Anoushka Shinde **Social Media Mining** **Fall 2024**

Title: Social Network Cluster and Content Analysis of ISIS in Twitter (1262 words)

Introduction:

The rise of the Islamic State of Iraq and Syria (ISIS) has been accompanied by a sophisticated propaganda campaign leveraging social media platforms, particularly Twitter. This study aimed to gain insights into the pro-ISIS Twitter network by identifying influential accounts and analyzing the content of their tweets using a dataset comprising over 17,000 tweets from more than 100 pro-ISIS Twitter accounts collected from Kaggle, spanning the period following the November 2015 Paris Attacks up until 2019. The data collection method aligned with approaches used by Ali (2016) in identifying terrorist affiliations via social network analysis and Peele (2015) in exploring how ISIS propagates its message online through structured metadata. The results were visualized with Gephi, scaling influential users larger than others to provide a clear depiction of the network structure, inspired by the methods outlined by Benigni et al. (2017), who effectively used similar techniques to detect and map extremist communities.

Research Question: Who are the top influencers in the pro-ISIS twitter network? What are the pro-ISIS users talking about?

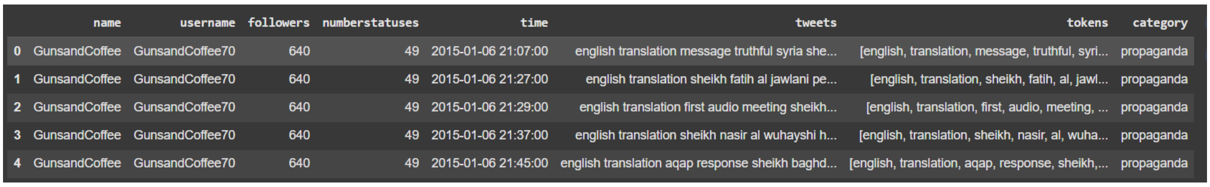
Method

Data : This study utilized a dataset found on Kaggle, comprising over 17,000 tweets collected from more than 100 pro-ISIS Twitter accounts globally, following the November 2015 Paris Attacks and has been updated up until 2019. The dataset consists of columns like: name, usernames, descriptions, locations, follower counts (when the dataset was created), status counts (when the dataset was created), timestamps, and tweet text. Data storage is done in a csv file that is 4.5MB huge with 17410 rows and 8 columns. By focusing on follower-following relationships, I can construct a social network graph and analyze the community structures for identifying influential accounts.

Analysis:

To understand what ISIS fanboys talk on Twitter, I am going to do sentiment analysis using python with supervised learning. Starting with preprocessing, the location and description column were dropped because they had missing values and random content that wasn;t useful. Then, the time column was standardized to a proper datetime format. Simultaneously, I removed all duplicates, urls, mentions ,hashtags, special characters, numbers, stop words and normalize the case. Lemmatization is carried out to reduce the words to their base form, followed by tokenization of the tweets.

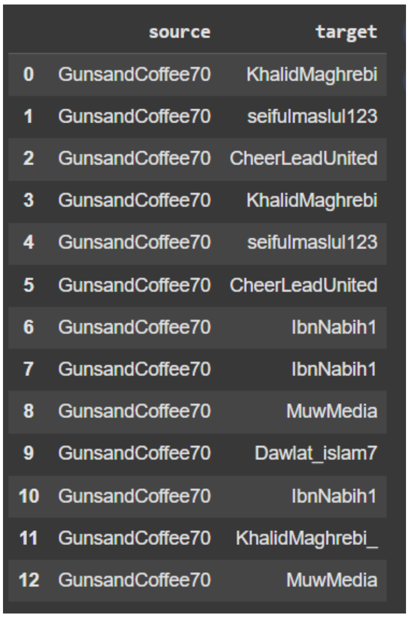
Empath library helped identify dominant themes in the text by analysing linguistic patterns to define 3 primary custom categories: propaganda(promoting violence or extremist ideologies), weapon, terrorism, crime, and religion. Using Empath scores, labels will be assigned. I will use Logistic Regression with TF-IDF embeddings to predict thematic categories. The model will be evaluated using accuracy, precision, recall and F1-score.



	name	username	followers	numberstatuses	time	tweets	tokens	category
0	GunsandCoffee	GunsandCoffee70	640	49	2015-01-06 21:07:00	english translation message truthful syria she...	[english, translation, message, truthful, syri...	propaganda
1	GunsandCoffee	GunsandCoffee70	640	49	2015-01-06 21:27:00	english translation sheikh fatih al jawlani pe...	[english, translation, sheikh, fatih, al, jawl...	propaganda
2	GunsandCoffee	GunsandCoffee70	640	49	2015-01-06 21:29:00	english translation first audio meeting sheikh...	[english, translation, first, audio, meeting, ...	propaganda
3	GunsandCoffee	GunsandCoffee70	640	49	2015-01-06 21:37:00	english translation sheikh nasir al wuhayshi h...	[english, translation, sheikh, nasir, al, wuha...	propaganda
4	GunsandCoffee	GunsandCoffee70	640	49	2015-01-06 21:45:00	english translation aqap response sheikh baghd...	[english, translation, aqap, response, sheikh,...	propaganda

Figure 1: Data after processing it.

Before loading the data in Gephi, I had to create in-degree(mentions received) and out degree(mentions made) to create a network data. I did this by extracting the mentions from tweets and creating an edge table where the source is the username of the user making the mention and the target is the username being mentioned. The in-degree and out-degree where calculated on the basis of the count of incoming edges and outgoing edges for each user respectively.



	source	target
0	GunsandCoffee70	KhalidMaghrebi
1	GunsandCoffee70	seifulmaslul123
2	GunsandCoffee70	CheerLeadUnited
3	GunsandCoffee70	KhalidMaghrebi
4	GunsandCoffee70	seifulmaslul123
5	GunsandCoffee70	CheerLeadUnited
6	GunsandCoffee70	IbnNabih1
7	GunsandCoffee70	IbnNabih1
8	GunsandCoffee70	MuwMedia
9	GunsandCoffee70	Dawlat_islam7
10	GunsandCoffee70	IbnNabih1
11	GunsandCoffee70	KhalidMaghrebi_
12	GunsandCoffee70	MuwMedia

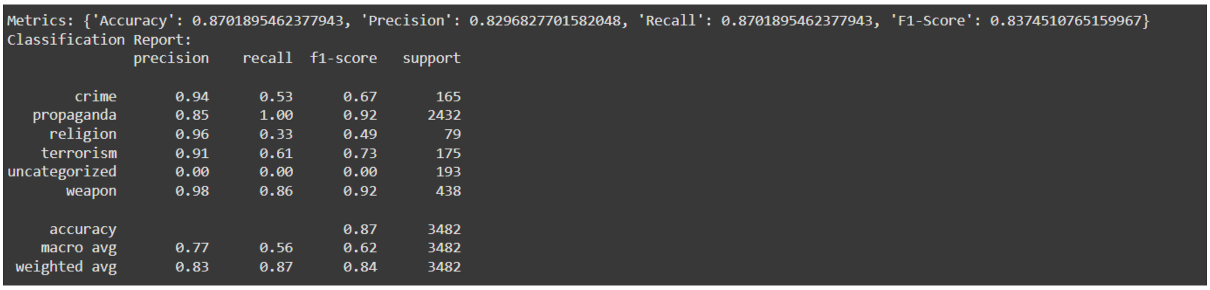
Figure 2: Data after modification for Gephi

Then, I loaded the directed graph data in Gephi where it created the nodes me. The nodes were the unique usernames. This is going to help me create a directed social network graph where nodes represent Twitter users and edges represent follower-following relationships.

I used **Social Network Analysis (SNA)** to identify the most influential accounts in the pro-ISIS Twitter network Centrality measures, such as closeness centrality, . To identify tightly knit subgroups, I used the **Louvain algorithm** for community detection. For most tagged username, closeness centrality was applied to rank the nodes based on colour. They were also ranked with in-degree on size. So as the number of mentions to a user increases, their size increases. The same ranking was added for the labels to avoid overcrowding. For the same reason, I also added a filter under 'Topology', for 'in-degree'. This filter helped to focus on only those people who had a count of 3 or more mentions.

Results:

It was discovered that there were 12208 tweets categorised as propaganda, 2178 for weapon,856 for terrorism, 792 for crime, and 419 for religion. The results show the performance metrics of a sentiment analysis model that categorizes tweets into themes like terrorism, crime, religion, etc. The model achieved an accuracy of 77-87% and a weighted average F1-score of 84%. Propaganda had the highest F1-score at 92%, while religion and uncategorized had lower scores around 49-73%. The macro average precision, recall and F1-score were 83-87%.



```
Metrics: {'Accuracy': 0.8701895462377943, 'Precision': 0.8296827701582048, 'Recall': 0.8701895462377943, 'F1-Score': 0.8374510765159967}
Classification Report:
               precision    recall  f1-score   support

        crime       0.94      0.53      0.67       165
   propaganda       0.85      1.00      0.92      2432
     religion       0.96      0.33      0.49        79
    terrorism       0.91      0.61      0.73       175
uncategorized       0.00      0.00      0.00       193
       weapon       0.98      0.86      0.92       438

     accuracy                           0.87      3482
    macro avg       0.77      0.56      0.62      3482
 weighted avg       0.83      0.87      0.84      3482
```

Figure 3: Output of Logistic regression

The network analysis identifies the top influencers in the pro-ISIS Twitter network based on the number and strength of their connections. The nodes represent Twitter users, with larger, more central nodes being the most influential. The green nodes, such as Rami AlLolah and Nidalgazaui, appear to be the key influencers with many connections radiating out to other users in the network.

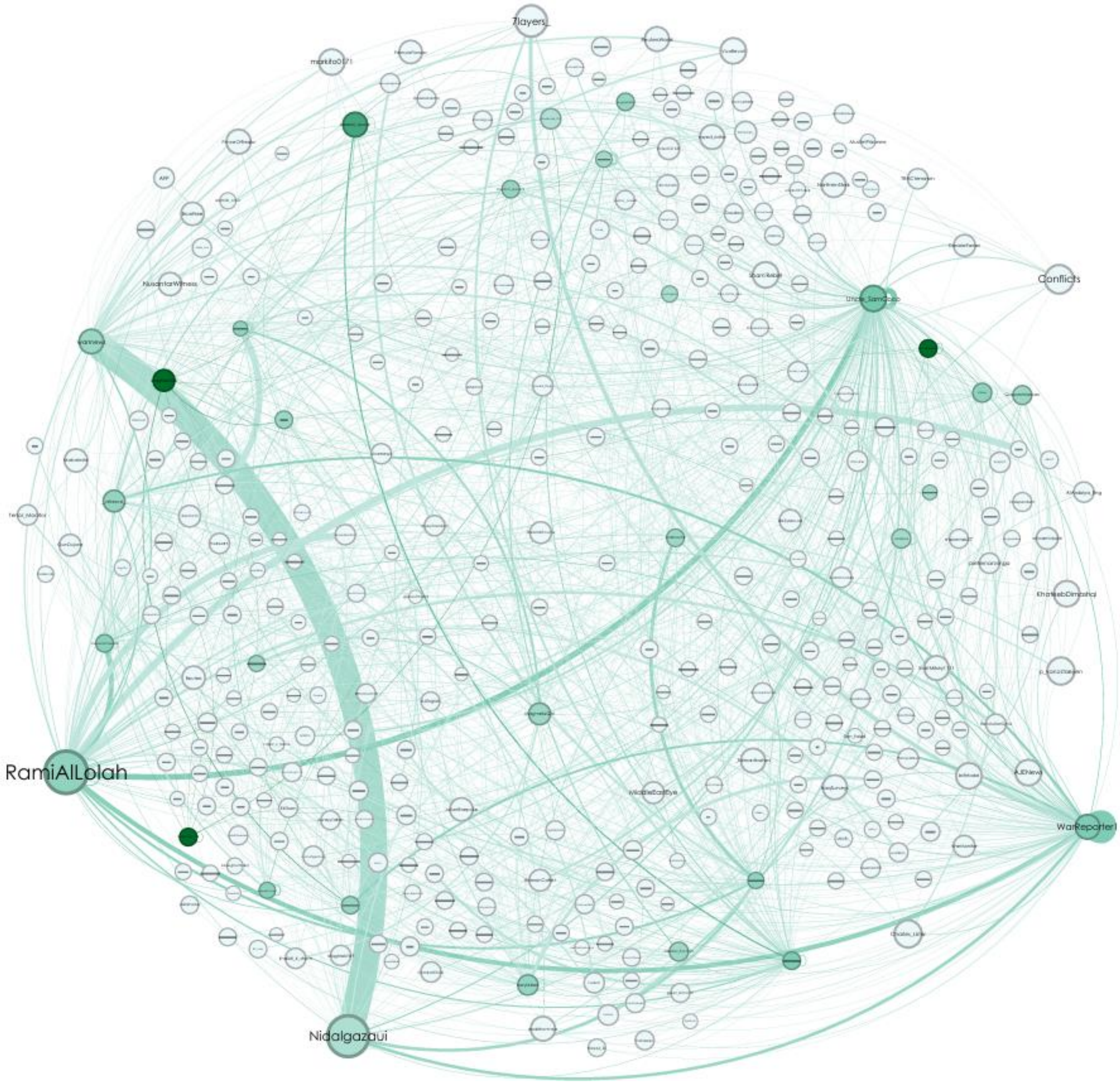


Figure 4: Top influencers in the pro-ISIS twitter network

Conclusion and Limitations:

The results from using ‘Empath’ library suggest that ISIS supporters on Twitter primarily engage in spreading propaganda, with some discussions of weapons, terrorism, crime, and religion. The high performance of the sentiment analysis model in identifying propaganda tweets indicates that ISIS fanboys consistently use language patterns associated with promoting their extremist ideology and narratives. However, recall and F1-score are low for the uncategorized class, suggesting the model struggles to identify those samples. The crime class also has mediocre recall, indicating some crime samples are misclassified. Small differences in support across classes could introduce bias.

The Social Network Analysis identified the most influential accounts in the pro-ISIS Twitter network using centrality measures. Closeness centrality and in-degree ranking revealed key nodes with high connectivity. The Louvain algorithm detected tightly knit subgroups within the network. Visualization choices like node color, size, and labels based on centrality and in-degree highlighted important actors while managing visual complexity. However, the dataset is a partial network sample and may not represent the full Twitter network on this topic, potentially introducing biases. Expanded data collection on nodes, edges, and network coverage could enable more nuanced analysis and bias mitigation.

References:

Kaggle Dataset : https://www.kaggle.com/datasets/fifthtribe/how-isis-uses-twitter/data
Benigni MC, Joseph K, Carley KM (2017) Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. PLoS ONE 12(12): e0181405. https://doi.org/10.1371/journal.pone.0181405

Ali, Govand A., "Identifying Terrorist Affiliations through Social Network Analysis Using Data Mining Techniques" (2016). *Information Technology Master Theses*. 1. https://scholar.valpo.edu/ms_ittheses/1

Peele, Elizabeth. *Forming Your Terrorist Network: Isis, Twitter, and the Terrorist Propaganda Campaign*. 2015. https://doi.org/10.17615/dnph-vm42