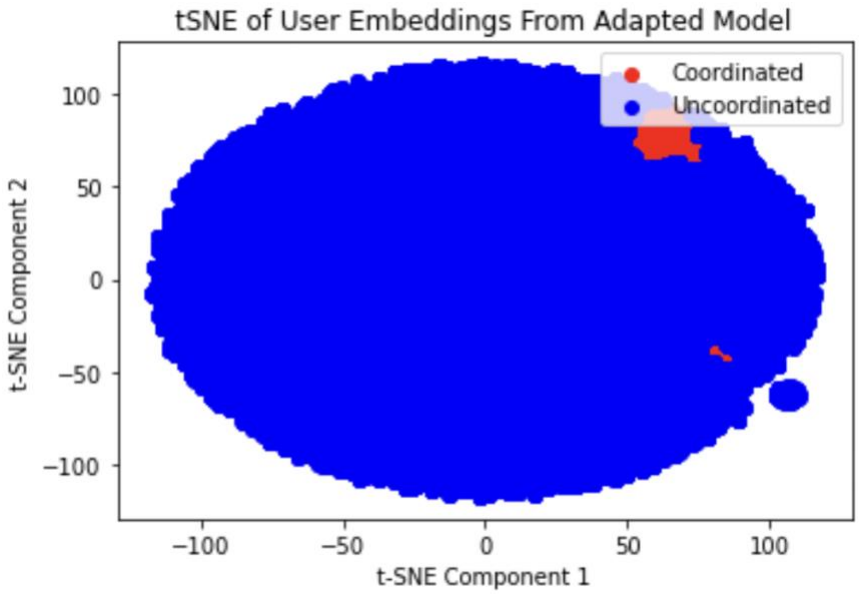

Image 1.5: t-SNE User Embeddings From Adapted Model

From the images, it is clear that the coordinated accounts belong to specific regions in the 2-dimensional plot for both models. The region is well defined with barely any overlap between the data points of coordinated and uncoordinated accounts, unlike the PCA visualizations.

It is expected that the 64-length vector user embeddings are unique as the vector majorly encodes the user's Twitter activity in time. This is captured in Images 1.4 and 1.5 as the data points are more uniquely identifiable than the data points in the PCA visualizations. This can be attributed to t-SNE's nature of preserving structure by representing points in the lower dimension based on their similarity in the original dimension.

CHAPTER 3: EVALUATION

The results of the AMDN-HAGE model are evaluated through a set of aggregated descriptive metrics. Each evaluation metric is applied to four groups of people: coordinated accounts in the original model, uncoordinated accounts in the original model, coordinated accounts in the adapted model, and uncoordinated accounts in the adapted model. The adapted model refers to the AMDN-HAGE model appended with the type of tweet feature, while the original model refers to the baseline model this thesis is centered around.

Evaluation Metric 1: Distribution of Hashtags/Suggestive Hashtags

The distribution of hashtags was a metric chosen to understand the different group's interactions with the journalist. For each group, a histogram was plotted of the top 10 most used hashtags by Twitter users in that group. The original model's uncoordinated and coordinated group of accounts had the hashtag distribution shown in Images 1.6 and 1.7.

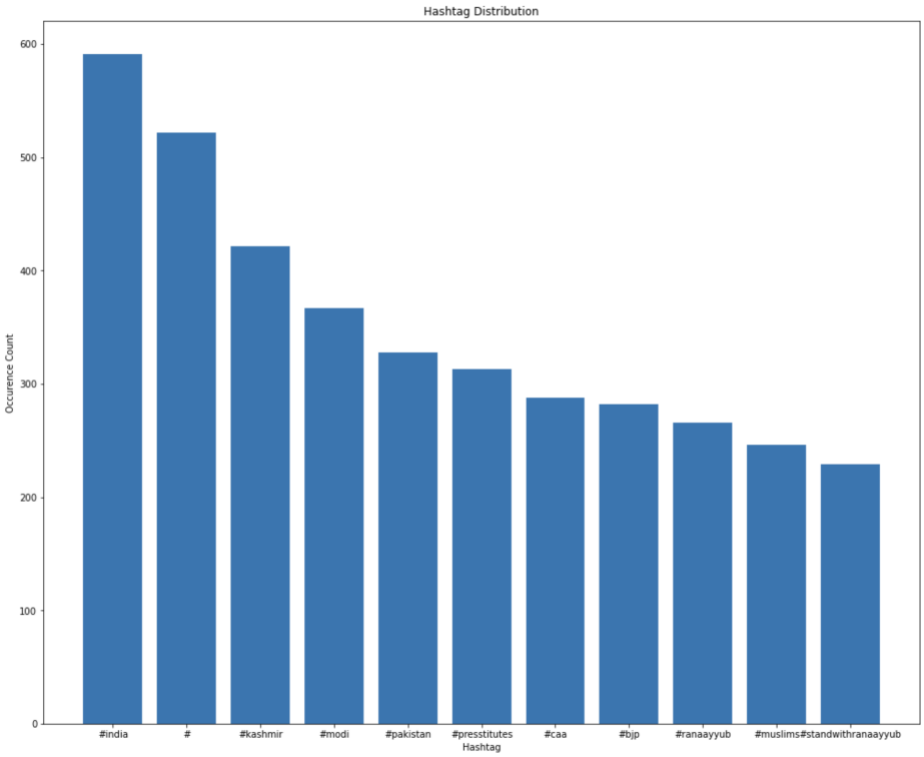

Image 1.6: Original Model Uncoordinated Group Hashtag Distribution

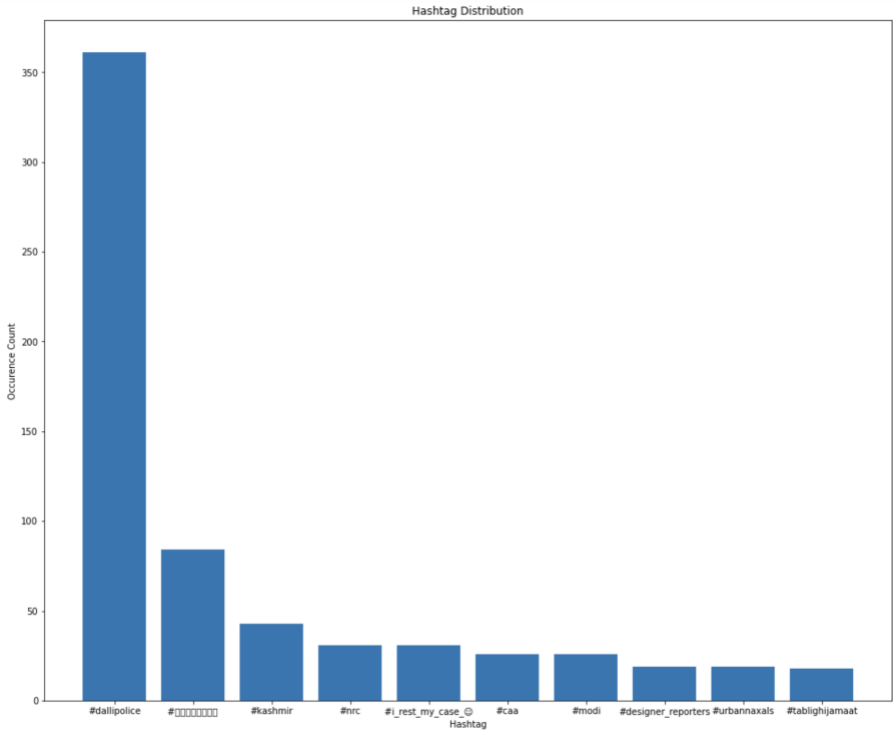

Image 1.7: Original Model Coordinated Group Hashtag Distribution

The two hashtags in Image 1.6, #presstitutes and #standwithranaayyub, are suggestive, while none appears so in Image 1.7. #presstitues is a derogatory attack on her profession of working with the press, while #standwithranaayyub shows her support. The count occurrences for #presstitutes and #standwithranaayyub in the coordinated group are 11 and 0 respectively. When averaged over the number of users in each group (Count occurrences/unique users), we get the values in Table 1.1.

Account Group	#presstitute	#standwithranaayyub
Coordinated Original Model	0.0052	0
Uncoordinated Original Model	0.0017	0.0012

Table 1.1: Suggestive Hashtag Ratios For Original Model

The high ratio of #presstitute and zero-valued ratio for #standwithranaayyub for the coordinated group suggests that the model has performed well as coordinated accounts should have a higher ratio for hate speech. The larger ratio for #presstitute over #standwithranaayyub for the uncoordinated group suggests that while there are supportive texts in this group there are more instances of hate speech, coordinated or uncoordinated.

Similar hashtag distributions were plotted for the coordinated and uncoordinated groups from the adapted model. They are depicted in Images 1.8 and 1.9.

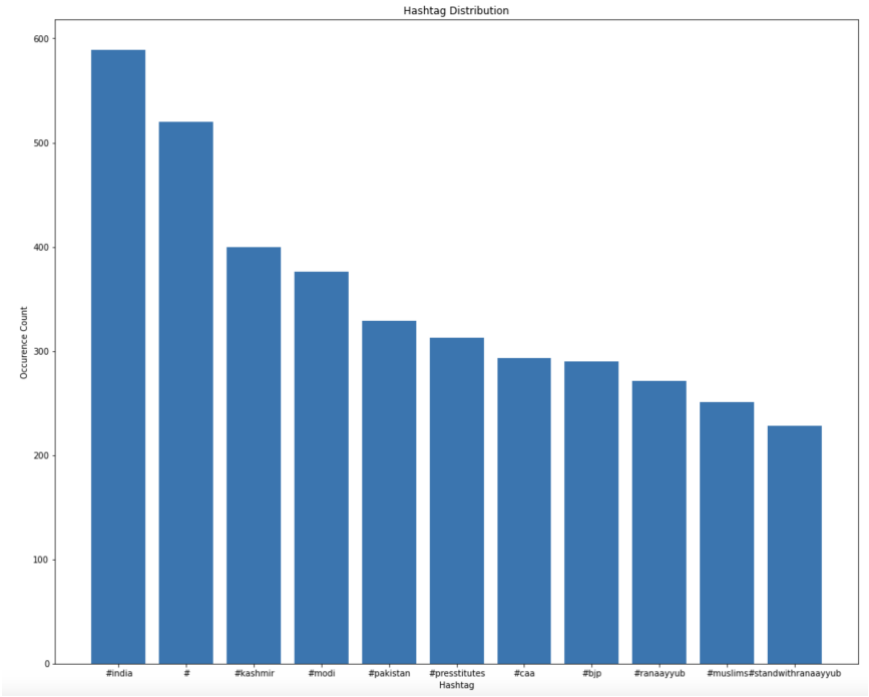

Image 1.8: Adapted Model Uncoordinated Group Hashtag Distribution

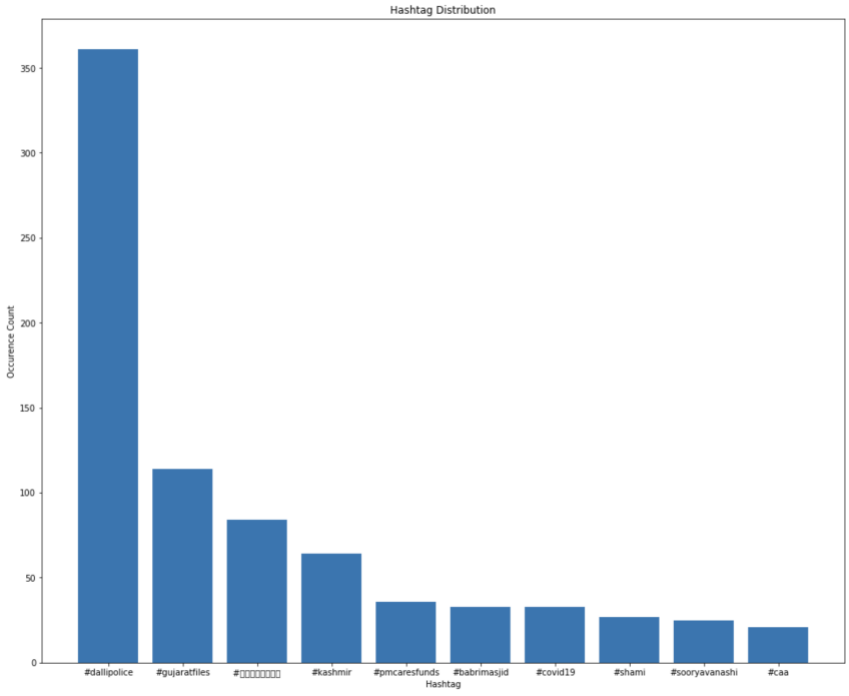

Image 1.9: Adapted Model Coordinated Group Hashtag Distribution

Similar suggestive hashtag ratio values are displayed in Table 1.2.

Account Group	#presstitute	#standwithranaayyub
Coordinated Adapted Model	0.0032	0.0002
Uncoordinated Adapted Model	0.0016	0.0012

Table 1.2: Suggestive Hashtag Ratios For Adapted Model

The original model has more promising results in this metric with a larger discrepancy between the coordinated and uncoordinated groups. The adapted model captures the trend correctly but with a lesser distinction between the two groups.

Evaluation Metric 2: Average Sentiment Value

The sentiment values are a set of three values, negative, neutral, and positive. These values were obtained for each tweet from a pre-trained model (Barbieri et al., 2020). For each author, the author's sentiment values were averaged across all his/her tweets on Rana Ayyub's timeline. Then the average sentiment values were computed for each model's coordinated and uncoordinated groups by averaging over its user's sentiment values. The results are shown in Table 1.3 below.

Account Group	Positive	Neutral	Negative
Coordinated Original Model	0.1204	0.5423	0.3372
Uncoordinated Original Model	0.1430	0.5507	0.3063
Coordinated Adapted Model	0.1237	0.5568	0.3195
Uncoordinated Adapted Model	0.1431	0.5505	0.3064

Table 1.3: Average Sentiment Values Across Groups

From the table, it is clear the negative sentiment of the coordinated groups is greater than the negative sentiment of the uncoordinated groups. Similarly, uncoordinated groups have a higher positive sentiment. Across the two metrics, suggestive hashtags, and sentiment values, it can be observed that the coordinated and uncoordinated groups are not the stark opposite in behavior. This could be reasoned by a mixture of two arguments. The uncoordinated people could be influenced by the coordinated group, thereby perceiving the journalist in a negative light. Secondly, the uncoordinated groups might not like the journalist for what she stands for, thereby acting independently. Due to these arguments, the test for the model's performance lies in the existence of a difference between coordinated and uncoordinated group metrics and not in the magnitude of that difference.

Evaluation Metric 3: Copied Tweet Ratio

This metric pays attention to the number of copied texts within the four groups across the two models. It is expected that coordinated groups will have a larger percentage of copied tweets.

To capture this idea, a ratio of the number of copied tweets over total number of tweets within each group was computed. The results are displayed in Table 1.4.

Account Group	Copying of Tweets Ratio
Coordinated Original Model	0.088
Uncoordinated Original Model	0.100
Coordinated Adapted Model	0.215
Uncoordinated Adapted Model	0.089

Table 1.4: Copying of Tweets Ratio

The results are different from the initial expectation as for the original model a larger percentage of people in the uncoordinated group copied text as opposed to the percentage who copied text in the coordinated group. However, for the adapted model the coordinated group had a lot more copied texts in comparison to its counterpart. This is explained by the addition of the tweet-type feature. It can be understood that there is a greater correlation between tweet type and copied text as opposed to tweet timings and copied text.

Evaluation Metric 4: Follower-Following Ratio and Tweet Counts

Another metric that can be used is the following-follower ratio and the tweet counts. These values are averaged across each group. The average tweet counts of the coordinated group are expected to be higher, while the average following-follower ratio could be indifferent to coordination. The data for these metrics are presented in Table 1.5.

Account Group	Average Following-Follower Ratio	Average Tweet Count
Coordinated Original	0.5079	22010
Uncoordinated Original	0.4877	12714
Coordinated Adapted	0.4503	17298
Uncoordinated Adapted	0.4889	12736

Table 1.5: Following-Follower Ratio and Tweet Counts

As expected, there is no clear pattern within the following-follower ratio across the account groups, while the tweet counts are higher amongst the coordinated group. That can be justified as coordinated groups have an agenda incentivizing them to actively spread disinformation.

Evaluation Metric 5: Average Number of Tweets Across All Conversations Participated In

This value is computed for each account. The value for the account group is obtained by averaging this value over all its members. It can be argued that coordinated groups should have a higher value for this metric as they try to bombard the journalist's timeline with tweets that suit their agenda. However, there can also exist the counterargument that coordinated accounts attempt to comment on every post of Rana Ayyub's, and that uncoordinated people tend to indulge in specific discussions, hence increasing their group's value for this metric. The results are shown in Table 1.6.

Account Group	Average Number of Tweets Across All Conversations Participated In
Coordinated Original	1.740
Uncoordinated Original	1.324
Coordinated Adapted	2.526
Uncoordinated Adapted	1.307

Table 1.6: Average Number of Tweets Across All Conversations Participated In

The results suggest that in both models the group identified as coordinated has much higher values for this metric. The large discrepancy between the coordinated and uncoordinated groups across both models suggests that there is a correlation between the average number of tweets across all conversations and coordination, possibly invalidating the extent of the counterarguments made earlier.

Evaluation Metric 6: Normalized Average Degree of Collaboration

This value attempts to capture how often a pair of users are found replying together across several Twitter conversations. The metric is calculated for each user and is updated on a conversation basis. For each conversation, a complete subgraph is constructed with the nodes as participating users in the chosen conversation and the edge weights as a proxy collaboration value. The edge weight between users U_1 and U_2 is computed as the minimum number of tweets from the two users under that Twitter conversation tree. For example, if user, U_1, had tweeted 10 times under this conversation, while user, U_2, tweeted 6 times, the edge weight between the users would be 6. In this manner, a subgraph is constructed for each of Rana Ayyub's Twitter conversations.

These graphs are then aggregated to obtain a single supergraph whose nodes are all users in Rana Ayyub's Twitter timeline, while the edge weights capture degree of collaboration between two users across all the journalists' conversations. The edge weight between users U_1 and U_2 in the supergraph is computed by adding the edge weights of the two users across all the subgraphs constructed for each conversation.

From this supergraph, the degree of collaboration, a node-level metric, for each user is computed by taking the average of the edge weights for the node representing that user. In a similar manner, the degree of collaboration is calculated for all users. These values are used to find the normalized average degree of collaboration for the coordinated and uncoordinated groups in the original and the adapted model, which are presented in Table 1.7.

Account Group	Normalized Average Degree of Collaboration
Coordinated Original	0.658
Uncoordinated Original	0.342
Coordinated Adapted	0.714
Uncoordinated Adapted	0.286

Table 1.7: Normalized Average Degree of Collaboration

There is a clear trend across both models that the coordinated group has a much higher normalized average degree of collaboration when compared to the corresponding value for the uncoordinated group.

This trend is expected as coordinated attackers work together to bring down the opinion of the journalist. Several of these attackers are expected to disguise themselves as normal citizens and collaboratively speak against the journalist's tweets. This results in the journalist losing credibility with the uncoordinated users due to the large backlash caused by the coordinated users. By this argument, the adapted model is better off at identifying the coordinated group as the discrepancy between the metric value for coordinated and uncoordinated groups is larger for the adapted model in comparison to the original model.

Although the trend is as expected and the discrepancy is clear, the current computation for this metric linearly adds the edge weights between two users U_1 and U_2 across the subgraphs. It can be argued that if two users are seen tweeting together across several conversations there is a higher probability that they are coordinated than uncoordinated. This notion can be incorporated

into the metric by log scaling the edge weights of the supergraph based on the number of conversations the two users are seen together in. The log scaling of the edge weight is described in equation (6)

$$E_i^* = E_i * \log(1 + n) \qquad (6)$$

where E_i^* is the updated edge weight, E_i is the original supergraph edge weight, and n is the number of conversations in which users U_1 and U_2, which form the edge E_i, are seen tweeting together in.

This change to the supergraph's edge weights will also help correct the scenario wherein several passionate uncoordinated users tweet extensively in very few niche conversations, resulting in a large average degree of collaboration values from those few subgraphs constructed. The larger value can be attributed to the lower number of neighbors for such users in the supergraph and the large edge weights due to their extensive tweeting in the few conversations.

The assumptions made earlier that rationalized log scaling the edge weights in the supergraph are partially justified through the normalized Log-scaled average degree of collaboration values presented in Table 1.8 as the discrepancy is larger between the coordinated and uncoordinated groups.

Account Group	Normalized Log-Scaled Average Degree of Collaboration
Coordinated Original	0.762
Uncoordinated Original	0.238
Coordinated Adapted	0.789
Uncoordinated Adapted	0.211

Table 1.8: Normalized Log-Scaled Average Degree of Collaboration

CHAPTER 4: CONCLUSION

The evaluation of the two models for the several metrics highlights a few interesting ideas. Firstly, some metrics could help validate the model's performance, like the suggestive hashtags, sentiment values, copied tweet ratio, tweet counts, and average degree of collaboration. For each of these metrics, we expect a distinction between the coordinated and uncoordinated groups for the reasons listed in Table 1.9.

Metric	Reason for distinction
Suggestive Hashtags	It is expected that on average coordinated groups use hateful hashtags more, while the uncoordinated group tweet supportive hashtags.
Sentiment Values	The coordinated group is more likely to use hateful speech, resulting in a larger average negative sentiment and lower positive sentiment. The opposite is true for the uncoordinated group.
Copied Tweet Ratio	Coordinated groups can be expected to copy each other's tweets to push their agenda, while uncoordinated people just share their individual thoughts.
Average Tweet Counts	With an agenda to pursue, coordinated groups are expected to tweet very often as opposed to uncoordinated groups.
Average Degree of Collaboration	Coordinated users are expected to tweet in the same conversations to amplify their viewpoint. Uncoordinated users are not expected to be seen tweeting together in several conversations. Their behaviour

	is expected to be more random, so it is unlikely that they are seen replying together with other such users.

Table 1.9

The expected distinction between coordinated and uncoordinated groups across both models existed in these metrics. This is a strong indication that the original and the adapted model performed well in distinguishing between coordinated and uncoordinated groups. However, this indication is slightly weakened by the Copied Tweet Ratio discrepancy for the original model as the uncoordinated group had a higher ratio in comparison to the coordinated group. On the other hand, the discrepancy between the Copied Tweet Ratio for the two groups of the adapted model was the clearest across all metrics.

The change in this metric over the original and adapted model gives insights into the impact of adding the tweet type feature in identifying coordination. The performance improvement suggests a stronger correlation between copied text and tweet type over copied text and tweet timings. This difference across the models could suggest that the original model does not perform well in identifying coordinated accounts as Copied Text is a very strong metric to distinguish between coordinated and uncoordinated. However, based on the other metrics in Table 1.9, the original model's coordinated group marginally performs better than the adapted model's coordinated group. Hence, it can be concluded that both models are mostly successful in distinguishing between coordinated and uncoordinated accounts when aggregated, however, both models likely misclassify some coordinated accounts as uncoordinated. In short, an inferred conclusion would be that the true positive rate and the false negative rate of both models could be high. There are

hints that there are a lot more coordinating accounts that are not picked up by both models. This is intuitive as coordination can occur across other domains that are not only time, like textual features, dialogue structures etc.

The two other metrics considered, the following-follower ratio and the average number of tweets across all conversations, do not have an expected distinction between the coordinated and uncoordinated groups as objective arguments can be made for both sides as to why either of the groups should have the higher metric value. Hence, these details do not validate either model. However, the metrics help us better understand the dataset and the behavior of coordinated and uncoordinated groups.

The Log-scaled average degree of collaboration metric also helps us understand the aggregate behavior of coordinated and uncoordinated groups. The scaling caused a larger discrepancy between the metric value for the coordinated and uncoordinated groups in both models. This strongly suggests that coordinated attackers tweet together in several conversations to amplify the opposition to the journalist. They seem to have chosen the strategy of strength in numbers, compromising on potential suspicion.

This thesis can be concluded with the idea that the metrics in Table 1.9 help us understand the capabilities of the models. When aggregated, the metrics suggest that the coordinated groups work collaboratively. Following that, we can now understand the behavior of coordinated and uncoordinated people in terms of other metrics like, the following-follower ratio and the average number of tweets across all conversations participated in. This understanding helps evaluate

either the argument or counterargument made earlier in this context that initially rendered these metrics indifferent to coordination.

CHAPTER 5: FUTURE WORK

This thesis highlighted the existence of time-coordinated accounts in Rana Ayyub's Twitter timeline. The conclusion hints that there could be a lot of coordinated accounts that were not picked up by the original or the adapted model. This is reasoned with the argument that coordination can occur across other vectors like dialogue structures, and textual features and not solely over time. Future work revolves around developing models that encode these new features to improve the accuracy of classification.

The work can be extended further by trying to identify the strategies the coordinating group adopts at various points in time to maximize the harm to the journalist. A positive correlation between the observable harm caused to the journalist and the coordinating group's current attack strategy is very indicative of the group's ill intent.

A future model could be developed that identifies the coordinating group as the set of accounts that adapts its attacking strategy to maximize harm caused to the journalist. The premise of such a model lies in the idea that an effective coordinated group with malign intent will seek to change the behavior of the journalist they attack in a particular manner. This could range from changing the tone of the journalist to making him/her deactivate the account. These effects on the journalist are observable, hence the coordinated group is expected to change their attacking strategy if the current strategy does not lead to desirable effects. The future model can encode and update an interpretable strategy vector for each Twitter user and observe which users' vectors change in a synchronized manner, mimicking coordinated attack strategies. Such a model

will be able to better justify an account classification as the attack strategy adopted by the user can be interpreted over time, correlating to the harm caused to the journalist. Such an interpretable model could provide compelling evidence that can be grounds for Twitter to remove the accounts.

Another extension would be to run this model for other journalists and evaluate the results using the metrics used in this thesis. The results can give insights into how different coordinated attacks against journalists are in various countries. A comparison between coordinated attacks on male and female journalists can also be worth exploring.

REFERENCES

1. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news. *ACM Transactions on Intelligent Systems and Technology*, *10*(3), 1–42. https://doi.org/10.1145/3305260
2. Schoch, D., Keller, F. B., Stier, S., & Yang, J. (2022). Coordination patterns reveal online political astroturfing across the world. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-08404-9
3. Al-Rawi, A., & Rahman, A. (2020). Manufacturing rage: The Russian Internet Research Agency's political astroturfing on social media. *First Monday*. https://doi.org/10.5210/fm.v25i9.10801
4. Keller, F., Schoch, D., Stier, S., & Yang, J. (2017). How to manipulate social media: Analyzing political astroturfing using ground truth data from South Korea. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 564–567. https://doi.org/10.1609/icwsm.v11i1.14941
5. Sharma, K., Zhang, Y., Ferrara, E., & Liu, Y. (2021). Identifying coordinated accounts on social media through hidden influence and group behaviours. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. https://doi.org/10.1145/3447548.3467391
6. Posetti, J., Aboulez, N., Bontcheva, K., Maynard , D., & Shabbir, N. (2022, May 4). *The chilling: Global trends in online violence against women journalists*. UNESCO. https://en.unesco.org/publications/thechilling
7. Sharma, B. (2022, July 15). *What 8.5 million tweets targeting Rana Ayyub tell us about online violence & the failure to stop itBetwa*. Article 14. https://www.article-14.com/post/what-8-5-million-tweets-targeting-rana-ayyub-tell-us-about-online-violence-the-failure-to-stop-it-62d104dd20f4b
8. Saurabh.jaju2. (2022, June 23). *Comprehensive guide on T-SNE algorithm with implementation in R & python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/
9. Tjøstheim, D., Jullum, M., & Løland, A. (2023). Statistical embedding: Beyond principal components. *Statistical Science*, *1*(1). https://doi.org/10.1214/22-sts881
10. Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for Tweet Classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.148