State-Backed Information Operations Analysis Using Pre-trained Transformer-Based Models

Hailie Nguyen hln2020@nyu.edu

Jon Xufan Ma xm595@nyu.edu

Yichen Shi ys5538@nyu.edu

Ghana Shyam Bandi gb2705@nyu.edu

Abstract

Leveraging pretrained transformer-based models and clustering techniques, we establish a generalizable approach for detecting suspicious content and identifying state-backed information operations (IO) on Twitter. Additionally, we implement narrative extraction techniques to summarize and identify overarching themes within state-backed IOs. Results show that content pertaining to IO accounts are noisy, warranting further investigation beyond tweet content to detect information operations. Nonetheless, finetuning pre-trained models on more IOs can lead to an increase in model performance.

1 Introduction

State-backed information operations (IO) involve coordinated efforts, including fake accounts, disinformation, propaganda, and deceptive tactics, orchestrated by governments or state-sponsored entities. They aim to shape narratives, create confusion, and manipulate perceptions, especially on social media and online platforms. This tool has become prominent for state actors to influence public opinion and advance geopolitical interests. Understanding the scope and methods of state-backed information operations is crucial for safeguarding online discourse integrity and protecting the democratic foundations of the digital age.

2 Related Work

Related work on the topic of state-backed information operations includes a range of studies published since Twitter launched their first archive of foreign information operations in 2018. Barrie et al. [1] examined domestic engagement with IO accounts in Saudi Arabia and reveal that domestic engagement with IO tweets was significantly lower than that with the average Saudi Twitter user. DiResta et al. [2] delved into the activity analysis and narrative extraction of a state-backed operation attributed to the Saudi Arabian digital marketing firm Smaat, and uncovered tweets criticizing Qatar's government and tweets criticizing Jamal Khashoggi, among other themes. These studies inspire our research by highlighting the importance of understanding state-sponsored information operations and how they shape digital-age narratives.

3 Approach

Our project's primary goal is to create a classification model to identify state-backed information operations on Twitter. We utilize existing pretrained transformer-based models to establish a generalizable approach for detecting suspicious content associated with such operations. Additionally, we aim to implement clustering and narrative extraction techniques to summarize and identify overarching themes within state-backed information operations.

NYU DS-GA-1011 NLP with Representation Learning Final Report

4 Experiments

We fine-tune Ada and BERT using two approaches and benchmark their performance across four metrics. Additionally, we simulate a real-world scenario through a third approach to evaluate BERT's performance. We experiment with various data preprocessing methods and record their impact on BERT. Details of our training data construction and experimental setup are outlined below.

4.1 Data

1. **Positive class:** Since 2018, the Twitter Moderation Research Consortium [3] has released a cohort of datasets on potential foreign information operations. These operations consist of persistent platform manipulation campaigns in violation of Twitter's platform manipulation and spam policy. Manipulation that Twitter can reliably attribute to a government or state-linked actor is considered an information operation (IO). Tweet from the following operations constitute our positive class:

   - Russia East Africa (REA) (December 2021) - 50 Accounts
   - People's Republic of China - Xinjiang (CNHU) (December 2021) - 2048 Accounts
   - Russia IRA North Africa (RNA) (December 2021) - 16 Accounts

   The datasets encompass multiple languages, numerous URLs, a high volume of retweets, and special characters like iOS emojis and text-based emoticons. They also feature duplicates and near-duplicates, deliberately crafted to evade duplication detection (e.g., 'british vlogger the life of #xinjiang #uygurs and other minorities in china lcqg' and 'british vlogger the life of #xinjiang #uygurs and other minorities in china wnrp'). Despite some tweets being noise without clear narratives, we retain them for the project, assuming all tweets are associated with IOs due to the challenge of removing random tweets.

2. **Negative class:** We use the following data sets to construct our negative class:

   - Random Twitter Dataset [4]: This dataset consists of 1.6 million random tweets, mainly in English. Its diverse content reflects the varied nature of user-generated content. The tweets from this dataset are used as negative, non-IO related tweets, representing informal, personal content.
   - Diplomatic Discourse Online - Twitter [5]: This data set consists of tweets from 500+ Russian and Chinese diplomats in multiple languages. They serve as a political, non-IO focused resource. We combine these tweets with tweets from the Random Twitter Dataset to form our negative class.

   The non-IO data from the Random Twitter Dataset [4] is entirely in English. While it shares characteristics of tweets such as hashtags and handlers, it contains fewer emojis and URLs. In contrast, the non-IO data from Diplomatic Discourse Online - Twitter [5] more closely resembles IO data, containing multiple languages, numerous emojis, and URLs.

4.2 Data Preprocessing

In data preprocessing, we initially perform a three-step procedure on both the training and test datasets: dropping NaNs from tweets, removing duplicate tweets, and eliminating single and double quotes. We report the performance of both Ada and BERT using this basic preprocessing framework. Additionally, an advanced preprocessing procedure is applied, manipulating handlers, emojis, URLs, and retweet signals. For the training dataset, near-duplicate removal using TF-IDF and MinHash is also employed. We report and compare BERT's performance between these approaches.

4.3 Experimental details

1. **IO Classification/Detection**

   We construct the training data for Ada and BERT through various approaches. To address class imbalance, in Approaches 1 and 2, we balance the positive and negative classes during sampling. Approach 3 simulates a real-world scenario where the majority of tweets do not belong to an IO. Due to cost constraints, we did not explore Approach 3 with Ada and defer it to future work. We assess their performance using the Russia North Africa (RNA) dataset.

   (a) **Training data**

Model	Ada0	Ada1	BERT0	BERT1	BigBERT
Approach 1	✓	-	✓	-	-
Approach 2	-	✓	-	✓	-
Approach 3	-	-	-	-	✓

Table 1: Summary of finetuned models

i. **Approach 1:** Full REA dataset and balanced non-IO tweets
   - **Positive class:** 7,723 tweets from the REA dataset
   - **Negative class:**
     - 3,500 random sample tweets [4]
     - 3,500 random samples from Russian diplomats [5]

ii. **Approach 2:** Sample of REA and CNHU datasets and balanced non-IO tweets
   - **Positive class:**
     - 1,000 random samples from the REA dataset
     - 1,000 random samples from the CNHU dataset
   - **Negative class:**
     - 1,000 random sample tweets [4]
     - 500 random samples from Russian diplomats [5]
     - 500 random samples from Chinese diplomats [5]

iii. **Approach 3:** Full CNHU dataset and 1.4 million random tweets
   - **Positive class:** 30,830 tweets from the CNHU dataset in 2020 and 2021
   - **Negative class:** 1,440,000 random tweets [4]

(b) **Test data**
   - **Positive class:** 1,000 random samples from the RNA dataset
   - **Negative class:**
     - 500 random samples from Random Tweet Dataset [4]
     - 250 random samples from Russian diplomats [5]
     - 250 random samples from Chinese diplomats [5]

2. **Clustering & Theme Extraction**

   Our framework for clustering and theme extraction follows the following three steps:

   (a) **Embedding Generation:** We leverage OpenAI's text-embedding-ada-002 model to create meaningful representations of IO tweets.

   (b) **Clustering:** To organize and group similar embeddings, we employ K-Means clustering, a widely-used algorithm for partitioning data into clusters.

   (c) **Theme Extraction:** To extract thematic information from clustered data, we sample 5 tweets from each cluster and leverage text-davinci-003 for summarization.

4.4 Evaluation method

We use accuracy, F1 score, precision, and recall to evaluate our model performance.

- Accuracy $= \frac{\text{True Positives}+\text{True Negatives}}{\text{Total Population}}$ Precision $= \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$
- Recall $= \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$ F1 Score $= 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

5 Results

5.1 Classification results

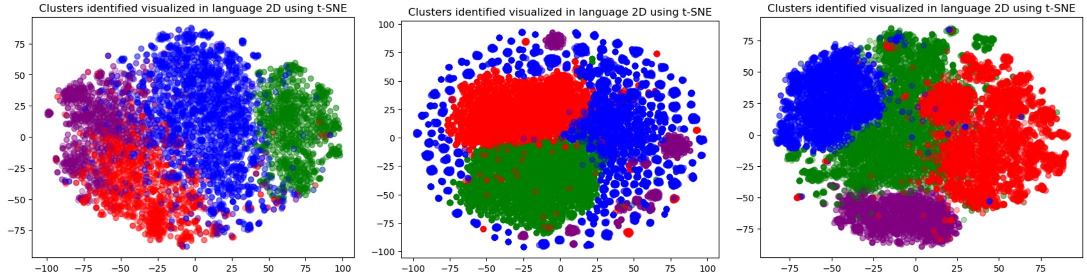
Metric/Model	Ada 0	Ada 1	BERT 0	BERT 1	BigBERT
Precision	0.993	0.967	0.80	0.95	0.97
Recall	0.493	0.969	0.70	0.95	0.97
Accuracy	0.738	0.969	0.70	0.95	0.97
F1	0.659	0.968	0.67	0.95	0.97

Table 2: Validation results on a hold-out dataset from the training set.

Metric/Model	Ada 0	Ada 1	BERT 0	BERT 1		BigBERT
				Basic Preprocessing	Advanced Preprocessing	
Precision	0.872	0.933	0.70	0.77	0.78	0.73
Recall	0.294	0.581	0.62	0.71	0.73	0.86
Accuracy	0.622	0.767	0.62	0.71	0.73	0.86
F1	0.440	0.716	0.58	0.70	0.72	0.79

Table 3: Results when generalizing to a new operation (Russia North Africa - RNA).

5.2 Clustering results



(a) Russia East Africa (REA)

(b) China Xinjiang (CNHU)

(c) Russia North Africa (RNA)

Figure 1: Clusters of embeddings generated using text-embedding-ada-002

Table 4: CNHU's common themes extracted by text-davinci-003

Cluster	Theme	Sample Tweets
0	Xinjiang's counter-terrorism measures protect human rights.	@ErkinSidick @Dolkun_Isa @Erkin_Azat @adrianzenz #Xinjiang's counter-terrorism measures protect human rights https://t.co/pzp6NisIsC Xinjiang's counter-terrorism measures protect human rights #xinjiang https://t.co/HQjWl6hONn
1	Admiration for someone's remarkable memory and a celebration of nature's beauty.	他有惊人的记忆力. https://t.co/1q3J1b9Etv 浊酒一杯温如言
2	Hard work and appreciation for parents.	Whats on tonight. Your parents are still working hard for you; this is the reason you are strong today.
3	Xinjiang is a safe and vibrant place, with a rich culture and history; efforts of the Chinese government to combat terrorism and extremism in the region.	CGTN documentary exposes terrorism, extremism in Xinjiang, says China's Foreign Ministry https://t.co/Xr6VCDF7yp CUXI ##Xinjiang A documentary filmmaker captured the daily lives of locals in the capital city of #Xinjiang during those tumultuous times. https://t.co/cYJea4HMKd

6 Analysis

1. **Clustering Analysis**

   Our clustering analysis indicates a diverse content spectrum, encompassing tweets with distinct political narratives alongside others that appear trivial without a discernible political agenda. For example, in themes extracted from CNHU, clusters 0 and 4 appear to carry some political agenda, while clusters 1 and 2 exhibit trivial content. It is important to note that our theme extraction process relies on the content of 5 sampled tweets per cluster. The unclear exact number of clusters introduces variability in the associated themes for each cluster. Our embedding visualizations suggest potential overlapping themes among clusters, challenging the notion of clear-cut, defined themes.

2. **Data Preprocessing Analysis**

   For data preprocessing, we observe a decrease in accuracy when URLs or emojis are removed. This suggests that these elements may carry significant contextual information relevant to the classification task. Conversely, removing handlers and retweet signals improves accuracy. This could be due to the reduction of noise in the data, as handlers and retweet signal may not contribute much to the tweet content. Removing near duplicates in the training set at a threshold of 0.8 also enhances overall accuracy. This implies that near duplicates might cause the model to overfit on repeated patterns or biased representations of the data. In addition to basic preprocessing, removing handlers, retweet signals, and dropping near duplicates collectively improves the accuracy, recall, and F1 score by approximately 2%.

3. **IO Classification/Detection Analysis**

   Results on a novel IO test set indicate that both Ada and BERT are more effective in predicting non-IO tweets compared to IO tweets. 4 This could be due to the noisy nature of IO tweets, which can contain trivial content, as suggested by our clustering and theme extraction results. Interestingly, training both Ada and BERT on a more diverse set results in improved performance on the novel test set, despite the training data's smaller size. This suggests that broader training could enhance the models' effectiveness in detecting novel operations.

   BigBERT's remarkable performance in accuracy, recall and F1 score may be attributed to a substantial amount of information learned from random tweets. However, the improvement is not drastic considering the significantly larger training set. In fact, BERT1 and Ada1 appear to outperform BigBERT in terms of precision. Nonetheless, BigBERT's performance reveals an interesting insight: increasing the size of non-IO data can enhance the model's ability to detect IO content.

4. **Further Analysis**

   Our results show that Ada and BERT perform better when trained on two IOs compared to training on just one, despite the smaller data size. This raises the question: will adding more IOs to the training data improve the model's performance? To explore this, we follow Approach 2, maintaining the same non-IO dataset. In addition to the two IOs used during training (REA and CNHU), we add data from four additional IOs. We keep the sample size for all six IOs, and the total size matched the non-IO counterparts. Training is again conducted on BERT and Ada, and testing on a new operation (RNA). However, we find no improvement in performance compared to training on just two IOs. We suspect this might be due to a shift in the tweet language distribution between the training and test datasets.

   Figure 3 in the Appendix show the percentage of tweet languages in our IO data. REA's lack of diversity in languages could cause the models to underperform on unseen languages. In the dataset with two IOs (REA + CNHU), the data includes tweets in a wider range of languages. It also shares some dominant languages with the test set, such as English (62%) and French (12%), providing sufficient data for the model to generalize well to the test set. For the six-IO dataset, while it includes almost every language, some dominant languages in the test set, like French (20%), have only a 5% representation in the training set. Given our dataset's limited size, even though we included more languages to enhance the model's generalization ability, the training might not have been comprehensive enough.

7 Discussion and Conclusion

In conclusion, our results show that content pertaining to IO accounts are noisy, warranting further investigation beyond their tweet content to detect such operations. Our model performance in IO content detection improves when they are trained on two IOs (REA and CNHU) compared to one (REA), though they are more effective at identifying non-IO tweets than IO tweets, as indicated by the relatively high precision combined with low recall. This can be attributed to the noisy content of the IO data, which includes many trivial tweets. To enhance our models, we need more diverse training data that better aligns with the test data distribution, or a larger dataset containing a sufficient number of tweets in various languages, enabling the model to learn and generalize to unseen tweets. In general, many factors outside of tweet content could impact the detection of information operations, such as account activity, spam level, interactions with other accounts, among others. Future work may consider these aspects in their IO detection analysis for improved performance.

References

[1] C. Barrie and A.A. Siegel. Kingdom of trolls? influence operations in the saudi twittersphere. *Journal of Quantitative Description*, 1:1–41, 2021.

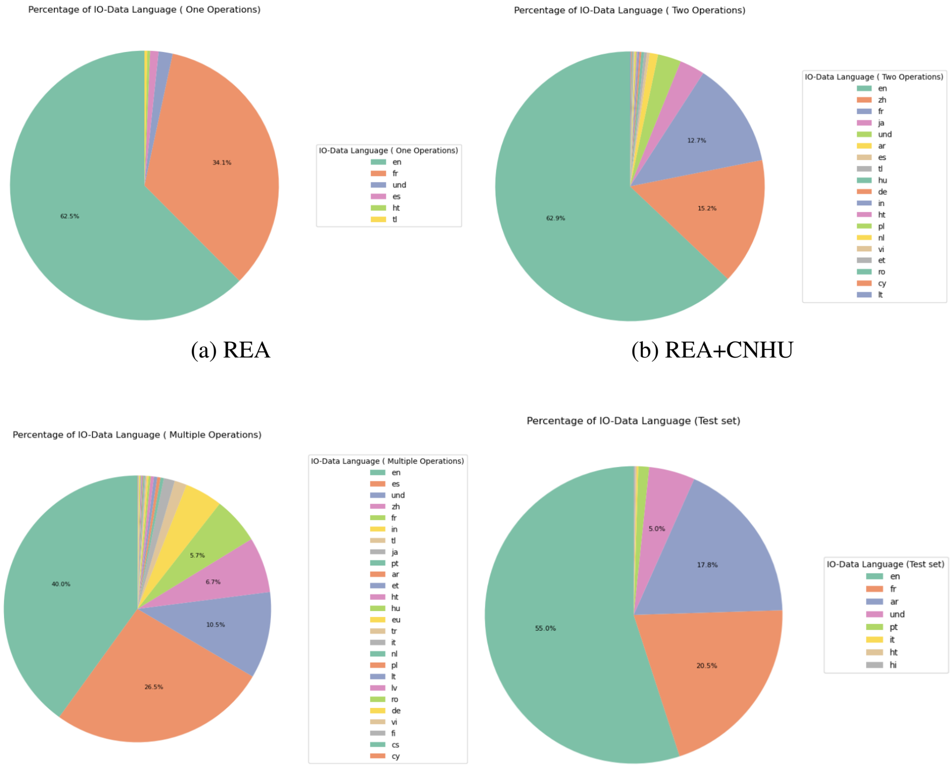
[2] K.H. Renée DiResta, Shelby Grossman and Carly Miller. Analysis of twitter takedown of state-backed operation attributed to saudi arabian digital marketing firm smaat, 2019.

[3] Twitter moderation research consortium, 2023.

[4] Majid Ahmad Khan. 1.6 million random tweets. *1.6 million random tweets collected from Twitter two years ago (pre-elon's era)*, 2021.

[5] Keeley Erhardt. Diplomatic Discourse Online - Twitter, 2023.

A Appendix



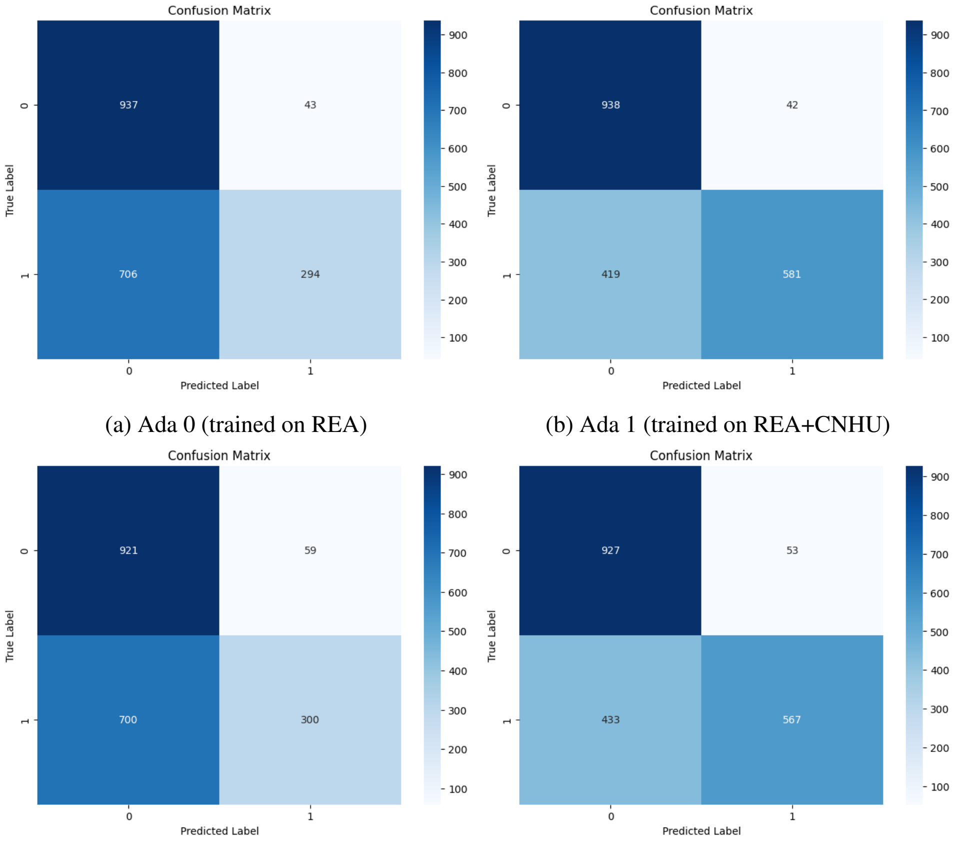
(a) REA

(b) REA+CNHU

(a) Six IOs

(b) Test (RNA)

Figure 3: Percentage of tweets by language for various IOs



(a) Ada 0 (trained on REA)

(b) Ada 1 (trained on REA+CNHU)

(c) BERT 0 (trained on REA)

(d) BERT 1 (trained on REA+CNHU)

Figure 4: Confusion matrix when tested on RNA