IADIS International Journal on WWW/Internet
Vol. 20, No. 2, pp. 48-64
ISSN: 1645-7641

APPLYING SOCMINT TO EXTRACT CYBER THREAT INTELLIGENCE FROM THE RUSSIA-UKRAINE CONFLICT

Bipun Thapa
Marymount University, 2807 N Glebe Rd, Arlington, VA 22207, United States of America

ABSTRACT

The paper applied SOCMINT (Social Media Intelligence) techniques to discover cybersecurity-related information from the contemporary Russia-Ukraine conflict. Using open-source tools and APIs, datasets created were assessed through topic modeling, thematic analysis (word cloud), Logit function, and neural network classification. The topic modeling and word cloud yielded trifling insights, but Logit and neural network classifier, MLP, suggested statistically significant features that were important to the outcome of the tweets with reasonable accuracy of 91%. Through the use of synthetic data (GaussianCopula) and feature selection(stepAIC), the model was extended to improve accuracy, which resulted in 96% accuracy, though, such competent performance requires further investigation. While deciphering the right intelligence is a challenge due to the unruly nature of social media, this nascent technique can be helpful with the proper framework and approach.

KEYWORDS

Social Media Intelligence, Cyber Threat Intelligence, Topic Modeling, Machine Learning, Synthetic Data, Logit

1. INTRODUCTION

There is a vital necessity for robust Cyber Threat Intelligence (CTI) due to evolving attacks (Husari et al., 2019), which can derail normalcy for organizations and governments. Accumulating threat intelligence is generally a prelude to solution discovery; thus, organizations and vendors are more acquiescent in sharing their knowledge and collaborating responses, albeit with varying efficacy(Sauerwein et al., 2021). Social media intelligence, an extension of Open-source intelligence (OSINT), provides a new conduit to intelligence for CTI (Omand et al., 2012)by sifting through social media data to measure the pulse of the cyber security posture.

Various vendors postulate (*Using Social Media (SOCMINT) in Threat Hunting*, n.d.) (Using Social Media (SOCMINT) in Threat Hunting, no date) the integration of SOCMINT in the existing OSINT ecosystem where the latter provides context to the cyber threats in question like time, location, and trends, though for SOCMINT to be consequential, it has to be complete, accurate, relevant, and timely. From a defender's perspective, employing SOCMINT capabilities is another tool in the arsenal to protect assets (Kropotov & Yarochkin, 2019) (Kropotov and Yarochkin, 2019).

Easily deployable Application Programming Interfaces (APIs)(*Getting Started — Tweepy 4.2.0 Documentation*, n.d.)(Getting started — tweepy 4.2.0 documentation, no date) for social media(*APIs for Scholarly Resources | Scholarly Publishing - MIT Libraries*, n.d.), and usability of programming languages like Python(Lakshmi, 2018) and its burgeoning libraries enable the construction of customized CTI frameworks. More notably, the open-source and inexpensive resource rollout acts as a catalyst for the adoption.

The storyline of the Russia-Ukraine conflict has been Ukraine's admirable resistance (Khan, 2022) to all vanguards, including its effective social media management to educate the global diaspora and combat Russian cyber attacks (*Ukraine's Digital Ministry Is a Formidable War Machine | WIRED*, 2022). Willing digital volunteers have assisted in fighting for Ukraine to win the 'information war'(McLaughlin, 2022), and Western media often tout that Ukraine is 'winning' the social media battle(Trouillard, 2022).

With the Russia-Ukraine conflict as the context, this exploratory research attempts to extricate cyber-related intelligence to find a meaningful understanding of the subject. The study aims to satisfy the SOCMINT concept by curating the dataset from publicly available cybersecurity information (Twitter), analyzing it, and synthesizing its essence.

President Volodymyr Zelenskyy and President Vladimir Putin are the lead thespians of the conflict from each side, and by leveraging their following on Twitter, the cyber-related dataset with ample entries will be curated. The datasets will be the basis of analysis; however, it is not established to prove the active hypothesis that Ukraine is winning the social media war. An intermittent outage of social media in Russia would under sample tweets; hence, it would not deliver a realistic foundation for such commentary (Milmo, 2022).

Adding synthetic data or fake data could add learning opportunities for the model as the injection of new data based on existing characteristics provides an opportunity to understand complex relationships - it is predicted that it will be more relevant in future modeling, as 60% of all analytics will use synthetic data by 2024(Toews, 2022)

Therefore, this research only attempts to create a repeatable, modular framework that is a proof-of-concept for the SOCMINT-infused CTI framework and potentially derive valuable experiences. This research endeavors to address three research questions:

RQ1 - Can Cybersecurity intelligence be derived from Twitter data?

RQ2 -What independent variables are important alliance indicators for Ukraine and Russia, and can they correctly predict the alliance?

RQ3 - Can adjusted feature selections and synthetic data improve model accuracy?

In Section 2, the paper investigates contemporary work on the subject, followed by Section 3, which describes the methodology of the research. Section 4 presents the results and analyzes to discover limitations and future scope (Section 5), and finally, Section 6 summarizes the essence of the findings.

2. RELATED WORK

As part of the intelligence-gathering family, SOCMINT, coined by Sir David Omand and fellow researchers (Ivan et al., 2015), is a combination of tools that leverages social media tools to uncover meaningful intelligence to aid the investigation. OSINT or open-source intelligence is often put in the same category as SOCMINT, but one key attribute separates the two, the latter can analyze both public and private data, whereas the former is strictly focused on publicly available data (*Social Media Intelligence*, n.d.).

Assessing the varying emotions of participants is plausible due to the plentiful crowd-sourced data, real-time perspicuity, and supposing intents from diverse groups. Eclectic techniques are inscribed in intelligence lexicons; Signals Intelligence (SIGINT), Imagery Intelligence (IMINT), and Geomatics (GEOINT) are pertinent examples. The SOCMINT, by nature, could collectively capture a combination of the intelligence above; a single entry could have an image, geographical information, and textual intelligence (Mahood, 2015).

To successfully identify perished Russian soldiers in Ukraine, Clearview AI, a US-based company, scrapped social media images to match the pictures of dead soldiers as a courtesy to the ailing family (Dave, 2022). Thailand maintains a dedicated task force to continuously monitor the public dissent towards monarchy and political groups, with a system structured to reward the whistleblowers (*Three Surveillance Technologies That Protesters Need to Know about - IFEX*, 2019). Egypt, around mid-2014, with suspicious premises to determine ‘security hazards,’ facilitated technologies to monitor social media with insidious purpose. Venezuela imprisoned numerous people by observing social media discord that was harmless in the other jurisdictions, like posting dollar rates or personal opinions differing from the ruling body.

‘Kansas City No Violence Alliance, an initiative to comprise a predictive instrument for future offenders, uses social media intelligence in its model (*Social Media Intelligence*, n.d.). Squeaky Dolphin, a presumed GCHQ product, allegedly compromised data cables to monitor comments about prominent British personalities through YouTube and Facebook content (Kelion, 2014). China’s social credit system incorporates payment delinquency, public habits, non-compliance with local laws, and social media behavior to control individuals’ access to society (Kobie, 2019).

To improve security posture and confidence, or conversely, to create chaos and uncertainty, the use cases of SOCMINT are equally applicable because, for the most part, the analysis is conducted on public data. SOCMINT sources are disorderly and informal, from tweets, blogs, forum posts, chats, or any avenues available (Forrester & Hollander, 2016). In addition, excellent open-source network visualization tools, intelligence gathering APIs, and forensic instruments make it easier to collect information.

TWINT is an OSINT tool that utilizes scraped data from Twitter for specific criteria like usernames or hashtags to comprehend ongoing trends (Kropotov & Yarochkin, 2019). The major tech companies allow APIs to connect to their environment, for example, Tweepy for Twitter and PRAW for Reddit, which makes data analysis easy, albeit they impose rate limits so the insights cannot be visualized in their entirety(*Code Snippets — Tweepy 3.5.0 Documentation*, n.d.).

At the onset of the Russia-Ukraine conflict, considerable hacktivists and the cyber army have taken sides though there is no way to validate the claims. Such parties induce cyber warfare, attacks on supply-chain, DDoS to major banks and government sites, and data breaches to expose tactics. Recorded Future, an intelligence-gathering entity(Vail, 2022), explored the available information to understand the alliance as illustrated in Table 1 below with their corresponding Twitter handles, except that most Russian groups have had their account suspended due to increasing violations of community guidelines. This member list, although it may or may not be comprehensive, was synthesized from credible news sources and intelligence-gathering sites like Recorded Future, Anomali, and ThreatQuotient.

Table 1. Cyber Group Alliance

Group	Alliance	Twitter Handle
Anonymous	Ukraine	@YourAnonOne
IT Army of Ukraine	Ukraine	@ITarmyUA
Belarusian Cyber Partisans	Ukraine	@cpartisans
Secjuice	Ukraine	@Secjuice
Conti leaks	Ukraine	@ContiLeaks
RedBanditsRU	Russia	@RedBanditsRU
Sandworm	Russia	unknown or suspended
Freecivilian	Russia	unknown or suspended
Digital Cobra Gang (DCG)	Russia	unknown or suspended
Zatoich	Russia	unknown or suspended

Ordinary Linear Squares (OLS) Regression model can be incorporated to understand the factors that impact Twitter behavior(Costa et al., 2021); could variables like the length of the tweet, mentions, or hashtags be significant when predicting the alliance by calculating its statistical significance? In a study of one hundred Twitter users analyzing online behavior, with an accuracy of 75.13 percent, the authors were able to predict their preferences and deeper suites of personalities like openness and agreeableness (Mahajan et al., 2022). Value systems or personal beliefs could be an essential predictor of why someone engages in retweets and are sometimes as effective as traditional machine learning models like Random Forest, XGB, and Logistic Regression(Kakar et al., 2021).

The presence of pro and anti-Kremlin bots are plentiful, and both parties were invariant in promoting their desired accounts on Twitter with one exception - pro-Kremlin's source of truth was their state media, and anti-Kremlin derived their content from areas that the state could not control (Stukal et al., 2019). Infodemic on social media is another challenge as the dissemination of information is rapid with credibility issues (Wang et al., 2021).

There is some argument in the literature that while SOCMINT provides some situational awareness in the aftermath of the event (Dover, 2020), it is incapable of predict the immediate threats and creating a significant foundation to yield intelligence (McLoughlin et al., 2020**).** In addition, there are privacy issues even with publicly available data that could be ripe for misuse, and adversaries like Al-Shabaab are highly active on social media providing an intelligence-gathering framework for their use as well makes SOCMINT a double-edged sword (Momi, 2021).

The use cases of SOCMINT are abreast in the literature - however, organizing data and deducing intelligence is difficult the dynamism of data, lack of validity in the public domain, and possibilities of duplication making it somewhat risky to base the foundation. Nevertheless, we are addicted to public perceptions, and having some synthesis of the opinions could be valuable. The Russia-Ukraine conflict is the most impactful event to start in 2022, and surveying the cyber public opinion is beneficial with available tools to measure the pulse of space provides value.

Leveraging copula-based synthetic enhancement Meyer et al., were able to decrease the model error by up to 75% in a number of experiments for climate and weather prediction(Meyer et al., 2021). With the ability to model multivariate properties, copulas can be effective in reproducing dependency structure in the original dataset (Warnes, 2021). Synthetic Data Vault (SDV), an open-source library deployable in Python is an effective process to create synthetic data that has grown in popularity in the last decade (Montanez, 2018).

3. METHODOLOGY

In this exploratory research design, the first step is to curate the dataset (regular and synthetic) from Twitter to address the research questions. President Zelensky and President Putin lead the conflict from each side; therefore, numerous alliances have backed their cause. In this case (*Tweet Object | Docs | Twitter Developer Platform*, n.d.), two separate country-specific datasets are generated with the same columns. The Twitter data is highly malleable, and #hashtags in its ecosystem provide sorting and aggregation (Laucuka, 2018) of information - a topic modeling mechanism that encapsulates similar themes.

A synthesis of popular hashtags originating from the presidents and alliances is selected by analyzing Twitter intelligence. The factor analysis reduced the hashtags to the ones that provided explicit support. The research aims to uncover cyber-related intelligence from the dispute; therefore, war-related, popular hashtags paired with cyber-related hashtags are blended to create the most relevant dataset for each country. Table II describes the mapping of hashtags to the countries.

Table 2. Cyber Hashtag Harvesting

Country	Support	Cyber
Ukraine	#ukraine,#ukraine	#cyber
	#standwithukraine	#cyberattack
	#ukrainewar	#cybersecurity
	#ukraina	#hacking
	#ukrainerussiawar,	#cyberwar
	#strongertogether, #helpukraine	#cyberwarfare
Russia	#kremlin	#cyber
	#moscow #IStandWithPutin	#cyberattack
	#istandwithrussia	#cybersecurity
		#hacking
		#cyberwar
		#cyberwarfare

Python is a powerful programming language due to its rich libraries. Tweepy, twint, and snscrape are effective APIs that pull tweets based on the prescribed criteria - in this instance, Tweepy was used due to its official alignment with the company, although rate-limiting is a nuance. The feature selection or the columns extracted from a tweet are listed below in Table 3 below.

Table 3. Feature Dictionary

Columns Extracted	Definition	Data Type
AU1R0	Alliance, R =0, U = 1	Boolean
Time Hour	Time of the tweet	Int
Followers Count	users following user	Int
Tweets	Content of the tweet	String
Length	Total characters of the tweet	Int
Location	Location for this account's profile	String
Statuses Count	Tweets issued by the user	Int
Friends Count	Users this account is following	Int
Favorites Count	Number of Tweets this user has liked in the account's lifetime.	Int
Account creation	Data account was created	Int
Retweet Count	Retweeted by other users	Int
Favorite of the tweet	Favorite of the tweet	String
Account Verified	Account confirmed by Twitter	Boolean
Listed Count (public list)	Users adding people to their list	Int

The dataset curated will be pre-processed to eliminate repetitive tags, empty columns, and integrate one-hot encoding to convert the categorical variables to numerical for model integration. Latent Dirichlet Allocation and integration of Logit helped to understand which explanatory variable is statistically significant when predicting the alliance of the tweets. The alpha or p-value for statistical significance is 0.05; any independent variable yielding this value or lower is presumed to be statistically significant (negative or positive) to the dependent variable or, in other words, influences alliance (Sperandei, 2014). The result of the data collection aims to yield four datasets. Figure 1 illustrates the methodology.

Dataset 1 - Curated Ukraine Dataset

Dataset 2 - Curated Ukraine Dataset

Dataset 3 - Composite (Dataset 1 + Dataset 2)

Dataset 4 - Dataset 3 + Synthetic Data, resulting in 20,000 total observations

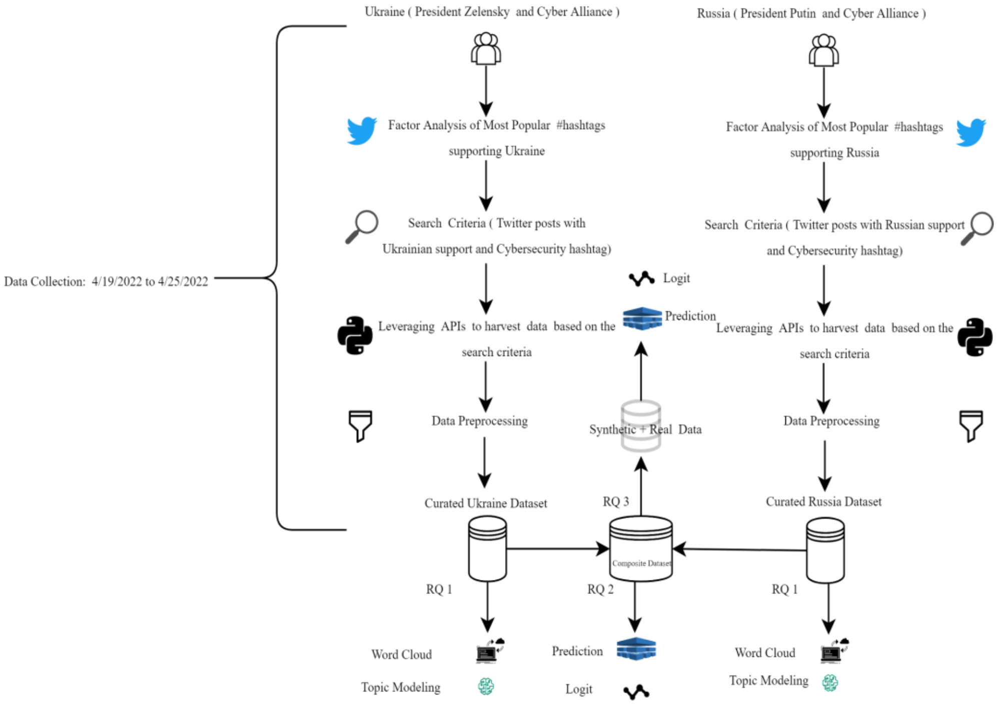

Ukraine (President Zelensky and Cyber Alliance)
Russia (President Putin and Cyber Alliance)
Factor Analysis of Most Popular #hashtags supporting Ukraine
Factor Analysis of Most Popular #hashtags supporting Russia
Search Criteria (Twitter posts with Ukrainian support and Cybersecurity hashtag)
Search Criteria (Twitter posts with Russian support and Cybersecurity hashtag)
Data Collection: 4/19/2022 to 4/25/2022
Logit
Prediction
Leveraging APIs to harvest data based on the search criteria
Leveraging APIs to harvest data based on the search criteria
Data Preprocessing
Synthetic + Real Data
Data Preprocessing
Curated Ukraine Dataset
RQ 3
Curated Russia Dataset
Composite Dataset
RQ 1
RQ 2
RQ 1
Word Cloud
Topic Modeling
Prediction
Logit
Word Cloud
Topic Modeling

Figure 1. Methodology Overview

4. RESULTS AND DISCUSSIONS

The framework with specific criteria, as defined in Figure 1, harvested 2017 tweets that aligned with Ukraine and 1007 that supported Russia. If any of the missing values from the columns were missing, the entire row of data was eliminated. The composite dataset was extended by 16,796 synthetic observations creating a final dataset of 20,000 observations with equal Russian and Ukrainian sampling.

4.1 RQ1 - Can Cybersecurity Intelligence be Derived from the Tweets?

Contrary to the initial supposition, confining data harvesting to strict criteria where the tweets had to mention programmed cybersecurity hashtags with Russia-Ukraine in the background, the intelligence yielded was not denotative. Most tweets mentioned very little about cybersecurity-related topics. An assumption that the tweets would discover pertinent information about vulnerabilities, risks, and threats due to escalating conflict was inaccurate. The intelligence gathered mostly revolved around the actors of the wars.

No meaningful intelligence was deduced. Figure 2 and Figure 3 assemble thematic visualizations of the conflict for each alliance. Both Ukrainian and Russian datasets were devoid of cyber intelligence and looked fairly similar in essence, although their hashtags advertised cyber-inclinations – the research gathered random data from the Twittersphere, but the unruly and uncorroborated tweets from randomly users are of major concern.

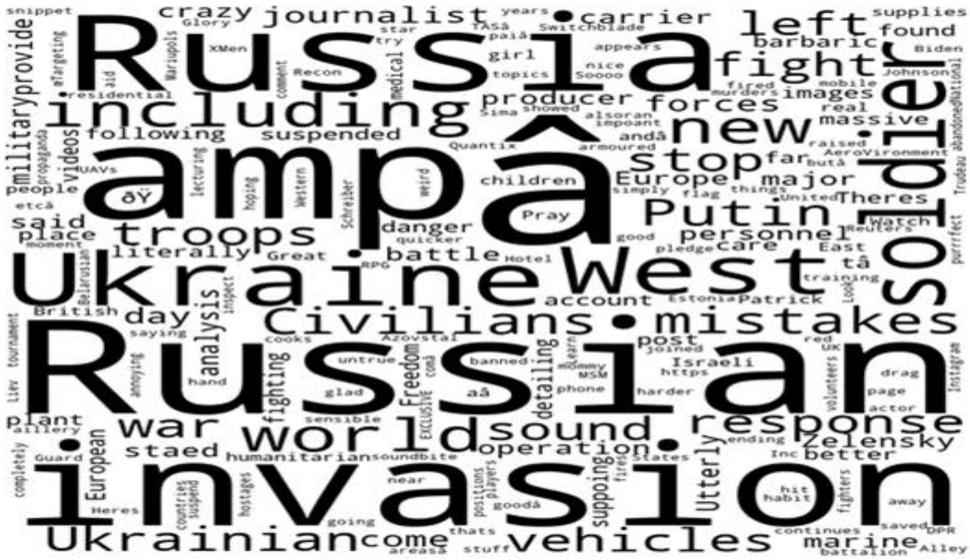


Figure 2. Word Map of Ukrainian Tweets

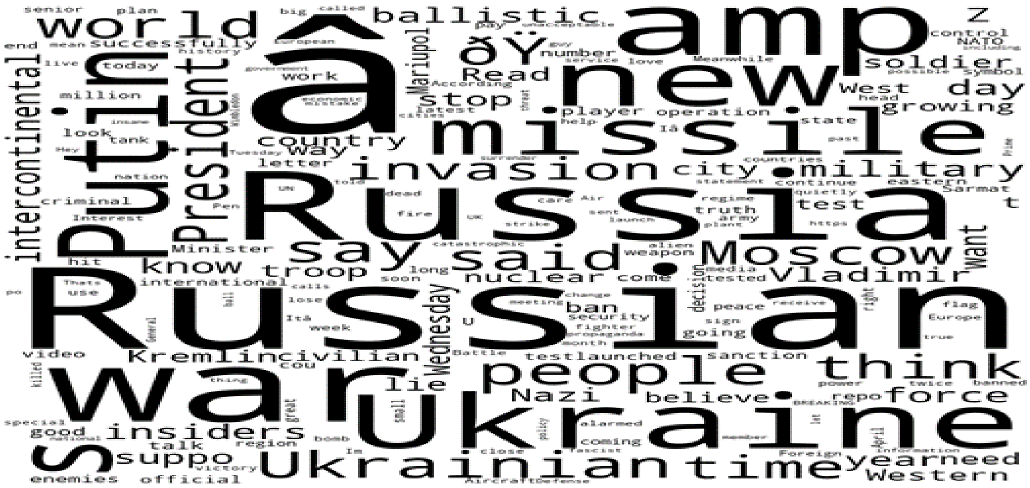
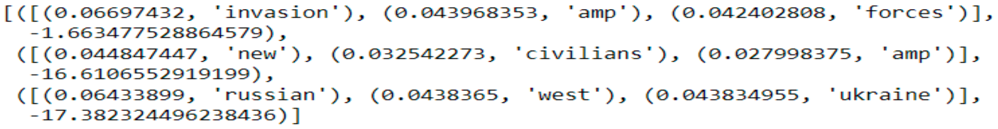


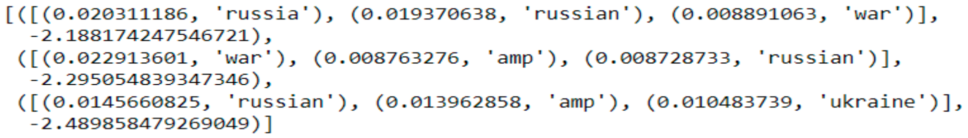
Figure 3. Word Map of Russian Tweets

Another way to explore textual content is by LdaMOdel(Blei, 2003), which uses a generative probabilistic model to classify discrete data. For the Ukrainian tweets, the texts were classified into three topics and three subtopics, as shown in Figure 4. Likewise, Figure 5 shows the classification of Russian tweets



```
[([(0.06697432, 'invasion'), (0.043968353, 'amp'), (0.042402808, 'forces')],
  -1.663477528864579),
 ([(0.044847447, 'new'), (0.032542273, 'civilians'), (0.027998375, 'amp')],
  -16.6106552919199),
 ([(0.06433899, 'russian'), (0.0438365, 'west'), (0.043834955, 'ukraine')],
  -17.382324496238436)]
```

Figure 4. Topic Modeling (LDA) Ukrainian Tweets



```
[([(0.020311186, 'russia'), (0.019370638, 'russian'), (0.008891063, 'war')],
  -2.188174247546721),
 ([(0.022913601, 'war'), (0.008763276, 'amp'), (0.008728733, 'russian')],
  -2.295054839347346),
 ([(0.0145660825, 'russian'), (0.013962858, 'amp'), (0.010483739, 'ukraine')],
  -2.489858479269049)]
```

Figure 5. Topic Modeling (LDA) Russian Tweets

Both models did not exhibit cyber content relevance, most topics were rudimentary and devoid of cyber-specific topics, further confirming that cybersecurity intelligence was negligible.

4.2 RQ2 - What Independent Variables are Important Alliance Indicators for Ukraine and Russia, and can they Correctly Predict the Alliance?

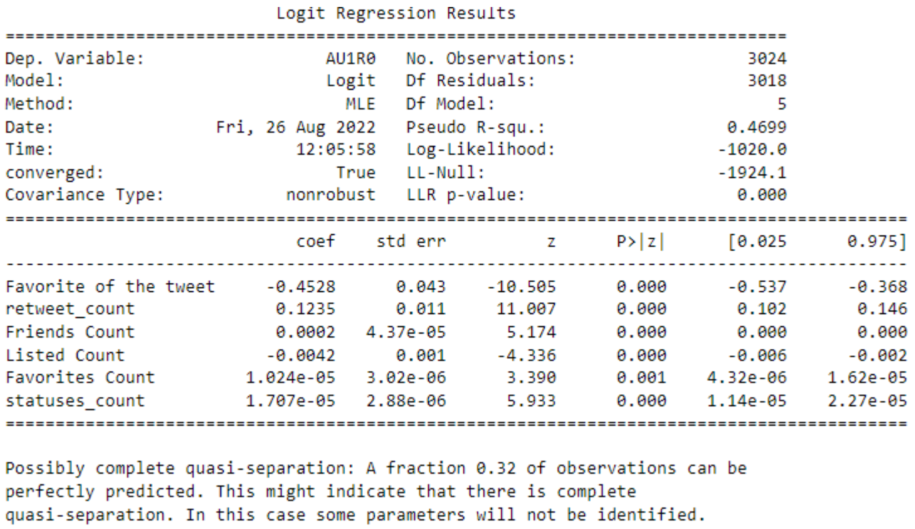
A composite dataset was created with Russia and Ukraine entries; the Ukrainian affiliation was denoted '1' or 'True' in the 'AU1R0' column and '0' and 'False' for Russian affiliation. Logit Regression, a binary classification model with conditional probability (Taboga, n.d.) was used to exhibit the relationship between 'AU1R0', a dependent variable, and the numerous explanatory variables. 'AU1R0' denotes the alliance - if a tweet has a '1' value in this variable, it means the tweet explicitly supported Ukrainian cyberspace initiatives.

Multicollinearity using Variance Inflation Factors (VIF) was used to eliminate competing features. In Table IV below, 'Time Hour', 'Length', and 'Account Creation', whose VIF score is greater than 4, are assumed to be noise in the modeling - hence eliminated to explain the dependent variable, 'AU1R0'. After iterating the model, it produced a high p-value for 'Followers Count' and 'Account Verified', which were deemed insignificant to the model, so it was eliminated as well.

Table 4. Multicollinearity Assessment (VIF)

Features	VIF Score
Time Hour	5.72
Followers Count	1.15
Length	4.52
Statuses Count	1.60
Friends Count	1.24
Favorites Count	1.65
Account creation	10.15
Retweet Count	1.00
Favorite of the tweet	1.43
Account Verified	1.63
Listed Count (public list)	1.25

If the p-value is ≤0.05, then the independent variable is significant, and thus, will impact the direction of the dependent variable, negatively or positively. As shown in Figure 6, the p-value of ‘Favorite of the tweet’, ‘retweet_count’, ‘Friends Count’, ‘Listed Count’, ‘Favorites Count’, and ‘statuses_count’ are important independent variables in predicting the Ukrainian alliance.

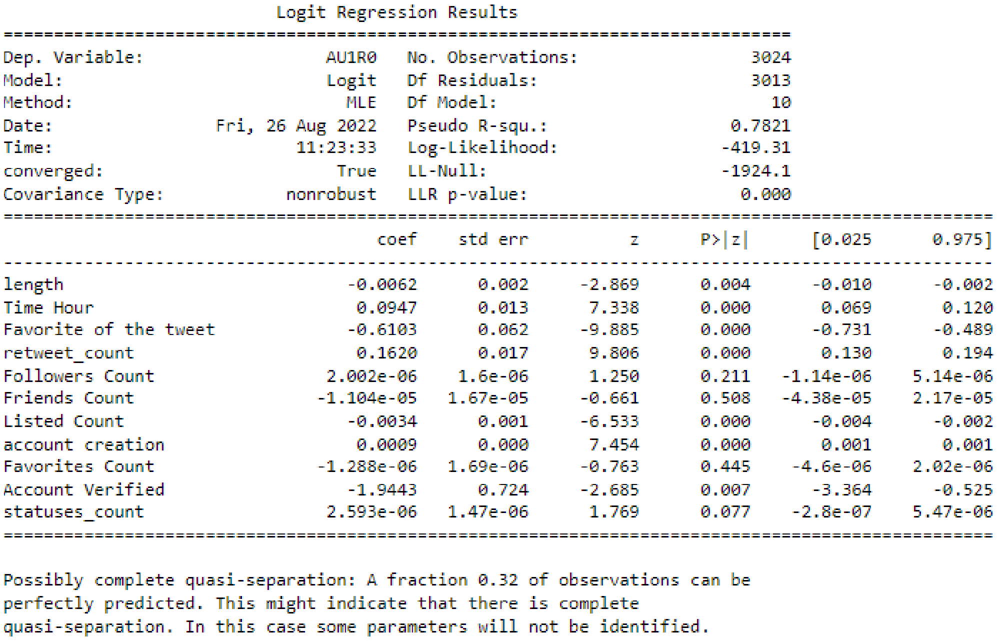


```
                           Logit Regression Results
==============================================================================
Dep. Variable:                   AU1R0   No. Observations:                 3024
Model:                           Logit   Df Residuals:                     3018
Method:                            MLE   Df Model:                            5
Date:                 Fri, 26 Aug 2022   Pseudo R-squ.:                  0.4699
Time:                         12:05:58   Log-Likelihood:                -1020.0
converged:                        True   LL-Null:                       -1924.1
Covariance Type:             nonrobust   LLR p-value:                     0.000
=======================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Favorite of the tweet    -0.4528      0.043    -10.505      0.000      -0.537      -0.368
retweet_count             0.1235      0.011     11.007      0.000       0.102       0.146
Friends Count             0.0002   4.37e-05      5.174      0.000       0.000       0.000
Listed Count             -0.0042      0.001     -4.336      0.000      -0.006      -0.002
Favorites Count        1.024e-05   3.02e-06      3.390      0.001    4.32e-06    1.62e-05
statuses_count         1.707e-05   2.88e-06      5.933      0.000    1.14e-05    2.27e-05
=======================================================================================

Possibly complete quasi-separation: A fraction 0.32 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Figure 6. Logit for Tweets (Multicollinearity)

Another approach was to eliminate multicollinearity analysis and include all numerical variables to see if it yielded different results. Figure 7 reflects the output with 'Favorite of the tweet', 'Time Hour', 'length', 'retweet_count', 'Followers Count', 'listed count', and 'account creation' being statistically significant or, in other words, extremely important to deduce the alliance. Surprisingly, 'Friends Count' and 'statuses_count' weren't shown to be significant in this model. The Log-Likelihood and Pseudo R-Squ. of the model actually came out to be better without addressing multicollinearity.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                  AU1R0   No. Observations:                 3024
Model:                          Logit   Df Residuals:                     3013
Method:                           MLE   Df Model:                           10
Date:                Fri, 26 Aug 2022   Pseudo R-squ.:                  0.7821
Time:                        11:23:33   Log-Likelihood:                -419.31
converged:                       True   LL-Null:                       -1924.1
Covariance Type:            nonrobust   LLR p-value:                     0.000
====================================================================================
                            coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------
length                   -0.0062      0.002     -2.869      0.004      -0.010      -0.002
Time Hour                 0.0947      0.013      7.338      0.000       0.069       0.120
Favorite of the tweet    -0.6103      0.062     -9.885      0.000      -0.731      -0.489
retweet_count             0.1620      0.017      9.806      0.000       0.130       0.194
Followers Count        2.002e-06     1.6e-06     1.250      0.211    -1.14e-06    5.14e-06
Friends Count         -1.104e-05    1.67e-05    -0.661      0.508    -4.38e-05    2.17e-05
Listed Count             -0.0034      0.001     -6.533      0.000      -0.004      -0.002
account creation          0.0009      0.000      7.454      0.000       0.001       0.001
Favorites Count       -1.288e-06    1.69e-06    -0.763      0.445     -4.6e-06    2.02e-06
Account Verified         -1.9443      0.724     -2.685      0.007      -3.364      -0.525
statuses_count         2.593e-06    1.47e-06     1.769      0.077     -2.8e-07    5.47e-06
====================================================================================

Possibly complete quasi-separation: A fraction 0.32 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
```

Figure 7. Logit for Tweets (Without Multicollinearity)

A multilayer perceptron (MLP) is a readily deployable feedforward neural network often used in structured data that doesn't require intensive computing (Popescu et al., 2009). They are part of the neural network, with one or more hidden layers where classification or prediction is conducted on the output layer. An extremely flexible algorithm(Brownlee, 2016), in this case, is used to solve a binary class problem regarding the Ukrainian alliance using independent variables that address multicollinearity as stated in Figure 6.

A standard evaluation metric is deployed for the model after 80% of the data is dedicated to training and 20% to testing, where precision, recall, f1-score, and accuracy are calculated. The model infused with MLP was highly accurate in predicting the '1' or Ukrainian alliance with an average of over 90% in all metrics, while '0' was slightly lower but still respectable. An under sampling of '0' could have hurt the model's learning capabilities.

Table 5. Evaluation Metrics

	Precision	recall	f1-score	support
0	0.89	0.84	0.86	198
1	0.92	0.95	0.94	407
accuracy			0.91	605
macro avg.	0.91	0.89	0.90	605
weighted avg.	0.91	0.91	0.91	605

Figure 8 presents a visualization of the evaluation while the True Negative is 166, False Positive is 32, False Negative is 20, and True Positive is 387 for the given data.

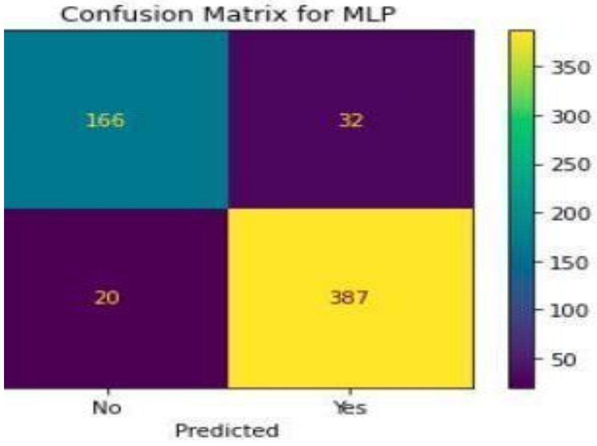
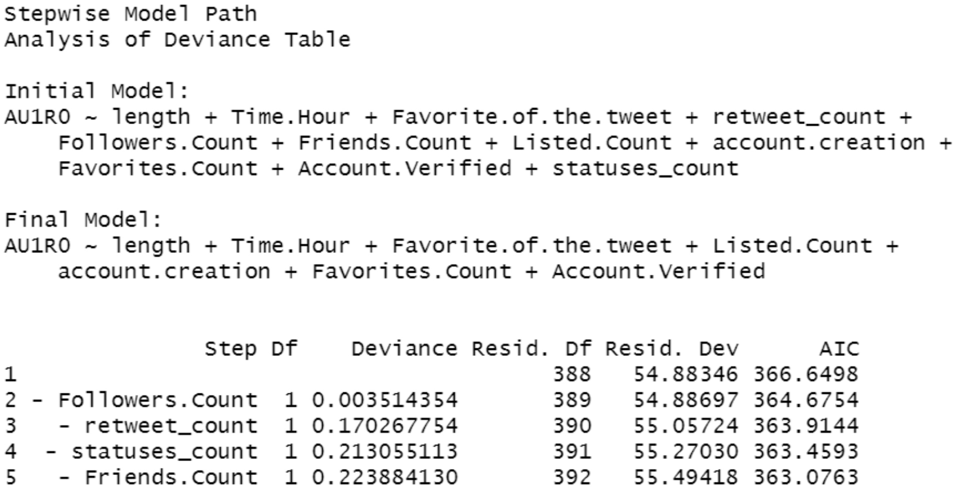


Figure 8. Confusion Matrix

4.3 RQ3 - Can Stepwise Feature Selections and Synthetic Data Improve Model Accuracy?

To improve model performance, selection features play an important part to build the best-performing model. Akaike Information Criterion (AIC) aims to explain variance in dependent variables (AU1R0) with the least number of independent variables under the assumption that a model with fewer parameters equates to a better-performing model. AIC takes into account log-likelihood and explanatory parameters to yield a score or in other words chooses a model that explains the most variance with the least number of parameters; with a model with the lowest AIC score indicating a more accurate model(Brownlee, 2020). Whilst AIC addresses overfitting and reduces computational stress, AIC score on its own is insignificant but the combination of model AIC scores can select the best candidate - a model with the lowest AIC score is presumed the best. Using the R package MASS, a stepwise selection is commissioned where model ‘full’ is created, and using the stepAIC() function eliminates independent variables that increase the AIC score(Zhang, 2016). In Figure 9, the ‘full’ model removes ‘Followers Count’,’retweet_count’,’statuses_count’, and ‘Friends Count’ to decrease the full model AIC score from 366.64 to 363.07.



```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
AU1RO ~ length + Time.Hour + Favorite.of.the.tweet + retweet_count +
    Followers.Count + Friends.Count + Listed.Count + account.creation +
    Favorites.Count + Account.Verified + statuses_count

Final Model:
AU1RO ~ length + Time.Hour + Favorite.of.the.tweet + Listed.Count +
    account.creation + Favorites.Count + Account.Verified


               Step Df    Deviance Resid. Df Resid. Dev      AIC
1                                        388   54.88346 366.6498
2 - Followers.Count  1 0.003514354       389   54.88697 364.6754
3   - retweet_count  1 0.170267754       390   55.05724 363.9144
4  - statuses_count  1 0.213055113       391   55.27030 363.4593
5   - Friends.Count  1 0.223884130       392   55.49418 363.0763
```

Figure 9. Stepwise Selection using R Package MASS

Using the same MLP classification technique and criteria to address RQ2 but with increased sample size through synthetic data injection and selection feature selection (AIC), the model produced improved results, as shown in Table 6.

Table 6. Evaluation Metrics (Synthetic Data + Feature Selection)

	Precision	recall	f1-score	support
0	0.94	0.99	0.96	2008
1	0.98	0.94	0.96	1992
accuracy			0.96	4000
macro avg.	0.96	0.96	0.96	4000
weighted avg.	0.96	0.96	0.96	4000

The newer model predicted an improved accuracy score from 0.91 to 0.96 and an exceptional Recall score of 0.99. The improvement can be attributed to better model fit and equal sampling of both the ‘Russia’ and ‘Ukraine’ alliances. A 10-fold cross-validation of accuracy score established the range from 0.953 to 0.989. However, overtly impressive evaluation metrics generally are cause for skepticism - some culprits could be data leakage (Brownlee, 2016) or incorrect relationships between the variables. In the confusion matrix of the model above, only 158 incorrect predictions were made for both classes, further establishing the efficacy or cynicism of the model.

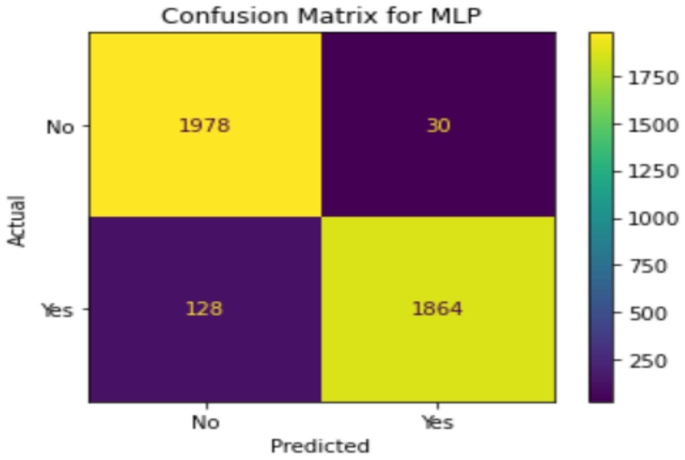


Figure 9. Confusion Matrix

5. LIMITATIONS AND FUTURE SCOPE

The research was unable to find a consequential CTI that was beneficial. Nevertheless, cybersecurity remains an interdisciplinary field, encapsulating all aspects of lives. One of the glaring limitations of the research was the probity of data collection - what makes social media rich in value is its abundance and availability, but at the same time, anyone can post anything, and the dubiety of diverse jurisdictions makes the data collection process unfair. The data was collected randomly in a relatively short period, and while Ukraine-centric data was plentiful, ongoing Twitter restrictions meant Russia-originated/supporting data was rare and often censored. Some trending hashtags at the onset of the conflict were removed, further missing out on the intelligence. The research tracked hashtags to find cybersecurity-centric tweets, but often it was futile in discovering useful content. An implementation of synthetic data is a possibility that could increase sampling and to better understand the model but complexity of dataset should be carefully addressed.

SOCMINT is in its infancy and here to stay; refined methodologies could yield more benefits for the future. Rather than chasing hashtags, a careful pre-selection of influential users that exhibit sincere cybersecurity content should be traced using graph and network theory. An understanding of the relationships from that sub-group will be more prudent in discovering intelligence as opposed to random samplings. Censorship is a huge issue for academics that want to leverage the power of public data - the current Twitter ownership change, in which the new owner will allow leniency towards data freedom to provide unbounded access, is a supposition that could provide research aspirations. Alternatively, creating synthetic data based on artificial intelligence could provide alternate data fuel.

6. CONCLUSION

Whilst the availability of data is tempting, various issues remain to uncover intelligence - questionable integrity and sporadic availability remain key issues. The public data is owned by private companies, and often, their policies dictate intelligence-gathering. A quick stoppage of data flow can derail a framework. When data is available, a refined approach is required to parse the information, as free data is not always the most valuable data. Nevertheless, the functional APIs, low open-source programming, and seamlessness of data discovery make the CTI integration appealing. This research provided sequential steps on building an intelligence-gathering framework whilst incorporating machine learning techniques to make it further discernible. Although it didn't provide substantial value to the Russia-Ukraine conflict largely due to weak data, a better part of the framework is dependable and can be replicated for future experiments.

REFERENCES

APIs for Scholarly Resources | Scholarly Publishing—MIT Libraries. (n.d.). Retrieved April 18, 2022, from https://libraries.mit.edu/scholarly/publishing/apis-for-scholarly-resources/

Blei, D. M. (2003). *Latent Dirichlet Allocation*. 30.

Brownlee, J. (2016, May 16). Crash Course On Multi-Layer Perceptron Neural Networks. *Machine Learning Mastery*. https://machinelearningmastery.com/neural-networks-crash-course/

Brownlee, J. (2019) *Probabilistic Model Selection with AIC, BIC, and MDL*. Available at: https://machinelearningmastery.com/probabilistic-model-selection-measures/ (Accessed: 15 October 2022)

Brownlee, J. (2016) 'Data Leakage in Machine Learning', *Machine Learning Mastery*, 1 August. Available at: https://machinelearningmastery.com/data-leakage-machine-learning/ (Accessed: 15 October 2022).

Code Snippets—Tweepy 3.5.0 documentation. (n.d.). Retrieved April 13, 2022, from https://docs.tweepy.org/en/v3.5.0/code_snippet.html

Costa, C., Aparicio, M., & Aparicio, J. (2021). Sentiment Analysis of Portuguese Political Parties Communication. *The 39th ACM International Conference on Design of Communication*, 63–69. https://doi.org/10.1145/3472714.3473624

Dave, P. (2022). *Ukraine uses facial recognition to identify dead Russian soldiers, minister says | Reuters*. https://www.reuters.com/technology/ukraine-uses-facial-recognition-identify-dead-russian-soldiers-minister-says-2022-03-23/

Dover, R. (2020). SOCMINT: A shifting balance of opportunity. *Intelligence and National Security, 35*(2), 216–232. https://doi.org/10.1080/02684527.2019.1694132

Forrester, B., & Hollander, K. den. (2016). *The role of Social Media in the Intelligence Cycle*.

Getting started—Tweepy 4.2.0 documentation. (n.d.). Retrieved November 1, 2021, from https://docs.tweepy.org/en/stable/getting_started.html#models

Husari, G., Al-Shaer, E., Chu, B., & Rahman, R. F. (2019). Learning APT chains from cyber threat intelligence. *Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security - HotSoS '19*, 1–2. https://doi.org/10.1145/3314058.3317728

Ivan, A. L., Iov, C. A., Lutai, R. C., & Grad, M. N. (2015). Social Media Intelligence: Opportunities and Limitations. *Social Media Intelligence*, 7.

Kakar, S., Dhaka, D., & Mehrotra, M. (2021). Value-Based Retweet Prediction on Twitter. *Informatica*, *45*(2). https://doi.org/10.31449/inf.v45i2.3465

Kelion, L. (2014). *Snowden leaks: GCHQ 'spied on Facebook and YouTube'—BBC News*. https://www.bbc.com/news/technology-25927844

Khan, I. (2022). *Zelenskyy Humanizes Ukraine's Plight in His Social Media Messaging—CNET*. https://www.cnet.com/news/politics/zelenskyy-humanizes-ukraines-plight-in-his-social-media-messaging/

Kobie, N. (2019). *The complicated truth about China's social credit system*. The Complicated Truth about China's Social Credit System. https://www.wired.co.uk/article/china-social-credit-system-explained

Kropotov, V., & Yarochkin, F. (2019). *Basic Social Media Intelligence (SOCMINT) Tools To Help Fight Disinformation*. http://www.mikekujawski.ca/2019/02/25/basic-social-media-intelligence-socmint-tools-to-help-fight-disinformation/

Lakshmi, J. V. N. (2018). Machine learning techniques using python for data analysis in performance evaluation. *International Journal of Intelligent Systems Technologies and Applications*, *17*(1/2), 3. https://doi.org/10.1504/IJISTA.2018.10012853

Laucuka, A. (2018). Communicative Functions of Hashtags. *Economics and Culture*, *15*(1), 56–62. https://doi.org/10.2478/jec-2018-0006

Mahajan, R., Mahajan, R., Sharma, E., & Mansotra, V. (2022). "Are we tweeting our real selves?" personality prediction of Indian Twitter users using deep learning ensemble model. *Computers in Human Behavior*, *128*, 107101. https://doi.org/10.1016/j.chb.2021.107101

Mahood, Lc. M. (2015). *SOCMINT: Following and Liking Social Media Intelligence*. 25.

McLaughlin, J. (2022). *Social media volunteers aim to help Ukraine win the information war: NPR*. https://www.npr.org/2022/03/17/1087137578/social-media-volunteers-aim-to-help-ukraine-win-the-information-war

McLoughlin, L., Ward, S., & Lomas, D. W. B. (2020). 'Hello, world': GCHQ, Twitter and social media engagement. *Intelligence and National Security*, *35*(2), 233–251. https://doi.org/10.1080/02684527.2020.1713434

Milmo, D. (2022). *Russia blocks access to Facebook and Twitter | Russia | The Guardian*. https://www.theguardian.com/world/2022/mar/04/russia-completely-blocks-access-to-facebook-and-twitter

Momi, R. (2021). *SOCMINT: Social Media Intelligence a New Discipline? - Grey Dynamics*. https://www.greydynamics.com/socmint-social-media-intelligence-a-new-discipline/

Omand, D., Bartlett, J., & Miller, C. (2012). Introducing Social Media Intelligence (SOCMINT). *Intelligence and National Security*, *27*(6), 801–823. https://doi.org/10.1080/02684527.2012.716965

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). *Multilayer Perceptron and Neural Networks*. *8*(7), 11.

Sauerwein, C., Fischer, D., Rubsamen, M., Rosenberger, G., Stelzer, D., & Breu, R. (2021). From Threat Data to Actionable Intelligence: An Exploratory Analysis of the Intelligence Cycle Implementation in Cyber Threat Intelligence Sharing Platforms. *The 16th International Conference on Availability, Reliability and Security*, 1–9. https://doi.org/10.1145/3465481.3470048

Social Media Intelligence. (n.d.). Privacy International. Retrieved April 12, 2022, from http://privacyinternational.org/explainer/55/social-media-intelligence

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, *24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Stukal, D., Sanovich, S., Tucker, J. A., & Bonneau, R. (2019). For Whom the Bot Tolls: A Neural Networks Approach to Measuring Political Orientation of Twitter Bots in Russia. *SAGE Open*, *9*(2), 215824401982771. https://doi.org/10.1177/2158244019827715

Taboga, M. (n.d.). *Logistic classification model (logit or logistic regression)*. Retrieved May 4, 2022, from https://www.statlect.com/fundamentals-of-statistics/logistic-classification-model

Three surveillance technologies that protesters need to know about—IFEX. (2019). https://ifex.org/three-surveillance-technologies-that-protesters-need-to-know-about/

Toews, R. (2022) *Synthetic Data Is About To Transform Artificial Intelligence*, *Forbes*. Available at: https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/ (Accessed: 13 October 2022).

Trouillard, S. (2022). *Following the Ukraine war – and fighting it – on social media*. https://www.france24.com/en/europe/20220308-following-the-war-in-ukraine-%E2%80%93-and-fighting-it-%E2%80%93-on-social-media

Tweet object | Docs | Twitter Developer Platform. (n.d.). Retrieved April 19, 2022, from https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

Ukraine's Digital Ministry Is a Formidable War Machine | WIRED. (2022). https://www.wired.com/story/ukraine-digital-ministry-war/

Using Social Media (SOCMINT) in Threat Hunting. (n.d.). Retrieved April 18, 2022, from https://www.anomali.com/blog/using-social-media-socmint-in-threat-hunting

Vail, E. (2022). *Russia or Ukraine: Hacking groups take sides—The Record by Recorded Future*. https://therecord.media/russia-or-ukraine-hacking-groups-take-sides/

Wang, H., Li, Y., Hutch, M., Naidech, A., & Luo, Y. (2021). Using Tweets to Understand How COVID-19–Related Health Beliefs Are Affected in the Age of Social Media: Twitter Data Analysis Study. *Journal of Medical Internet Research*, *23*(2), e26302. https://doi.org/10.2196/26302

Zhang, Z. (2016) 'Variable selection with stepwise and best subset approaches', *Annals of Translational Medicine*, 4(7), pp. 136–136. Available at: https://doi.org/10.21037/atm.2016.03.35.