Using Open-Source Intelligence and Machine Learning to Analyze Cyber-Attack Trends in the African Cyberspace

Jabu Mtsweni
Council for Scientific and Industrial Research and Stellenbosch University, South Africa

Abstract

Cyber-attacks continue to evolve, circumventing modern security solutions and exploiting discovered software vulnerabilities. Businesses and governments tend to respond in a reactive manner when dealing with these attacks, mostly due to limited situational awareness, lack of threat intelligence sharing, and lack of timely and actionable threat intelligence. The African cyberspace is vulnerable of cyber threats and risks that require timely monitoring to enable decision makers to understand the cybersecurity threat landscape on continuous basis. This paper, therefore, aims to use open-source intelligence from social media and machine learning techniques to analyze the emerging cyber-attack trends in Africa and demonstrate the value of such information for situational awareness that could be used by decision makers to mitigate cyber threats and risks. The paper contributes strategic and technical guidelines for exploiting OSINT to timely respond to cyber-attacks.

1. Introduction

The African cyberspace is seen as a haven for cyber-crime and cyber-attacks due to the lack of cybersecurity capabilities (people, processes, and technologies), lack of reporting and awareness. Based on multiple business and research reports, cyber-attacks have been increasing exponential in Africa [1] and there seem to be no end-in-sight with over 52% of companies in Africa said to be unprepared for a large-scale cyber-attack, according to the Dell Technology [2]. In Kenya, it is reported that cyber-attacks have doubled from 2021-2022 [1]. In South Africa, it is estimated that cyber-attacks are costing the economy over R2 billion every year [3]. The report by Ugandan authorities indicates the cyber-crime financial impact of over UGX 19 billion in 2022 [4]. On average, it is reported that organizations take over 200+ days to discover cyber-attacks in their environment [2], and in most cases informed by hackers or other third parties.

The Dell Technologies report released in 2023 [2] claims that cyber-attacks in Africa have increased by 600% over the last two years (2020-2022) due to cyber unpreparedness of organizations. The report further indicates that most companies (61%) are impacted by ransomware attacks leading to an estimated financial loss of over $4 billion every year, with most data breaches caused by human error. These insights are supported by the Verizon Data Breach report of 2023 indicating that over 83% of data breaches are financially motivated [5].

The availability and sharing of cyber threat intelligence in the African cyberspace is limited, if not non-existent. This puts many organizations at risk as they would get to know about the cyber-attacks or threats when it has impacted them, which can be very costly.

Open-source intelligence (OSINT) is gaining momentum in different domains to address modern challenges such as cyber-crime and cyber-attacks. However, in Africa it is still underutilized for dealing with cybersecurity challenges.

This paper demonstrates how open-source threat intelligence from social media could be used to understand and determine the cyber threat landscape in the African cyberspace on continuous basis for situational awareness purposes. This could in turn aid both the public and private sector organizations in proactively mitigating cyber threats and risks, but also sharing threat intelligence with their stakeholders to minimize the extent of cyber breaches.

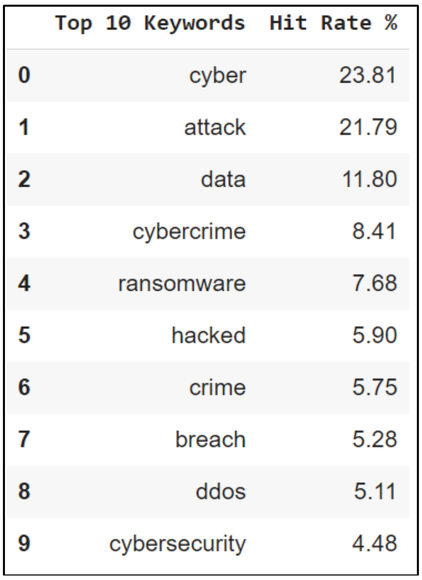
The rest of this paper is structured as follows: Section 2 summarizes the research approach taken to conduct the study including data collection, analysis, and results reporting. Section 3 provides a cybersecurity literature review overview, and Section 4 provides an exploratory analysis of the cyber-attack social media data using statistical techniques and sentiment analysis.

In Section 5, we adopt natural language processing approaches to identify key topics from the cyber-attack data points and present these as prevalent cyber-attack trends in Africa. Section 6 discusses the research results and implications by recommending strategic and technical guidelines for situational awareness using OSINT. Section 7 concludes the paper with a summary and future research considerations.

2. Research approach

The main objective of the research presented in this paper was to determine and understand the threat landscape and trends in the African cyberspace using open-source intelligence from social media.

Table 1. Data collection based on selected keywords.



	Top 10 Keywords	Hit Rate %
0	cyber	23.81
1	attack	21.79
2	data	11.80
3	cybercrime	8.41
4	ransomware	7.68
5	hacked	5.90
6	crime	5.75
7	breach	5.28
8	ddos	5.11
9	cybersecurity	4.48

The data was collected over a 3-month period (May 2023 – July 2023) on social media using commonly known cyber security related words which returned the hit-rates as demonstrated in Table 1. It should be noted that some of the keywords appeared in combination such as cyber-attack and DDoS attack.

The research approach adopted in this paper was both quantitative and qualitative analysis as discussed in the subsections that follow. Machine learning techniques in topic modelling and clustering were adopted to determine the common topics in the OSINT data.

2.1. Open-source intelligence data collection

Open-Source Intelligence (OSINT) deals with the collection and/or processing of publicly available information and data from different sources for data intelligence purposes [6] [7]. These sources include the Internet and other social media platforms such as Facebook, Twitter, and websites [8].

The authors of this paper collected data for this research using a web platform called “Talkwalker” by setting up the search criteria using the keywords depicted in Table 1. The data was exported to a “csv” file for exploration and deeper analysis.

Social media, as a source of data collection, is used often in research [9] [10] [11] [12] and it was found to be useful and relevant to this study because social media and news websites report on daily basis and information is generally received without much delay.

Twitter was the main source for data collection including the top news websites that were referenced in the tweets. To ensure that data related to African countries was considered, only data geo-tagged to African countries were collected and data in “English” for purposes of the analysis. It should, however, be noted, that in some instances the tweet could be geo-located to be in Africa, but the content may be referring to events happening outside the African continent, and thus why deeper analysis of the data including triangulation of coordinates was important in this study.

The data points collected included (amongst other data items) date, content, authors details, post type (e.g., text, videos, and images), engagements (e.g., likes, comments, and retweets), followers, region, country, and city).

2.2. Trend, sentiment, and topic analysis

The data was analyzed using different tools and techniques focusing on descriptive statistics, trend analysis, sentiment analysis and topic modelling. Trend analysis refers to the use of historical data to identify patterns and trends related to a subject of interest. In this paper, we chose a time-series as an approach to investigate the cybersecurity trends in Africa within the data collection period.

Sentiment analysis, which refers to the identification and quantification of emotions, moods, and context of a data point, is quite common in natural language processing (NLP) and computational linguistics [10]. On social media, sentiments are quite important to understand what is being shared as well as its implications. In this study, we employ latent sentiment analysis (LSA) to measure sentiments of the data points from social media using different textual and emotional signals [13]. This is done to generally identify cyber attack impact in the African cyberspace.

Lastly, topic modelling is a topic analysis technique in machine learning that is used to analyze latent topics and themes in textual documents based on the frequency of words [14], and these techniques are commonly used in text mining. In this study, we found topic modelling to be relevant to enable us to broadly understand the cyber-attack trends in Africa.

2.2. Data analysis tools and techniques

Several tools and techniques were used to analyze the data in this study. Excel and Google Colaboratory Platform were employed as platforms for analysis. The python programming language was used in Google Colab, and the Data Analysis Package was used in Excel. For latent sentiment analysis (LSA), we used the VADER [15] sentiment analysis framework, specifically catered for social media data. The comprehensive analysis is provided in Sections 4 and 5. To perform topic modelling, we applied the Latent Dirichlet Allocation (LDA) technique [14] in Google Colab. This is explained further in Section 5.

The next section discusses the related research literature pertaining to the cyber threat landscape in Africa.

3. Related research

Cybersecurity, cybercrimes, and cyber-attacks are observed as being the top trending research topics between 2011 to 2021 [16]. In the same context, open-source intelligence research in the cybersecurity domain has been steadily emerging over the recent years, increasingly used by governments and businesses for strategic planning and implementation.

According to the research work in [17], OSINT resources are leveraged to shadow threat actors using public information to generate intelligence pertinent to power systems cyber threats in the US, which involved constructing power system model, validating the model, and processing the model to identify critical locations of power systems.

In their research, Hayes and Cappa [18] demonstrated the importance of using OSINT in conducting risk assessments to prevent cyber-attacks. In their research, they conducted vulnerability assessments on critical infrastructure for a private company in the United States using open-source intelligence, and their results showed that OSINT can be very useful in providing and improving a security posture of an entity. The main contribution from their work are OSINT usage recommendations to identify and mitigate cyber threats. The similarities between this work and ours is the use of OSINT, however the focus of our study was on understanding cyber-attack trends to proactively improve one’s cybersecurity program.

In [19], social media analytics are used to understand the trend of Distributed Denial of Service (DDoS) cyber-attacks from a sentiment and motivation perspectives. In their work [19], they argue that cyber-attacks are not only technical events, but that they need to be associated with social, political, economic, and cultural motives to effectively prevent them. Social media data is therefore seen as an arsenal for insights in understanding and preventing cyber-attacks based on sentiment analysis.

In their study [20], social media analysis is used to identify Common Vulnerabilities and Exposures (CVE). In this study, Twitter was found to be the most common platform for engaging on cybersecurity vulnerabilities. They conclude that security vulnerabilities could be identified with high probability by tracking open-source intelligence on social networks such as Twitter, GitHub, and Reddit.

As may be noted from the related research, there are no related studies in Africa that attempts to utilize social media OSINT to understand cyber-attacks in the continent. This study contributes to closing this research gap using various techniques.

4. Cyber-attack trends in the African cyberspace

The cybersecurity trends in Africa are changing as the continent accelerates its digital transformation. Nevertheless, the cybersecurity reports in Africa are still largely produced by international companies based on their specific footprint. In this section, we attempt to show another view on cyber-attack trends in Africa using open-source intelligence.

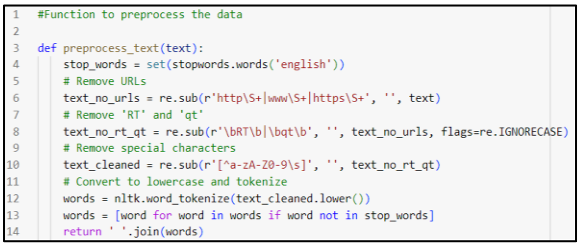
4.1 Data analysis overview

During the data collection phase, 16,384 cyber-attack related data points in “English” were collected contributed by over 10,296 unique authors with the total engagement score of over 25,372. Only 6% of the total data points had an engagement score of 1 or more. This suggests a low engagement and awareness on cybersecurity issues in the African cyberspace on social media. The majority (82%) of the tweets were also retweets, suggesting low reporting of cyber-attacks on social media.

Furthermore, most of the data points indicated negative sentiments and clustering also showed a positive correlation between engagements and followers. The influencers of cyber-attack reports in Africa appear to be entrepreneurs, authors, media personalities. However, our analysis of the data indicates that there is a limited number of cybersecurity experts (based on the description of authors) reporting about cyber-attack incidents, which may also be the reason why there is such low engagement around this domain.

4.2. Data preprocessing

Data cleaning is an important part of data analysis. The collected data, saved in a “csv” file was cleaned via a python script (see Figure 1) exploiting stop-words to remove all common words, and language detection libraries to remove all words that were not in “English”. We also used regular expressions to remove words such as “RT” (for retweets), links, punctuations, and name of twitter handles, and others. In addition, empty columns were dropped. Digits were not removed since they play an important role in understanding the impact of cyber attacks such as number of data exposed by a breach.



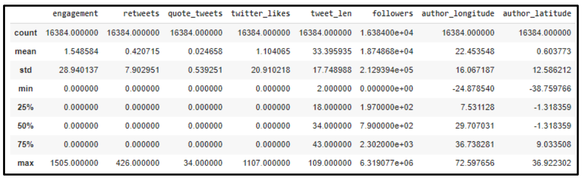
```python
#Function to preprocess the data

def preprocess_text(text):
    stop_words = set(stopwords.words('english'))
    # Remove URLs
    text_no_urls = re.sub(r'http\S+|www\S+|https\S+', '', text)
    # Remove 'RT' and 'qt'
    text_no_rt_qt = re.sub(r'\bRT\b|\bqt\b', '', text_no_urls, flags=re.IGNORECASE)
    # Remove special characters
    text_cleaned = re.sub(r'[^a-zA-Z0-9\s]', '', text_no_rt_qt)
    # Convert to lowercase and tokenize
    words = nltk.word_tokenize(text_cleaned.lower())
    words = [word for word in words if word not in stop_words]
    return ' '.join(words)
```

Figure 1: Text preprocessing

4.3 Exploratory analysis

Data exploration is quite useful in understanding the data and detecting any issues before comprehensive analysis. In this research, we performed descriptive statistics and text analysis on the data to have a view on what could be the keywords that appear the most and what the stats may be saying about the data after cleaning.

4.3.1. Descriptive statistics. Figure 2 shows a summary of the descriptive statistics of the data used in this research. The stats provide the count of the total records, mean, standard deviation (std), min and max.



	engagement	retweets	quote_tweets	twitter_likes	tweet_len	followers	author_longitude	author_latitude
count	16384.000000	16384.000000	16384.000000	16384.000000	16384.000000	1.638400e+04	16384.000000	16384.000000
mean	1.548584	0.420715	0.024658	1.104065	33.395935	1.874868e+04	22.453548	0.603773
std	28.940137	7.902951	0.539251	20.910218	17.748988	2.129394e+05	16.067187	12.586212
min	0.000000	0.000000	0.000000	0.000000	2.000000	0.000000e+00	-24.878540	-38.759766
25%	0.000000	0.000000	0.000000	0.000000	18.000000	1.970000e+02	7.531128	-1.318359
50%	0.000000	0.000000	0.000000	0.000000	34.000000	7.900000e+02	29.707031	-1.318359
75%	0.000000	0.000000	0.000000	0.000000	43.000000	2.302000e+03	36.738281	9.033508
max	1505.000000	426.000000	34.000000	1107.000000	109.000000	6.319077e+06	72.597656	36.922302

Figure 2:Descriptive statistics

The descriptive statistics give us some insights into the usefulness of the data, and from the above, that count of records is 16,384. The standard deviation on engagement, which is far from the mean, suggests diversity in relation to cyber-attack content engagement in Africa. A further exploration indicates that there are data points with zero engagements, which impacts on the engagement scores.

4.3.2. The occurrence of cyber-attacks words. Based on the keywords in Figure 3, it is evident that cyber-attack dominates the dataset. Moreover, other common aspects such as data breach, cyber-crime, DDoS attack, and ransomware attack seem to be prevalent in cyber engagements in Africa, however this is expected, since these are the keywords used for data collection.

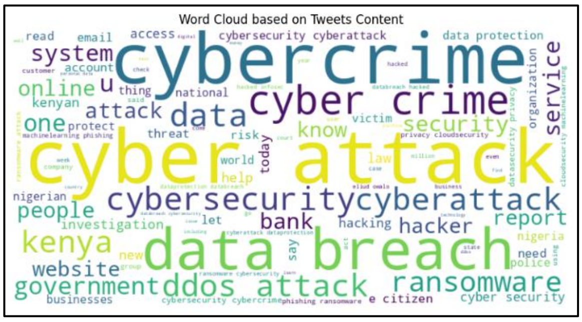


Figure 3: World cloud based on Tweets [own source]

4.3.3. Distribution of data per country in Africa. The data analyzed suggests that most (40%) of the cybersecurity posts were related to Kenya (see: Figure 4), followed by Nigeria and South Africa. These stats seem to contradict other cybersecurity reports such as those by Interpol [3] that indicate that many of the cybersecurity incidents are detected in South Africa.

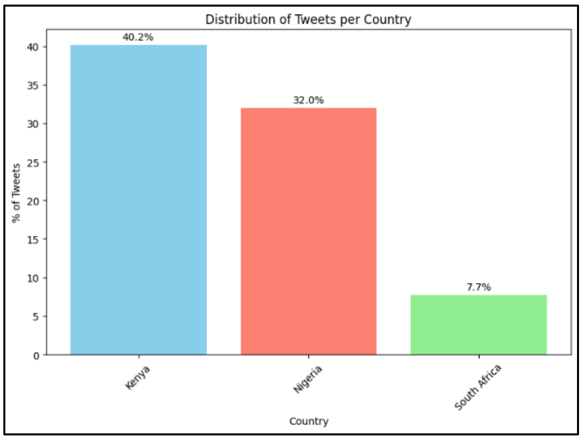


Figure 4: Distribution of content per country

4.3.4. Sentiment analysis. The content of the tweets was analyzed using the VADER sentiment analysis technique [15]. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis technique that is specifically attuned to sentiments expressed in social media. It is a rule-based sentiment analyzer in which the terms are generally labeled as per their semantic orientation as either positive or negative. It has been shown to outperform other sentiment analysis techniques on a variety of datasets, including social media data, and thus was chosen for this research.

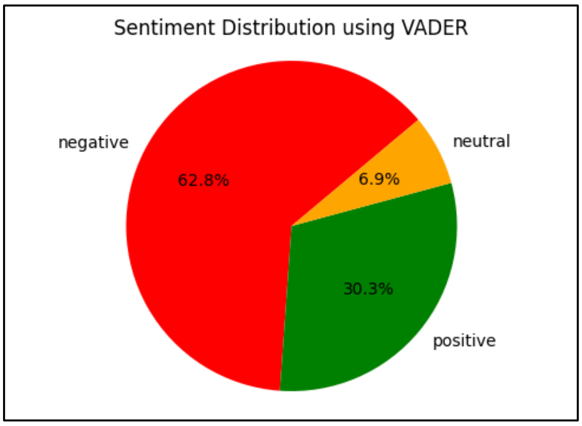


Figure 5: Overall Sentiment Analysis

Our analysis on the collected data indicates that most of the cyber-attack related posts have a strong negative sentiment (62.8%) as depicted in Figure 5. This sentiment is not a surprise considering that cyber-attacks have a negative impact on organizations and users, whether because of service delivery disruptions, financial implication, and/or reputational damage that happen due to data breaches and/or ransomware attacks.

4.3.5. Data source type. The cybersecurity Twitter content is by all accounts generated by mobile devices (95%), and the rest generated using external websites and desktop computers. The automated bots accounted for less than 1% of the content generated. In terms of gender, 65% of the data points were generated by males with female contributing 18%. It should be noted that not all social media data points are labeled in terms of gender and location, so in some cases these stats may not always be 100% accurate, but nevertheless they do give an indication of some reality. According to [21] most of the Twitter registered users are males, and as such this analysis is not far off.

5. Emerging cyber threat landscape in Africa

In this section, we demonstrate the cyber-attack trends in the African cyberspace using the social media data as discussed above. We demonstrate this in three ways (1) trend analysis using a time-series, (2) topic modelling, and (3) clustering and correlation.

5.1. Trend analysis

We grouped the posts' date and associated engagement to determine the trend and peak points as shown in Figure 6. The social media engagements are mostly below a mean of 2 and spikes going to as high as a mean of 6.

The cyber-attack cyclic pattern in the African cyberspace indicate that cyber-events are influenced by news reports, political debates on cyber laws, data breaches, cyber-crime arrests, or investigations by authorities.

Examples: on 02-June-23, the peak point was influenced by the reported data breach that exposed half a million customers of Incredible Connection, HiFi Corporation, and other retailers in South Africa. On 27-July-23 to 29-July-23, the peak points were influenced by the government of Kenya admitting to a cyber-attack that brought its e-Citizen platform down for couple of days leading to interruption of public services such as issuing of visas and other online services. This was accompanied by other online reports of the Kenyan Bureau of Standards (KEBS) suffering a massive data breach with over 739GB data of KEBS published on dark web.

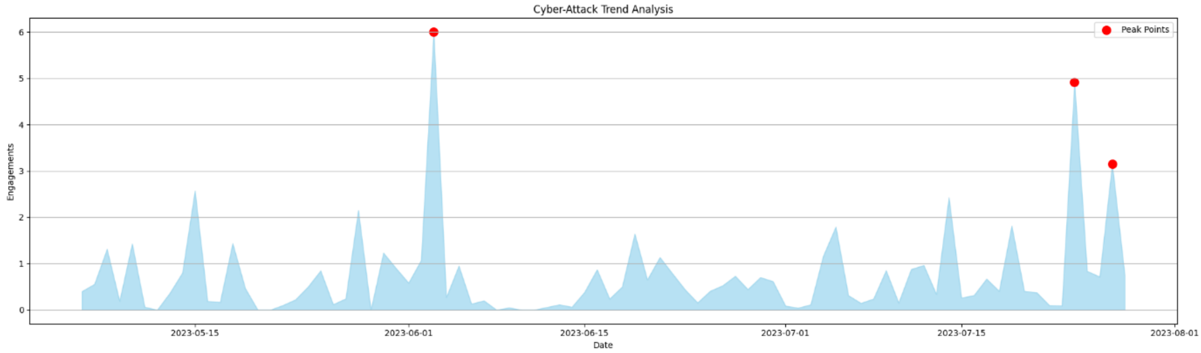


Figure 6: Cyber-attacks timeline trend in Africa

5.2. Topic modelling

The data collected for this research was quite rich. However, the number of issues covered by the data were also quite diverse as highlighted in the trend analysis. We focused on the content with an engagement score of 10 or more and z-score with 3 standard deviations from mean. This was done to take care of outliers that normally appear in peak points where retweets may dilute the topics identified.

For topic modelling, we employed unsupervised machine learning approaches in the Latent Dirichlet

Allocation (LDA) and the Non-Negative Matrix Factorization (NMF). The reason we chose the two was for topic validation and comparison of the topics across a large corpus of cyber-attacks.

LDA [14], is a commonly used machine learning technique for topic modeling. It is designed to discover hidden topics within a collection of text documents. LDA assumes that each document is a mixture of various topics, and that each topic is a mixture of words. In essence, the goal of LDA is to automatically identify these topics and their associated word distributions. On the other hand, NMF [22] can be described as topic modelling technique that decomposes a given matrix into two non-negative matrices. This is useful for dimensionality reduction, feature extraction, and interpretability.

We applied these techniques scaled at 10 topics with a range of two words. The maximum document frequency used for inclusion of words in a topic was 85% and the minimum document frequency was 2. For both techniques, we limited features extraction to 1000.

We also applied the Term Frequency-Inverse Document Frequency (TF-IDF) scores [23] on LDA to extract dominant topics. This was to understand their relevance in the context of cyber-attacks. The topics that emerged from the analysis using the two techniques indicate the realistic cyber-attack trends in Africa.

5.2.1. LDA generated topics. The Latent Dirichlet Allocation (LDA) technique was applied to preprocessed tweets to uncover themes and topics related to cyber-attacks in Africa. The cleaned content was converted into a list and thereafter vectorized (i.e., converted from text to numerical data) using the *TfidfVectorizer* plugin in Python for feature extraction and topic analysis. The cyber-attack related topics that were discovered using the LDA technique are shown below.

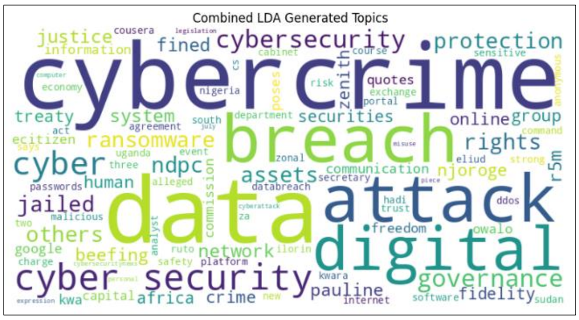


Figure 7: LDA cyber-attack generated topics [own source]

Based on the results depicted in Figure 7, the following sampled cyber-attack trends are observed in Africa, which are correlating with the trends reported in security vendors' reports, related research, and regular news reports.

- **Cybercrime**: this topic is at the center stage in Africa from the legal, attack, and awareness perspectives. The trend suggests that governments in Africa are ramping up efforts to create awareness about different types of cyber-crime. An example of this is the campaign by the Economic and Financial Crimes Commission (EFCC) in Nigeria appealing to students not to be involved in cyber-crime. The topic of cyber-crime also gained attention in Kenya in July 2023 when "*Pauline Njoronge*", a digital blogger and political commentator was arrested and social media users speculating that she is being charged for cyber-crimes and cyber harassment. Other trends in this topic include the development of the cyber-crimes' treaty and laws in different countries including the United Nations cyber-crimes convention in which several African countries are also participating. The convention seeks to balance criminalizing the use of ICTs to commit digital offenses vis-à-vis promoting human rights, including freedom of speech.
- **Data breach**: this topic transcends across most African countries impacting on government and businesses digital platforms. In a case in Kenya involving Equity Bank, the court ruled to allow members of the public to join a data breach lawsuit against the bank. During the same period, a data breach affecting over half-a-million users in South Africa was reported. During the same period, the Department of Justice in South Africa was fined ZAR5million for failing to put security measures in place after a data breach. In Nigeria, the National Data Protection Commission (NDPC) revealed that it was investigating several institutions including big banks such as Zenith and Guarantee Trust Bank for contravening the Nigerian Data Protection Act (NDPA), which carries penalties up to 2% of the gross revenue. The Uganda Securities Exchange is also facing prosecution for unauthorized disclosure of personal data.
- **Data governance**: Due to the prevalent nature of data breaches, several countries in Africa already have data protection laws, and in some cases already being implemented as we have observed in South Africa, Nigeria, Kenya, and Ghana. This is a positive step that provides digital users in Africa with remedies for invasion of the personal information.

- **Ransomware attacks**: The topic analysis reveal that ransomware attacks in Africa are prevalent. During the period of this study, Kenya Bureau of Standard (KEBS) suffered a massive data breach that exposed over 700GB of data on dark web. Moreover, a suspected ransomware attack on Kenya Power was reported to having been carried by the Blackcat ransomware group in 2021. In addition, it was also reported that Kabarak University was also attacked in 2022 by the same group where university systems were breached, and student information stolen.
- **Digital services disruption:** This topic is influenced by the disruption of the Kenyan eCitizen platform that was shut down for few days due to a Distributed Denial of Service (DDoS) attack that came from the Anonymous Hackers in Sudan. At least more than 5,000 public digital services were affected [24]. This event also demonstrates how nation-state align attacks are also emerging in Africa.
- **Cybersecurity awareness:** On the positive sentiment, the trends in Africa also indicate that organizations are taking an effort in creating cybersecurity awareness, with topics such as data protection, password hygiene, online security courses, as well as research by universities featuring.

From the topic analysis, a trend is also observed that governments are most impacted by cyber-attacks in Africa. However, the private sector is not left out, as they seem to be playing a role in causing data breaches.

5.2.2. NMF generated topics. The Non-Negative Matrix Factorization (NMF) technique was also applied, using same stages as the LDA, to the social media corpus. The only difference in this technique was that we applied that the *CountVectorizer*, which has a simpler application in that it converts a collection of text documents into a matrix of token counts without considering their importance. NMF was preferred for topic modelling in addition to the LDA technique because of its benefits to uncover underlying thematic structures in the social media text, and due to its strength of non-negative constraint, it was found to provide topics that are more interpretable than with LDA.

The topics that were discovered using the NMF technique are shown below.

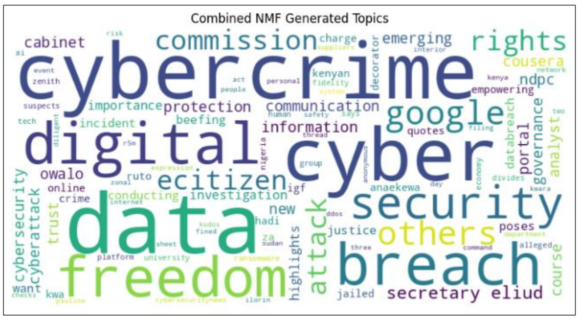


Figure 8: NMF cyber-attack generated topics [own source]

As many be observed from Figure 8, the topics generated through NMF are not far from LDA, and as such the topics' explanation will be the same.

5.3. Clustering and correlation

In understanding the cyber-attack trends in Africa, we further performed k-means clustering on the social media data to cluster posts based on engagement and followers. This was done to determine the relationship between these two indicators. As may be observed in Figure 9 below, cluster 0 commands most data points, which we can categorize as a "catch-all" cluster. This suggests that the cyber-attack posts within this cluster are varied.

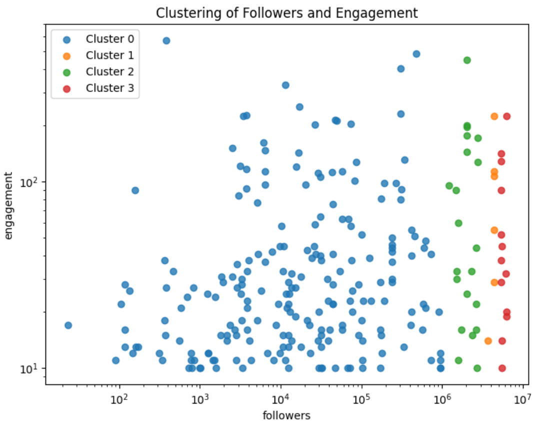


Figure 9: K-means clustering of engagement vs followers in cyber-attack data

Cluster 1 showed a positive correlation between followers and engagement. This means that as users with more followers report cyber-attack incidents, the engagement and spread of such information increased on social media. Cluster 2 showed no correlation between engagements and followers in relation to cyber-attack posts, and this was also the case with Cluster 3.

The clusters provided insights into how Africans on social media are engaging with cyber-attack reports. It is our conclusion that the engagement is quite low (less than 1%) even in instances where the reporter of the cyber-attack news has a large following. This is also an indication of the maturing levels of the African cyberspace that still requires growth and awareness in terms of cybersecurity.

6. Research discussions and implications

During the research period of this paper, it was evident that Africa is experiencing cyber-attacks impacting individuals, governments, and the private sector, although there was limited data from Twitter from other countries in Africa. The results of the study confirmed that Kenya, Nigeria, and South Africa have the highest social media posts that are related to cyber-attacks. These seem to be correlated with the reports in the media and other research studies. For instance, in early 2023, a report by Business Today [25] headlined that Kenya recorded over 278 million cyber-attacks in 3 months. The same report indicated that Kenya reported cyber threats had increased by 200% in one quarter. In South Africa, the Check Point Research estimates 1,450 attacks per week [1].

It is therefore apparent, based on the analysis provided in this paper and from research reports, that open-source intelligence could provide timely and relevant cyber-trends in the African cyberspace. It is, therefore, our recommendation that decision-makers consider the following strategic, tactical, and technical guidelines (see: Table 2) that may enable them to have a greater situational awareness of cyber issues within their operating environments.

Table 2: Guidelines for tracking cyber-attacks using open-source intelligence.

Guideline	Description
1. Monitor cybersecurity discussions	Cybersecurity leaders need to follow cybersecurity trends by experts, news reporters, including, researchers, and organizations on social media platforms. These often share insights, analysis, and real-time updates about cyber threats and attacks in Africa. Twitter, LinkedIn, Facebook, and specialized cybersecurity forums could be particularly useful for this purpose.
2. Track threat actors and groups	Cyber threat actors and hacking groups maintain a presence on social media to publicize their activities, promote their campaign, sell their malicious tools, or claim responsibility for cyber-attacks. Monitoring their profiles and interactions can provide clues about their capabilities, targets, and potential motives. This observation in Africa is seen with the Anonymous Hackers and Black Cat Ransomware Hacking Group.
3. Analyze security incident reporting	This study showed that governments, news reporters, organizations and individuals affected by cyber incidents are increasingly sharing their experiences on social media. By analyzing these reports, organizations can gain a wider understanding of cyber-attacks that are prevalent in the African cyberspace, including zooming into the industries targeted, and the tactics used by the attackers.
4. Follow cybersecurity news	In this research paper, it has been demonstrated that cybersecurity news websites have social media accounts to share the latest cybersecurity-related news and updates. By following these sources, organizations both in the public and private sector may stay informed about recent cyber incidents in Africa and in their sectors.
5. Participate in threat intelligence sharing platforms	Social media platforms, such as Twitter have a large list of relevant threat intelligence sharing groups. These groups allow cybersecurity professionals to share information, indicators of compromise (IOCs), and insights on emerging threats in real-time. And these could act as a source of threat intelligence for organizations, but also could be quite useful for open-source intelligence that may be used even for threat hunting.

While cyber-attacks OSINT could provide valuable intelligence, it is also important for organizations to consider the potential limitations and challenges of such public data. Information on social media may not always be accurate, and it could be subject to manipulation or misinformation campaigns. Thus, it is also critical for organizations to always verify the authenticity of the data before using it as a basis for cybersecurity decisions.

6.1. Research and data limitations

This research was only limited to the data collected from social media - Twitter for a 3-month period. The data collected was varied and did not only focus on direct cyber-attacks. Thus, the trend presented in this research deals with people, processes, and technology aspects of cyber-attacks. This may be a limitation as it may not give a comprehensive understanding of all cyber-attacks in the African cyberspace. In this research, we also applied various machine learning techniques tuned with different parameters, and as such, the results may differ slightly if different techniques and settings are used by other researchers.

7. Conclusion

Cyber-attacks are a reality in Africa and appear to be a regular occurrence. Over a 3-month period, we collected over 16,000 data points that dealt with cyber-attacks engagements in Africa. Over 6% of the data points had an engagement score of 1 or more, and majority had zero engagements indicating a low-cyber awareness in Africa. The cyber-attack trends in Africa show a cyclic pattern with sporadic peak points when cyber-attack issues are reported in the news or become part of legal and/or political discussions.

The collected data, defined as open-source intelligence, was analyzed using different statistical and machine learning techniques that allowed us to understand the data from different angles. The analysis revealed that Kenya is a country that demonstrates a high rate of cyber-attacks in Africa, with the government sector mostly impacted, based on public reports. In South Africa, data breaches seem to be quite high, albeit they are reported late. In Nigeria, cybersecurity engagements are quite high, but limited cyber-attacks are reported, except at an individual level where social media users report being attacked or compromised.

This research paper contributes to the field by demonstrating the cyber-attack trends in Africa using open-source intelligence and machine learning techniques. In addition, we provide cyber-attack themes that appear to be common across different Africa countries. The paper also concludes that the lack of reporting, data, and engagement in other countries on cyber-attack issues does not necessarily mean that those countries are safe or are not experiencing cyber-attacks. Further studies are recommended to track cyber-attack trends in other countries other than South Africa, Kenya, and Nigeria.

8. References

[1] Check Point. (2023). Check Point Software's 2023 Cyber Security Report. Check Point, United States.

[2] Akintaro, S. (2023). 52% of companies in Africa are unprepared for cyberattack – Report. Nairametrics, 19 July 2023. https://nairametrics.com/2023/07/19/52-of-companies-in-africa-are-unprepared-for-cyberattack-report/ (Accessed 23 July 2023).

[3] Interpol. (2021). African Cyberthreat Assessment Report: interpol's key insight into cybercrime in Africa. Interpol.

[4] PwC Uganda. (2023). The financial impact of cybercrime in Uganda, 18 July 2023. https://twitter.com/PwC_UG/status/1681281401973243907 (Accessed 23 July 2023).

[5] Verizon. (2023). Verizon 2023 Data Breach Investigations Report. https://www.verizon.com/business/resources/reports/dbir/2023/master-guide/.

[6] Hayes, D. R. and Cappa, F. (2018). Open-source intelligence for risk assessment. Business Horizons, vol. 61, no. 5.

[7] Yeboah-Ofori, A., and Brimicombe, A. (2018). Cyber Intelligence and OSINT: Developing Mitigation Techniques Against Cybercrime Threats on social media. International Journal of Cyber-Security and Digital Forensics (IJCSDF), vol. 7, no. 1, pp. 87-98, 2018.

[8] Evangelista, J. R. G., Sassi, R. J., Romero, M., and Napolitano, D. (2020). Systematic Literature Review to Investigate the Application of Open-Source Intelligence (OSINT) with Artificial Intelligence. Journal of Applied Security Research, vol. 16, no. 3, pp. 345-369.

[9] Niimi, A. (2022). Analysis of Location-Based Tweets Related to Covid-19 on Social Networking Services. International Journal for Information Security Research (IJISR), vol. 12, no. 1, pp. 1024-1031.

[10] Shu, K., Silva, A., Sampson, J., and Liu, H. (2018). Understanding Cyber Attack Behaviors with Sentiment Information on social media. In Social, Cultural, and Behavioral Modeling, Springer.

[11] Huang, S.Y., and Ban, T. (2020). Monitoring social media for Vulnerability-Threat Prediction and Topic Analysis. In IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guanzhou, China.

[12] Zaimy, N. A., Saip, M.A., and Fikri, M. (2023). Cybersecurity Threat in social media: A Bibliometric Analysis. Borneo International Journal, vol. 6, no. 1, pp. 80-86.

[13] Gagiano, M., and V. Marivate, V. (2023). Emotionally driven fake news in South Africa. In Proceedings of Society 5.0 Conference 2023.

[14] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, vol. 78, pp. 15169-15211, 2019.

[15] Roehrick, K. (2020). Vader: Valence Aware Dictionary and sEntiment Reasoner (VADER).

[16] Kyrdoda, Y., Marzi, Y., Dabić, G., and Daim, T. (2023). Cybersecurity Technology: An Analysis of the Topic from 2011 to 2021. In: Daim, T.U., Dabić, M. (eds) Cybersecurity. Applied Innovation and Technology Management, Springer, Cham.

[17] Keliris, A., Konstantinou, C., Sazos, M., and Maniatakos, M. (2019). Open-Source Intelligence for Energy Sector Attacks. In Critical Infrastructure Security and Resilience, Advanced Sciences and Technologies for Security Applications, Switzerland, Springer Nature, 2019, pp. 261-282.

[18] Hayes, D., and Cappe, F. (2018). Open-source intelligence for risk assessment. Business Horizons, vol. 61, pp. 689-697.

[19] Kumar, S., and Carley, K. M. (2016). Understanding DDoS cyber-attacks using social media analytics. In IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA.

[20] Schiappa, M., Chantry, G., and Garibay, I. (2019). Cyber Security in a Complex Community: A Social Media Analysis on Common Vulnerabilities and Exposures. In Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).

[21] Ruby, D. (2023). Twitter X Statistics In 2023 — (Users, Revenue & Trends), 11 July 2023. https://www.demandsage.com/twitter-statistics/. (Accessed 30 July 2023).

[22] Thiago, D. P. F., and Lopes, A. D. A. (2016). On the equivalence between algorithms for Non-negative Matrix Factorization and Latent Dirichlet Allocation. ESANN, p. 6.

[23] TF-IDF. (2011). Encyclopedia of Machine Learning, Encyclopedia of Machine Learning, pp. 986-987.

[24] Kenya News. (2023). Cyber Attack: 8 Major Cyber Attacks That Have Occurred in Kenya in Recent Years, KE Kenya News, 28 July 2023. https://www.ghanamma.com/ke/2023/07/28/cyber-attack-8-major-cyber-attacks-that-have-occurred-in-kenya-in-recent-years/ (Accessed 08 August 2023).

[25] Business Today. (2023). Kenya Records 278 million Cyberattacks in 3 Just Months, Business Today, 7 Feb 2023. https://businesstoday.co.ke/kenya-records-278-million-cyberattacks-in-3-just-months/ (Accessed 07 August 2023).

[26] Shihab, E., and Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and VADER sentiment. In proceedings of the international multiconference of engineers and computer scientists.

[27] Dixon, S.J. (2023). Number of global social network users 2017-2027. Statista, 13 February 2023. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (Accessed 08 August 2023).

[28] Taghandiki, K., and Mohammadi, M. (2023). Topic Modeling: Exploring the Processes, Tools, Challenges and Applications. TechRxiv.