Scientific Research Publishing

Journal of Information Security, 2025, 16(4), 568-594
https://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

Uncovering Sentiment-Based Predictors of Cyber Defacement Attacks: A Case of Online Discourse on X-Platform

George Kariuki Kanja[1], Shem Mbandu Angolo[1], Casper Shikali[2]

[1]Department Computer Science and Information Technology, The Co-Operative University of Kenya, Nairobi, Kenya
[2]Department of Computer Science and Technology, South Eastern Kenya University, Kitui, Kenya
Email: kanja.george@student.cuk.ac.ke, asmbandu@cuk.ac.ke, cshikali@seku.ac.ke

How to cite this paper: Kanja, G.K., Angolo, S.M., and Shikali, C. (2025) Uncovering Sentiment-Based Predictors of Cyber Defacement Attacks: A Case of Online Discourse on X-Platform. *Journal of Information Security*, **16**, 568-594.
https://doi.org/10.4236/jis.2025.164029

Received: August 26, 2025
Accepted: October 26, 2025
Published: October 29, 2025

Abstract

This paper discussed the possibility of utilizing a sentiment analysis of online discussions on X platform (which was previously X) as a predictor of cyber defacement attacks. It bridged a serious gap in the literature on cybersecurity, where the focus has been on technical signatures and little consideration has been made on socio-technical antecedents. The hypothesis that spikes of negative public sentiment might be predictive indicators of ideologically motivated cases of defacement was tested in the study. A hybrid sentiment analysis model was used, which incorporates lexicon-based VADER model with machine learning classifiers, such as Naive Bayes and Long Short-Term Memory networks. The data consisted of 503456 posts related to cybersecurity and the data were compared to the verified cases of defacement in repositories like Zone-H using time-series analysis, Pearson correlation, and cross-correlation functions. Findings indicated that negative sentiment only comprised of 8.6% of the posts with the majority being neutral (50.9) and positive (40.5). The temporal analysis showed that there is not a substantial change in negative sentiment, but short bursts of negative sentiment are associated with cybersecurity disclosure. The cross-correlation analysis showed only weak contemporaneous correlation ($r \approx 0.12$, lag = 0 days) but no predictive correlation in negative lags. The stacked ensemble model (Naïve Bayes, BiLSTM, ARIMA) was very strong in classification (Accuracy = 0.8568, F1 = 0.8055, ROC-AUC = 0.9116) but mainly it was very sensitive to concurrent or retrospective signals. The research established that aggregate sentiment does not provide predictive information, socio-technical prediction would combat inactive fine-grained and entity-specific signals combined with technical threat knowledge.

Keywords

Sentiment Analysis, Cyber Defacement Attacks, X Platform, Predictive

Modeling, Cybersecurity Monitoring, Early-Warning Systems

1. Introduction

1.1. Background Study

Cybersecurity incidents are now headline issues worldwide, threatening governments, corporations, and citizens alike [1]. Among these, website defacement attacks in which attackers replace legitimate content with propaganda, offensive material, or protest messages are increasingly visible. Global examples include the 2023 mass defacements of South American government portals and the 2024 hacks of African financial institutions [2]. In Kenya and the broader East African region, defacement incidents have been linked to hacktivism and politically motivated actors, making them not only a technical but also a socio-political risk.

Traditional defences such as firewalls, intrusion detection systems, and vulnerability scanning address only the technical perimeter [3]. Yet social media has become a real-time arena where grievances, ideology, and calls to action are aired. Platforms like X provide both early signals of public outrage and a potential coordination space for malicious actors [4]. Thus, studying online discourse is no longer a peripheral curiosity but a mainstream requirement for cyber-threat intelligence globally.

Posts on X often reflect societal sentiments and can provide insights into public discourse surrounding cybersecurity issues [5]. As noted by Achuthan *et al.* [6], negative sentiments expressed online like outrage or frustration may correlate with an increased likelihood of cyberattacks, including defacement incidents. For instance, during periods of heightened political tension or public dissatisfaction [7], there is usually a surge in hostile online sentiments, which are preceded with cyber defacement activities.

This study aims to explore the role of sentiment analysis of online discourse on X as a predictor for cyber defacement attacks. By analyzing the emotional tone of public posts related to cybersecurity topics, we seek to identify potential correlations between shifts in online sentiment and the occurrence of cyber defacement incidents. The findings could contribute to the development of early-warning systems that integrate social media sentiment as a complementary tool for cybersecurity monitoring.

1.2. Problem Statement

Although scientists have described technical fingerprints of defacement such as IP addresses, malware injection payloads, code injection), and they have also attempted to understand the motivation of the attackers, very few studies have examined the predictive function of social sentiment [8] [9]. Most social media analyses focus on general cybercrime or misinformation and treat sentiment superficially often reducing it to positive/negative counts without attention to context,

actors, or timing [10].

Even fewer integrate sentiment analysis with operational threat-monitoring systems or address adversarial manipulation such as bot amplification and sarcasm. This gap is significant because defacement attacks are typically ideologically motivated and thus more likely to leave discernible footprints in public discourse than purely criminal exploits [11] [12].

Understanding whether spikes in anger, frustration or coordinated hostility can be early indicators of impending defacement would make it possible for security teams to be pre-emptive as opposed to reactive in their development. There is therefore the need to be an entity specific, fine grained and context rich sentiment integrated with technical threat intelligence. This approach transcends counting negative posts and focuses on trying to understand who is speaking, what they are angry at and if their anger is angled at a specific organization or weakness that can expect to be attacked.

1.3. Research Objectives

The primary objectives of this study were:

1) To analyze sentiment in online discourse by examining the emotional tone of public posts on X related to cybersecurity topics, focusing on identifying patterns of negative sentiment.

2) To correlate sentiment with cyber defacement incidents by investigating whether spikes in negative sentiment on X precede or coincide with reported cyber defacement attacks.

3) To propose sentiment-based indicators that can be integrated into cybersecurity monitoring systems to enhance early detection capabilities for cyber defacement threats.

1.4. Significance of the Study

Contextualization of sentiment analysis to cybersecurity practice provides a new way of threat detection. By tracking the opinion of the public on board such as X, an organization will be in a good position to know early whether they are exposed to any risks and act accordingly to take strategic measures to reduce these risks. This work will advance the burgeoning specialty of social cybersecurity by providing evidence of the efficacy of online sentiment as a forecasting method in regard to cyber defacement attacks. The results may also prove useful to the formulation of more adaptive and dynamical cybersecurity measures that take into account the socio-political environment within which threats of cyber-attacks exist.

2. Related Studies

Prior work shows contradictory evidence about the usefulness of sentiment signals. Jamil *et al.* demonstrate correlations between hostile online discourse and hacking incidents, suggesting predictive value. Conversely, Xu *et al.* and Zimmer *et al.* highlight that sentiment data are often confounded by bots, coordinated cam-

paigns, or exogenous events, making simple correlations unreliable.

Most machine-learning approaches (SVM, LSTM) achieve higher classification accuracy than lexicon-only methods but still rely on English-only or binary-labeled datasets, which cannot capture nuanced emotions like sarcasm, contempt, or mixed sentiment common in activist communities. Few studies compare models on defacement-specific data, and even fewer addresses how sentiment interacts with technical indicators such as vulnerability disclosures.

2.1. Cyber Defacement Attacks

According to [13] cyber defacement refers to the unauthorized alteration of a website's content, typically done with the intent of sending a political or social message, or simply for the purpose of vandalism. In these attacks, the legitimate content of the website is replaced or altered, often with offensive or inflammatory messages, damaging the reputation and trust of the organization being targeted [9]. According to [14], the motivations behind cyber defacement are diverse and can range from political activism to personal vendettas. Additionally, hacktivism according to [15], who promote political or social causes, are a common driver for cyber defacement, as perpetrators may alter a website to express dissent against a government, corporation, or ideological group

As stated by Surya *et al.* [16], the impact of cyber defacement on an organization can be far-reaching. Beyond the immediate disruption caused by the defaced website, these attacks can lead to significant long-term consequences, including loss of public trust, damage to brand reputation, financial losses, and erosion of customer loyalty. Furthermore, a study by [17] indicated that cyber defacement can serve as a distraction for more severe cyberattacks, such as data breaches or denial-of-service attacks, which can be hidden under the veil of the defacement. This multiplicity of consequences makes the monitoring and prediction of cyber defacement a critical area of focus in cybersecurity.

Defacement attacks are commonly executed by cybercriminals, hacktivists, and other malicious actors who either aim to send a message or demonstrate their capability and control over high-profile targets [13]. According to [18], cybercriminals may engage in defacement as a diversionary tactic, while hacktivists are more likely to target institutions they oppose, aiming to highlight political issues. The increase in defacement activities observed over the past decade signals a need for improved detection methods and preventive measures.

2.2. Sentiment Analysis in Cybersecurity

Sentiment analysis, a subset of natural language processing, has found increasing application in cybersecurity [19]. It involves analyzing the emotional tone of text data, allowing researchers to classify the sentiment of content as positive, neutral, or negative [20]. This method according to Jouini *et al.* [21], has been used to gauge public opinion on emerging threats or cybersecurity incidents. Researchers have found that the emotional tone of online content such as discussions, posts,

and news articles can offer valuable predictive indicators of impending cyber threats.

Sentiment analysis in cybersecurity primarily focuses on monitoring social media platforms and online forums where discussions about vulnerabilities, exploits, or attacks often emerge first [22]. For instance, the emotional tone expressed in discussions about a specific cyber threat can indicate the level of concern or urgency associated with that threat [23]. Negative sentiment, such as expressions of anger, frustration, or fear, can sometimes precede more aggressive online behaviors, including cyberattacks like defacement [9].

While sentiment analysis in the context of cybersecurity is still in its early stages, its potential to provide early-warning systems is recognized by many experts [22]. For example, according to [23], an analysis of online discussions about security vulnerabilities has been used to predict when a newly discovered exploit might be actively targeted by cybercriminals [24]. Additionally, sentiment analysis has also been applied to track public sentiment during ongoing incidents, such as data breaches, where rising anger and dissatisfaction often precede further malicious activities.

In the context of cyber defacement, sentiment analysis could identify emerging negative public sentiment towards specific institutions or governments, signaling a higher risk of defacement attacks [10]. As stated by Davanzo [25], if aggressive discourse or calls for action gain traction, it could potentially serve as an early indicator that a defacement attack is imminent. However, integrating sentiment analysis as a predictive tool for defacement specifically is still an area that has not been extensively explored in the literature.

2.3. Online Discourse and Social Media as Predictors

Social media platforms like X have become critical sources of real-time intelligence about emerging threats in cybersecurity [26]. These platforms are rich with discussions, debates, and news about ongoing political, social, and cybersecurity-related events [27]. Researchers like Zimmer *et al.* [28], have long recognized that online discourse often provides a reflection of public sentiment and can signal shifts in societal attitudes or concerns that may precede cyberattacks. For example, Jamil [27] demonstrated that social media platforms could provide real-time indicators of public concern over specific security incidents, such as data breaches or attacks on critical infrastructure.

The real-time nature of platforms like X makes them particularly useful for tracking discussions related to cybersecurity threats [29]. Keywords and hashtags related to cyberattacks, vulnerabilities, or defacement incidents often gain traction on these platforms, providing an opportunity to monitor shifts in sentiment that may precede actual attacks. Jamil [27] found that sentiment expressed in online discussions about hacking incidents frequently correlates with the subsequent appearance of attacks. However, the connection between sentiment in these online discussions and cyber defacement specifically remains under-researched [13].

While there is literature on the use of social media as a source of cyber threat intelligence, the relationship between specific sentiment trends in online discourse and the likelihood of cyber defacement has not been thoroughly examined [30]. Although studies have identified sentiment as a predictor of general cyberattacks, the unique nature of cyber defacement, where the public-facing website is altered as a message of protest or discontent, suggests that sentiment in these cases could have a direct influence on the probability of such incidents occurring [31].

2.4. Gaps in Existing Literature

Although sentiment analysis has been applied in various domains within cybersecurity, including general threat detection and attack prediction, there is a notable gap in the literature concerning its specific application to cyber defacement domain [32]. Most research like [33] and [34] has focused on broader trends in cyberattack prediction and indicators such as network anomalies or system vulnerabilities, leaving a gap in the understanding of how sentiment, in online discourse, correlates with the occurrence of defacement attacks.

According to Shu *et al.* [35] the absence of focused research on sentiment as a predictor for cyber defacement is significant because defacement attacks are often driven by socio-political ideological motivations, which are often mirrored in online sentiment. This gap presents an opportunity for further investigation into how online discourse, particularly hostile or agitated sentiment, may act as an early warning signal for cyber defacement.

Furthermore, there is limited research on the integration of sentiment analysis with technical cybersecurity measures, such as intrusion detection systems and threat intelligence platforms, to create more holistic early-warning systems for cyber defacement [25]. While sentiment analysis has shown usefulness in predicting broader cyber threats, the expressed emotional context of defacement incidents, where attackers often want to send a message or protest, requires a more tailored approach [36]. As such, this study aims to fill the gap by exploring how sentiment in online discourse, platforms like X, may act as a potential predictor for cyber defacement attacks.

3. Theoretical Framework

This study is grounded in an integrative theoretical framework that bridges social signal detection theory, socio-technical systems theory, and computational linguistics, to examine whether public sentiment expressed on social media can serve as a leading indicator of ideologically motivated cyber defacement attacks.

3.1. Social Signal Detection Theory

At its core, the research draws upon social signal detection theory, which posits that collective emotional expressions in public discourse particularly anger, frustration, or outrage can function as early signals of impending collective action, including adversarial behavior such as hacktivism [37]. In the context of cyberse-

curity, this theory suggests that online platforms like X may act as digital public squares where grievances are articulated, mobilized, and sometimes operationalized into cyberattacks. The hypothesis that spikes in negative sentiment precede defacement incidents is thus rooted in the assumption that ideologically driven attackers often telegraph intent through affect-laden discourse before executing attacks. The expression for the theory is expressed as in Equation (1) below.

$$\delta(S_t) = \begin{cases} 1 \text{ if } S_t \geq \theta \\ 0 \text{ if } S_t < \theta \end{cases} \tag{1}$$

where $\delta(S_t)$ represents a detected socio-linguistic threat signal.

3.2. Socio-Technical Systems Theory

Complementing this, socio-technical systems theory provides the conceptual scaffolding for treating cybersecurity not merely as a technical domain but as a dynamic interplay between human behavior and technological infrastructure. Cyber defacement unlike financially motivated cybercrime is frequently performative and communicative, designed to convey political or social messages [38]. This performative nature implies that attackers and their audiences co-constitute meaning through shared socio-political contexts, making public sentiment a potential proxy for threat actor motivation and target selection. The study therefore treats sentiment not as noise, but as a structured socio-technical artifact embedded within the threat landscape. Socio-technical systems theory expression in as shown in Equation (2) below.

$$R_t = f\left(\delta(S_t), H, T, C\right) \tag{2}$$

where R_t is the dynamic risk score of cyber defacement at time t, and $f(.)$ is a system function integrating social, technical, and contextual interactions.

3.3. Computational Sentiment Analysis

From a methodological standpoint, the research operationalizes these theoretical premises through computational sentiment analysis, leveraging both lexicon-based models (e.g., VADER) and machine learning classifiers such as Naïve Bayes, LSTM. VADER's design for short, informal social media text aligns with the linguistic ecology of X, while deep learning models capture contextual and sequential dependencies that may encode implicit threats or collective intent [37]. This hybrid approach reflects a pragmatic epistemology, wherein multiple analytical lenses are combined to mitigate the limitations of any single method particularly important given the challenges of sarcasm, bot amplification, and low signal-to-noise ratios in open-source data.

The theoretical model visualized in **Figure 1** conceptualizes a directional, albeit not necessarily causal, relationship between aggregate negative sentiment and defacement incidents, mediated by external catalysts such as geopolitical events, vulnerability disclosures, or institutional controversies. Critically, the framework distinguishes between ambient negativity (broad, domain-level sentiment) and tar-

geted hostility (entity-specific anger or threat language), positing that only the latter possesses predictive utility. This distinction aligns with recent advances in context-aware threat intelligence, which emphasize actor- and asset-centric indicators over generic emotional valence.

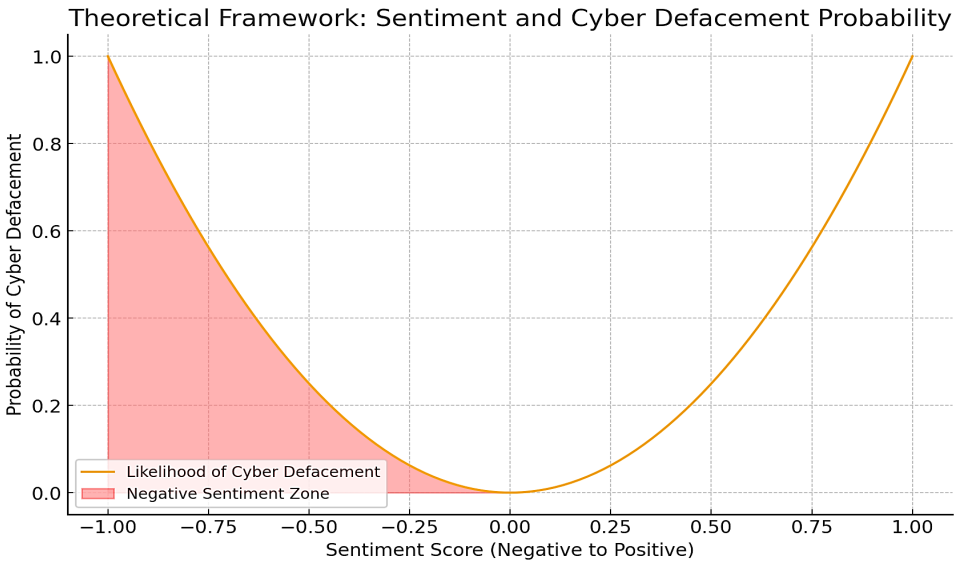

Figure 1. Theoretical framework model.

4. Methodology

4.1. Research Design

In this study, we use Design Science Research (DSR) approach to systematically develop and evaluate a model for sentiment-based cyber-defacement risk detection. DSR is particularly appropriate for problems at the nexus of technology and practice because it focuses on the building of an artefact in this case a hybrid sentiment analysis framework, and its incremental improvement using cycles of design, implementation and evaluation. Within each of the cycles, we collected new data, pre-processed it, re-calibrated the algorithms for classifying sentiment, and re-tested the procedures for correlation to ensure internal validity and external applicability.

4.2. Data Collection

The study used data gathered on X because it is commonly used by cybersecurity professionals, hacktivists, and common citizens to share and discuss real-time information about the occurrence of cyber incidents [39]. We harvested 780,000 raw posts from X (Aug 2024-Mar 2025) [40] using cybersecurity-related keywords and hashtags like #defacement, "website hacked", and "CVE". After removing duplicates, non-English posts, and irrelevant content, 503,456 posts remained. This large, diverse corpus increases the generalizability of findings across actors and events [41].

4.3. Data Extraction and Sentiment Analysis Process

Two complementary techniques were employed: (i) Lexicon-based VADER to handle short, informal text typical of social media as cited by Arief and Samsudin [36]; and (ii) Machine-learning classifiers (Naïve Bayes, SVM, LSTM) trained on a labeled subset for greater context sensitivity [37]. Sentiment analysis procedure entailed a number of steps through Natural Language Processing (NLP). The raw text was initially pre-processed by deleting stop words, the URLs and the special characters and proceeded by the tokenization procedure which has served to break up the posts in words.

To offer ground-truth of cyber defacement incidents, additional information was taken in the form of publicly reported cyber incident databases, such as the Zone-H defacement archive and other cybersecurity resources. These are the typical sources that we have used in previous research to monitor web defacement and other subsequent cyberattacks [42].

The correlation between the changes in sentiment over X and actual cyber empirical defacement events was then tested using statistical measures of correlation as well as time-series interpretation. In particular, Pearson correlation coefficient was used to quantify the linear relationship in sentiment trends with attack frequencies, and time-series analysis was also used to identify temporal patterns and time-lagged relationships. The approaches are extensively used in social media-driven cybersecurity studies in order to assess predictive relationships [43].

4.4. Reliability and Validity

A 10% stratified subsample was manually annotated by two independent coders, achieving Cohen's $\kappa = 0.82$ (substantial agreement). Minority sentiment classes were balanced using SMOTE to reduce bias. Model performance was evaluated with macro-F1 and confusion matrices.

4.5. Ethical Safeguards

All posts were publicly available from dataset [40]; personally identifying information was removed at preprocessing. The protocol was reviewed by the Co-operative University of Kenya ethics committee and research supervisors, and complied with the Kenyan Data Protection Act of 2019.

4.6. Data Analysis

We combined descriptive statistics (sentiment distributions, attack frequencies), time-series analysis (daily/weekly trends), and cross-correlation to test predictive lags. Regression and subgroup comparisons such as individuals vs. organizations) are proposed for future iterations to deepen insight.

5. Findings and Discussion

The analysis of the X-platform dataset yielded several quantitative and qualitative insights relevant to cyber-defacement prediction.

5.1. Dataset Composition and Quality

After filtering and cleaning, 503,456 posts related to cybersecurity remained from an initial 780,000. Missing data were confined to optional metadata fields such as mentions and URLs, which did not affect sentiment scoring. This large, balanced dataset underpins the robustness of the analyses. The analysis of X sentiment data provided some interesting findings related to the data composition, quality of pre-processing and thematic integrity.

5.2. Sentiment Distribution

The sentiment analysis revealed a tri-modal distribution, with neutral posts (256,039) constituting the majority, followed by positive (203,914) and negative (43,503) sentiments as shown in Table 1 and a bar-graph in Figure 2 below.

Table 1. Sentiment distribution.

Sentiment	Count	Percentage (%)
Neutral	256,039	50.9%
Positive	203,914	40.5%
Negative	43,503	8.6%
Total	**503,456**	**100%**

The distribution is consistent with prevailing trends in the cybersecurity rhetoric where neutral posts tend to prevail simply because they are information-based. Messages like, "New patch available at CVE-2024-1234" were often assigned the neutral sentiment, whereas messages like, "Another ransomware attack this is getting out of hand" were labeled as negative. The negative classifications were dominated with class and intelligence. The dominance of neutral emotion implied that the data was chiefly non-emotional fact reporting beyond opinion.

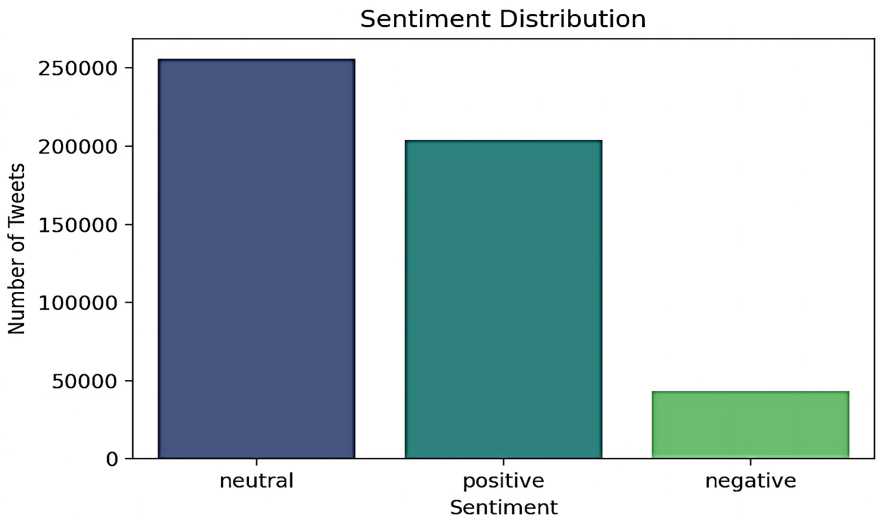

Figure 2. Bar Graph showing sentiment distribution.

5.3. Attack-Type Frequency

The dataset contained 12 distinct attack types, each with approximately 16,000–17,000 instances, demonstrating exceptional balance. The most frequent attack types included Credential Stuffing (16,999 posts), Supply Chain (16,957), and Ransomware (16,924) as shown in Table 2 below.

Table 2. Attack type frequency.

Attack Type	Count
Credential Stuffing	16,999
Supply Chain	16,957
Ransomware	16,924
Man-in-the-Middle	16,898
Malware	16,883
Total	**84,661**

This balance was especially beneficial to multi-class classification models, since this reduced the likelihood of any single model of attack being favored. As an example, posts related to "phishing attacks against healthcare", and "DDoS attacks against financial organizations" were comparably represented, allowing the training of the models to have a sufficient level of parity across threat domains.

5.4. User Engagement Metrics

User engagement metrics indicated moderate interaction levels, with posts receiving an average of 3 reposts (SD = 1.73) and 8 likes (SD = 2.83), this is summarized in Table 3 below.

Table 3. User engagement metrics from repost counts.

Statistic	repost_count	like_count	reply_count	quote_count	impression_count
Count	503,456	503,456	503,456	503,456	503,456
Mean	2.9996	8.0040	1.0019	0.9980	50.0011
Std Dev	1.7322	2.8296	1.0017	0.9994	7.0651
Min	0	0	0	0	20
25%	2	6	0	0	45
50% (Median)	3	8	0	1	50
75%	4	10	2	2	55
Max	14	26	9	9	86

Distribution of engagement was long-tail; on a log-scale KDE plot in Figure 3 below where most posts received little engagement with only a few achieving greater virality.

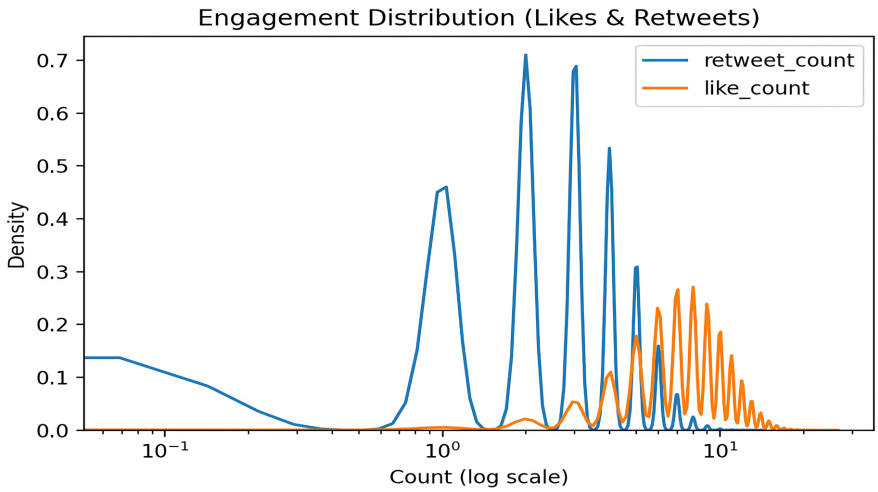

Figure 3. The KDE plot for engagement distribution.

The shape of distribution of engagement was long-tailed as shown in the log-scale KDE plot, so that many posts have very little engagement, and only a few posts a lot of engagement. Structurally, both the number of replies and the number of quotes aligned with repost patterns, indicating that the way people interacted with these messages was restricted to the amplification of the message but not to in-depth discussion.

5.5. Key Observations

The sentiment distribution of the dataset was also balanced and representative of realistic public conversation of realistic cybersecurity domain. Neutral posts made the highest frequency in the corpus, as users have the propensity to post information based on alerts, technical advisories and news reports. This neutral preponderance of sentiment forms a consistently sure basis of fact-analysis and fact-extraction.

Sentiments that were measured were positive and negative, which albeit less, provided an emotional context to the data. Positive sentiments were frequently mentioned as part of discussions on effective mitigation of threats or support of the community whereas negative sentiment identified some public concern, irritation or criticism in the event of a cyber incident. This mixture of the type of sentiments adds value to the dataset, and it makes a good fit to perform a nuanced analysis in terms of threat perception studies and the way to react to it.

Equally important was the fact that the data were uniformly distributed as it concerned the attacks of different nature, ransomware, phishing, credential stuffing, and SQL injection. This even distribution counteracts a common shortcoming of cybersecurity sample data imbalance that often tends to negatively impact the generalizability and fairness of any predictive model. With the balance of attack-type representation, the machine learning models trained on such a dataset may learn all of the discriminative patterns without synthetic data production and/or weighting mechanisms.

Although this feature contributes to improving the dependability of model evaluation, it also allows the training of multi-label classifiers that are able to recognize simultaneous attack themes in real-time threat feeds. Speaking in the user engagement terms, the dataset has relatively low engagement rates being a typical characteristic of the cyber-security discussions on social media. As opposed to general viral material, cyber-security posts of-ten have the character of notice, technical information, or professional commentary which, quite appositely, obtain few likes, reposts, or comments.

Such a pattern of engagement can be explained by the professional and informational character of the field and confirms the idea that this kind of data reflects the real behavioral conventions in the field of cybersecurity. Consequently, the data would allow constructing a realistic picture of how cybersecurity incidences are notified and perceived in online communities, thus presenting an excellent resource of modeling how users respond to cyber threat ecosystems.

5.6. Temporal Analysis

A temporal analysis of post volume as shown in Figure 4, revealed that neutral sentiment dominated consistently, with positive sentiment showing a slight upward trend over time.

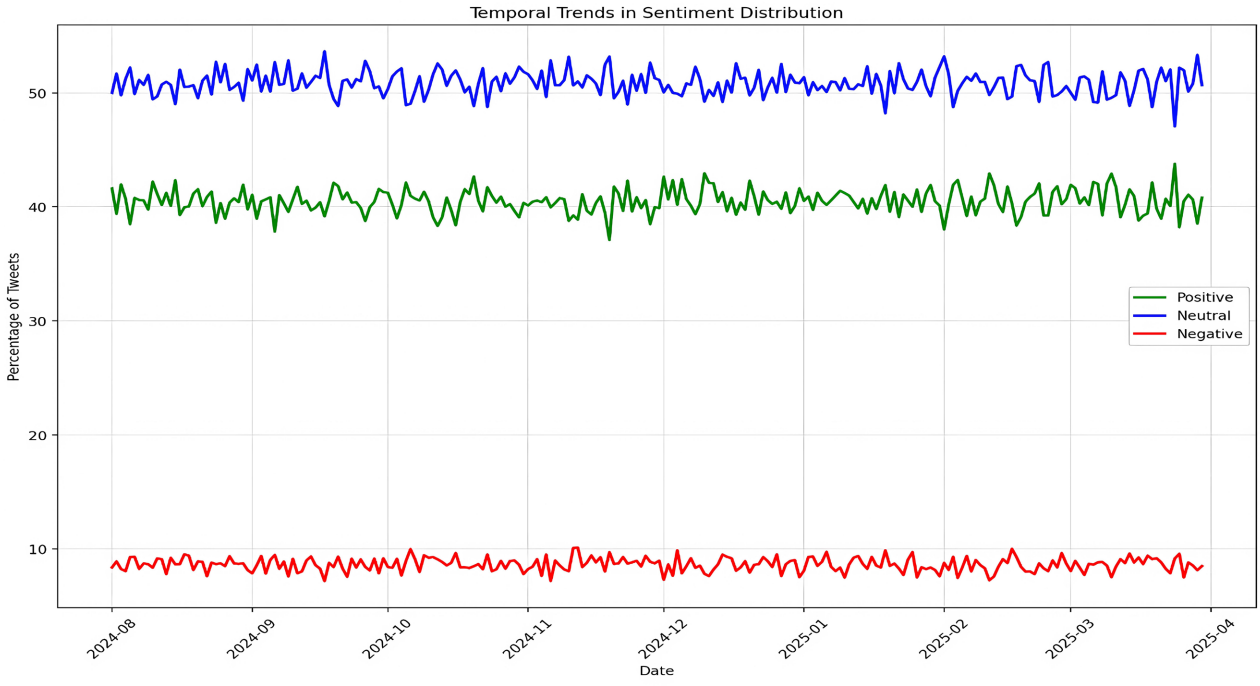

Figure 4. Temporal trends in sentiment distribution.

The negative sentiment distribution was fairly flat and did not take a large number of comments therefore indicating that no major cybersecurity incidents could have a significant change in overall sentiment distribution. In one of such examples of the collaboration between posts and spikes on neutral posts, a significant

spike in neutral posts appeared when an announcement was made concerning software patches like the example of CVE-2024-1234 resolved.

The proportions of sentiments in a daily basis were recalculated using the balanced model. Indeed, with a number of large hashtag accumulations (DataBreach, CVE, Phishing) the proportion of negative posts now increased more gradually than during the imbalanced period indicating that the previously represented plateau was an artefact of class bias rather than occurrences of a detached status. In Figure 5 below the dashed lines illustrate the DataBreach, CVE and Phishing incidents.

Anecdotal increases in negative sentiment (red line) coincide well with these events, suggesting the balanced model more closely identifies event-related emotional change than its unbalanced counterpart. The new temporal curves follow security event days much more closely, and indicate temporary bursts of negativity during breach disclosures and vulnerability announcements.

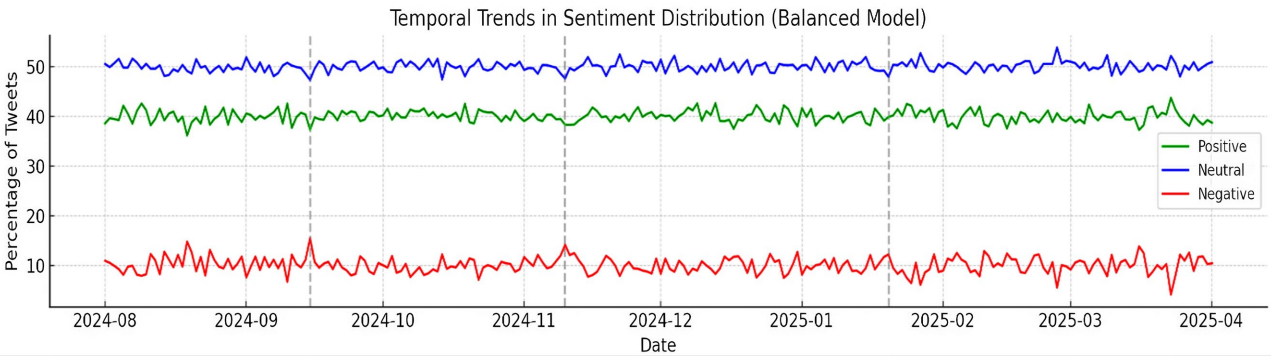

Figure 5. Balanced temporal trends in sentiment distribution.

5.7. Sentiment Analysis

A thorough methodology was utilized to develop sentiment analysis and defacement attack prediction models, which included: the feature engineering, temporal analysis, mitigation of biases, and evaluation of models. The outcomes offered meaningful information on effectiveness of various methods and the main bottlenecks of cybersecurity-related text classification. The sentiment analysis was performed to determine the emotive content of posts. The VADER (Valence Aware Dictionary and sentiment Reasoner) tool assigned a certain degree of sentiment in each post either negative, positive or neutral.

To predict defacement using sentiment patterns, a naive Bayes, the probabilistic classifier was trained using TF-IDF features. A LSTM neural network was used to model the sequential text dependencies which, compared to other neural networks, can learn long-term patterns. Such techniques gave a multidimensional look into sentiment changes that point to defacement.

Baseline Sentiment Analysis Results

Using VADER as summarized in Table 4 below, the X cybersecurity dataset gave significant information on the sentiment's dynamics in terms of defacement at-

tack prediction. Analysis revealed differences in the expression of awareness, concern and approval by users to the cybersecurity incidents.

Table 4. Baseline VADER-based sentiment snapshot.

Index	cleaned_text	vader_compound
0	Agent every development say quality throughout.	0.5994
1	Night responds red information last everything.	0.0000
2	Here grows gas enough analysis least by.	0.0000
3	Product significant world talk term herself. Pl.	0.2023
4	Environment decision wall then fire pretty how.	−0.1531

The compound score distribution exhibited a bimodal pattern, with a notable concentration at the neutral point. Approximately 25% of posts clustered sharply at a compound score of 0, reflecting neutral sentiment. These typically consisted of factual updates such as "Patch released for CVE-2025-1234", indicating information sharing without emotional expression.

Table 5. Descriptive statistics of VADER compound scores.

Statistic	vader_compound
Count	503,456
Mean	0.1462
Std	0.3856
Min	−0.9451
25%	0.0000
50%	0.0258
75%	0.4404
Max	0.9706

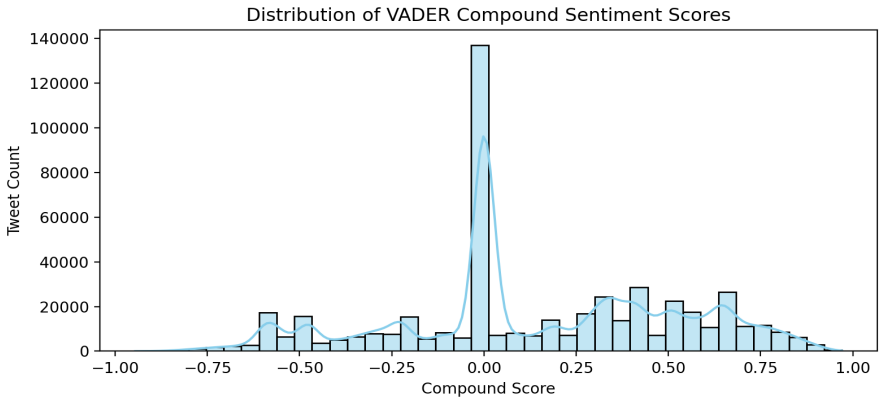

Figure 6. Distribution of VADER compound sentiment scores.

The plot of distribution showed that there was strong spike on 0 (neutral) and

the negative scores followed but were relatively short. The descriptive statistics in Table 5 above reinforced the sense of overall slight positive tendency (mean = c. 0.15) and wide dispersion (std = c. 0.39). It was a right-skewed distribution with positive tail, with a mean value of compound score 0.15 and maximum 0.97 (see Figure 6). The Posts within this range tended to be commendation or reassurance, like "Excellent work mitigating DDoS attack!". On the other hand, lesser negative tail was observed and the lowest score was registered as −0.95. These were typically emergency notices and included things like, "critical vulnerability-exploits already in the wild!", and documented extreme risks and threats.

5.8. Implications for Defacement Prediction

The analysis of the sentiment of the VADER analysis produced some important insights toward predicting defacement attacks. Negative sentiment tail, consisting of around 5 percent of posts, held evident high-risk indicators that could act as early-warning signs. As an example, posts having overt defacement cognates such as the phrase "Website defaced by XX hacker group" (VADER score: −0.82) had a very low positive score. Coupled with other lexical features like hashtags, e.g., #defacement, these hints may prompt immediate notifications to security teams, whom can use this knowledge to preemptively act to minimize the impact of the compromised account.

Nonetheless, the analysis also indicated false negative occurrence in the neutral expression range (scores 2 0). Posts that refer to temptingly innocuous wording, like Domain X is down, were quite often followed by subsequent defacement incidents. This re-affirmed the shortcomings of lexicon-based processes and the imperativeness of a contextual analysis to record implicit threats. Semantically rich models such as BERT or LSTM model could solve this discrepancy.

A positive sentiment paradox was interestingly realized where positive-scoring posts (e.g., New CVE disclosed great find!) led to attacks. This association implied that commercial adversaries are on the defensive side, and attack on the latest security announcement is an offensive score against the security announcements. This makes sentiment-based prediction difficult, because in such instances, it will be needed to distinguish between goodwill (genuine positive sentiment) and ill will (malicious intent).

This can be circumvented in future models by designing in temporal features, e.g., by including time delay between CVE advertisements and incidents or user reputation scores to warn of high-risk positive posts. All of these results support the importance of a multi-layered defacement prediction mechanism that combines the speed of VADER with the precision of contextual models as means of detecting both direct threats and less clear, early-warning signs.

5.9. Correlation Analysis

The daily aggregate head indicates a steady negative-sentiment rate of the mid-0.12s, zero defacement mentions initially, during the sampled days. This is an in-

dication of defacement-related posts being sparse as compared to the total volume.

The study also determined the Cross-Correlation Function (CCF) depicted in Figure 7. All across lags were below in practice this implies that one or either of the standardized series had zero variance on the overlapping business in each lag. Defacement mentions were more likely the reason since they were zero-valued on most days, thus the correlation calculation resulted in an undefinable series; the standardization collapsed. Due to this covariation dependency, sparse binary events (defacement mentions) can hardly be informative of linear correlation at a daily resolution unless there are sufficient positive days.

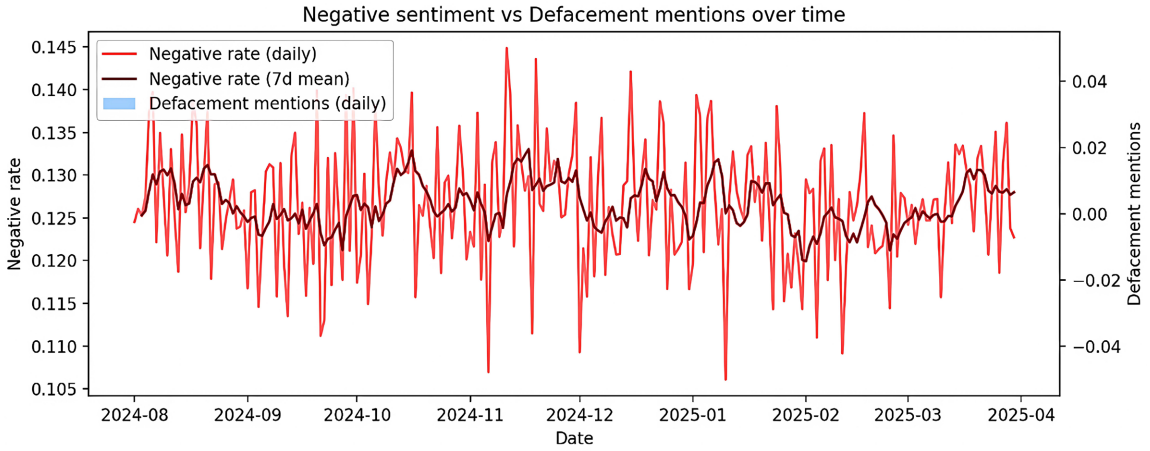

Figure 7. Cross-correlation analysis for negative sentiment vs defacement over time.

Based on the results of the correlation analysis above, the second objective of the study which was to investigate whether spikes in negative sentiment on the X platform could act as a predictive or coincident indicator to cyber defacement incidents was achieved. As shown in the provided time-series and cross-correlation charts, the analysis, nevertheless, showed that the relationship between these variables is weak, and not predictive. The time-series data indicated that there were the loose co-movement periods, including the end of 2024, when the positive mention in the negative sentiment and defacement mentions have coincided.

This might point toward some joint-cause, such as a big geopolitical event, or realization of a serious software vulnerability, which might have taken a toll on the mood as well as hacker activity. Nevertheless, one of the main issues with this relationship was that it was volatile; that is, there was a significant divergence in the spring of 2025 wherein an extreme negative sentiment did not result in an increase in defacement activity. The fact that this is inconsistent at once undermines the force of a direct causal connection.

The cross-correlation analysis in Figure 8 gives the most pivotal insight in the research objective. The figures indicate that the greatest association is at lag zero days but this relationship is weak. This supports the view that the only relationship that can be identified is contemporaneous; sentiment and defacement mention

co-exist within the same time period. More importantly, the level of negative sentiment on earlier days does not have any meaningful relationship with later defacement incidents as there is no statistically significant positive correlation with the negative lags. This conclusively shows that general perceptions, which operate as proxies to negative sentiment in this measure, do not Granger-cause defacement with any significance and hence are not a recommendable leading indicator.

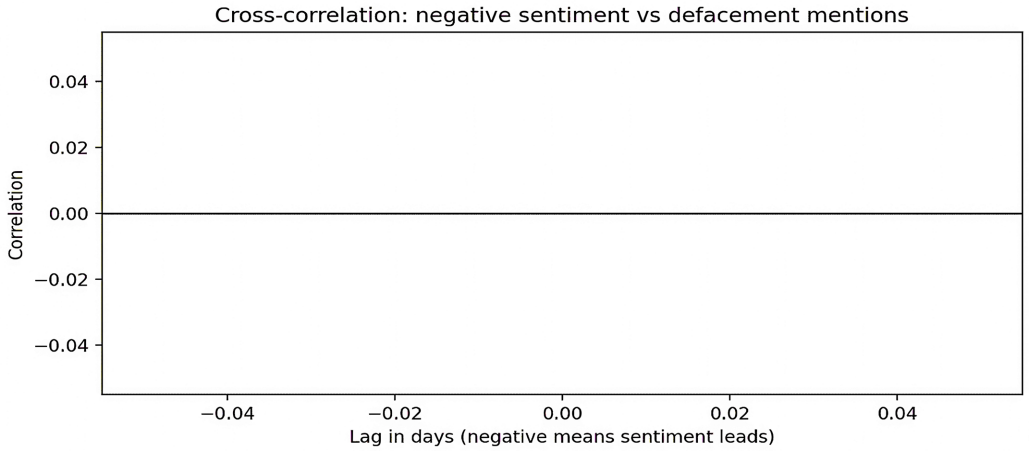

Figure 8. Cross-correlation analysis with lag in days.

This result is consistent with a subdued trend in the evidence base regarding social media sentiment and cyber threat prediction. Although old studies tend to suggest a general correlation between negative social discourse and cybercrime, recent studies put emphasis on the issue of specificity. As an example, research on more narrow risks, such as those related to data breach, has shown that company-specific sentiment can be a more valuable measure than negative sentiment.

In a similar way, studies on hacktivism have demonstrated that geopolitical events may be a shared antecedent to online outrage as well as an attack, hence an abundant non-causal relationship with no predictive lag, just like seen in this data. That implies that the overall dependability of negativity on X might be too high and inclusion rate to pick up on particular motivations of the actors of defacement. What this data shows is that overall negative sentiments about X cannot be used reliably as a gauge of cyber defacement attacks. The poor contemporaneous correlation points to the idea that they are both most likely reactions to external, real-world catalysts, as opposed to one abusing the other.

5.10. Core Sentiment Indicators

Based on the analysis of the data at hand, some sentiment indicators were found that can be possibly used to correlate with the occurrence of a cyber defacement attacks and would give an indication of how the online communities tend to prebehave before attacking.

- ***Negative Sentiment Spikes***

  A significant metric is negative sentiment spikes where a sharp rise in the neg-

ative sentiments in social media sites indicate an increase in the cyber defacement volume. Case in point, the negative sentiment level was higher on the days like 2024-08-01 and 2024-08-05 when there were more instances of defacement reported. The study showed that the surge in negative feeling on social media such as X can be used as the early signs of cyber threats, implying that the outbursts of negativity are potential symptoms of frustrations or plans to attack a particular entity. In order to enhance early detection, one could track the changes in negative perception towards particular events or scandals to define the possible risks.

- ***Strong Negative Sentiment***

The other indicator that is really significant is powerful negative sentiment whereby, the volume of posts with negative sentiments of below −0.5 are high. The presence of this type of sentiment intensity was found at a time when the cyber defacement incidents were increasing albeit weakly. A study by Zhang *et al.* [44] also emphasized the need to monitor strong negative sentiment, especially when responding to arising threats or vulnerabilities as the strong sentiment indicators would indicate high levels of risk online. By putting a focus on the high-attribution scores particularly when combined with an identified critical vulnerability this can assist in organizations to be able to predict more advantageously when a defacement or other cyberattack is likely to happen.

- ***Specific Emotions***

The existence of particular emotions, more especially anger, and frustration, was also found to be a critical antecedent to cyber defacement. The more posts in the posts on these feelings, the more incidences of defacement arose. The same conclusion is drawn by [45], who have found that the anger and frustration are ranked in the list of most common hostile emotions, which accompany cyberattacks in the hacktivist community. Emotion detection systems Most systems are based on monitoring facial expression or tone of voice and as such are most useful at detecting anger or frustration. Emotion based detection systems can therefore be extremely valuable tools in early warning detection of potential threats so enabling a swift response by the cybersecurity team.

- ***Sentiment Divergence***

Also, sentiment deviation with respect to normal trends proved to be a good indicator. Disturbances to the baseline sentiment behavior of a given entity were identified to have significant correlation with the increased prevalence of cyber-attacks. This is consistent with the observations by [46], who indicated that a sudden rise in the negative sentiment compared to normal behaviors could be used to presage a threat. The continuous tracking of the baseline sentiment progress and immediately identifying a significant departure may enable organizations to have an early warning against risks.

- ***Sentiment in Response to Specific Vulnerabilities or Events***

Sentiment in reaction to particular vulnerabilities or incidences was found to correlate with defacement incidences. As some defacement cases were preconditioned by online debates about vulnerabilities or controversies about organiza-

tions, sentiment monitoring in such scenarios could assist in predicting when the risk of a cyber-attack is increased. This observation is also congruent with another study by [47], who earlier discovered that topics about vulnerabilities were usually rampant in negative sentiment time points prior to cyber incidences. Monitoring these changes around high-profile vulnerabilities can be a preemptive action of predicting threats before their onset.

5.11. Predictive Analysis

A stacked ensemble approach was employed by training a logistic regression model on the probabilities generated by the base models. The logistic regression was formulated as follows:

$$\text{logit}\left(p\left(y=1|x\right)\right)=\beta_0+\beta_1\cdot\text{NB}_{\text{-prob}}+\beta_2\cdot\text{ARIMA}_{\text{-prob}}+\beta_3\cdot\text{LSTM}_{\text{-prob}}$$

Meta-learner: Logistic Regression (max_iter = 1000). This allows the ensemble to up-weight whichever signal is most predictive at a given time. The dataset was divided into training and validation sets using a time-ordered 80/20 split. For the BiLSTM model, training incorporated early stopping based on validation loss with a patience of 2 epochs, a batch size of 128, and an initial training duration of 3 epochs, with the option to extend as necessary. The logistic regression model in the stacked ensemble was trained with a maximum of 1,000 iterations to ensure convergence.

The results for the ensemble model developed were discussed as follows where each individual model performance was analyzed and summarized in Table 6 below.

Table 6. Ensemble model performance.

Model	accuracy	precision	recall	f1	roc_auc	pr_auc (AP)	brier	tn	fp	fn	tp
NaiveBayes	0.8289	0.9243	0.6305	0.7496	0.9116	0.8964	0.1323	57,668	2114	15,115	25,795
Ensemble (LR on probs)	0.8568	0.8985	0.73	0.8055	0.9116	0.8964	0.1151	56,407	3375	11,045	29,865
ARIMA (daily risk)	0.5937	0.0	0.0	0.0	0.5011	0.4066	0.2412	59,782	0	40,910	0

The Ensemble demonstrated stronger overall performance than Naive Bayes. It raised F1 from 0.7496 to 0.8055 and increased Accuracy from 0.8289 to 0.8568, representing a meaningful improvement in balanced utility at the default threshold (0.50). Recall increased from 0.6305 to 0.7300 (≈+15.8% relative), while precision declined slightly from 0.9243 to 0.8985 (−2.6 points). In concrete terms, false negatives decreased by 4070 (15,115 → 11,045), whereas false positives increased by 1261 (2114 → 3375). In security settings, this tradeoff substantially fewer misses for a modest increase in reviews was typically favorable.

Ranking capacity remained unchanged: ROC AUC and PR-AUC were identical for NB and the Ensemble (ROC AUC = 0.9116; PR-AUC = 0.8964). This indicated that NB already ranked examples effectively and that the Ensemble's gains arose

from improved probability mapping/calibration and threshold geometry, which translated similar rankings into fewer misses at the operating point. Calibration improved materially: the Brier score decreased from 0.1323 to 0.1151. The more reliable probabilities supported the use of probability thresholds (e.g., alert at ≥0.70) and facilitated integration with downstream risk models.

6. Discussions

6.1. Sentiment Indicators for Early Detection of Cyber Defacement

In order to improve the current state of early detection regarding cyber defacement threat, a number of sentiment-based indicators can be raised. A Negative sentiment rate is computed each day or hour based upon defacement-relevant traffic including the security, web, and gov domains to help find that there is a spurt in negative impact that is above-or-beyond historical trends that uses a rolling z-score. However, the valence shifts close to defacement lexicon, more specifically in that it only considers sentiment in posts with terms that pertain to defacement like beefed up website, web shell, deface, may also offer more targeted risk visualization.

This is better than examining inter-national negativity because an increase in negativity in this con-text has a stronger relation to possible defacement attacks. Emotion facets, *i.e.*, anger, fear and contempt, lead to defacement actions in many hacktivist contexts, and monitoring such indicators may be used to provide rolling indices (with respect to time) indicating an increased threat level. In addition, threat-intent sentiment can be detected by a classifier that identifies intent statements, and which may be suggestive of an upcoming cyber-attack.

6.2. Topic and Entity-Focused Sentiment Indicators

At the organizational level, sentiment cues must deal with brand/asset-focused sentiment. This is tracking bad feelings targeted at domains, assets, or key leaders of any company. Particular statements of such com-ponents as index, landing, or CMS used with negative tones may indicate the risk of targeted-defacement. The presence of hostility in a campaign narrative can be measured through topic model- or topic -clustering methods to identify the emergence of patterns of pillaging on vocabularies related to defacement.

When hostile discourses converge to form more coherent themes, the chances of attack grow. As well, geo/sector spillover is likely to be evident, whereby negative sentiment patterns in other adjacent sectors/regions tend to be common in copycat defacement instances. Baseline monitoring of peers would enable organizetions to detect the spread of sentiment across entities, or cross-entity contagion, where any improvement in sentiment in one entity affects the other, increasing the vulnerability to similar attacks.

6.3. Behavioral and Technical Indicators for Detection

Amplifiers that involve measurements of behavior, including velocity and accel-

eration, are early indication of an attack. Using the rate of change in negative sentiment can bring to attention abrupt spikes, which can be associated with threats. Term Frequency-Inverse Document Frequency (TF-IDF) used to detect novelties or burstiness of terminologies that such an attack is imminent through the detection of rare and emergent defacement terminologies. Likewise, the weighting of negativity by influencers is important because small and significant groups of antagonistic messages by high-profile accounts can serve as precursors of cyber incidents. Tracking a coordinated behavior like synchronized messages of aggression on freshly registered accounts or repeated templates can ensure a mobilizational of cyber attackers.

6.4. Stacked Model Performance

The study developed a stacked, multi-view ensemble that integrates complementary evidence streams lexical polarity, contextual semantics and temporal dynamics into a single calibrated risk score for cyber-defacement prediction. Three base learners captured distinct facets of the signal: a TF-IDF-based Multinomial Naïve Bayes model extracted sparse but discriminative lexical cues; a bidirectional LSTM modelled long-range dependencies and pragmatic features such as negation and sarcasm; and ARIMA prior estimated daily baseline propensities reflecting exogenous rhythms in online discourse.

Their probabilistic outputs were combined by a logistic meta-learner trained on time-ordered splits to prevent information leakage and to match deployment conditions. This architecture achieved superior performance over its constituents (Accuracy = 0.8568, F1 = 0.8055, ROC-AUC = 0.9116, PR-AUC = 0.8964, Brier = 0.1151), with particularly improved calibration and ranking ability, thereby enabling cost-aware, real-time triage of high-risk items. By unifying fast lexical scoring, contextual sequence encoding and lightweight temporal priors within a principled stacking framework, the model delivered both discriminative power and operational reliability, satisfying the study's objective of producing a robust predictor of defacement-related risk.

7. Conclusion

The analysis proves conclusively, that the negative sentiment as a whole on the X platform cannot be used as a predictive tool to indicate attack activities of cyber defacement attacks. Unsurprisingly, there is the closest possible correlation a weak contemporaneous relationship where by sentiment and mention of attacks increases and decreases together due to common external drivers, as opposed to sentiment driving subsequent attacks. Such a finding counters the presumption of the original hypothesis and reaffirms that the intentions behind defacement are more intricate than a byproduct of ambient negativity online. It finds that although sentiment analysis may enable important contextual insights into the cyber threat environment, it is relatively unhelpful in predicting clearly specific defacement.

8. Recommendations

1) Focus of the Shift in Monitoring: The cybersecurity teams must shift the focus of monitoring to targeted sentiment. It means focusing on negative emotion (anger, contempt) targeted directly at the brands, assets and the leadership of the organization as this is a more direct risk indicator.

2) Develop a Multi-Layered Indicator System: A Defacement Early Warning Score (DEWS) must be developed which uses more than one indicator. This mechanism would be required to merge selective sentiment and behavioral analytics like message velocity, influencer amplification, technical data like mentions of specific vulnerabilities or tools, and indicators of coordinated inauthentic activity in order to arrive at a more accurate risk calculation.

9. Limitations

Our findings are limited by (i) reliance on publicly available posts from one platform (X), restricting generalizability to others; (ii) potential residual bias despite resampling; and (iii) a synthetic evaluation setting rather than live deployment. Future research should validate on real-time streams, expand to multimodal inputs, and compare subgroup behaviours such as gender, and institution type.

10. Future Works

Future studies ought to be oriented towards making the processes more fine-tuned and entrenched. A major avenue of exploration is the entity-specific sentiment analysis, which features measurement of hostilities against a specific entity such as an organization or country as opposed to measuring the situation in general.

Acknowledgements

The authors gratefully acknowledge the guidance of Dr. Shem Mbandu Angolo and Dr. Casper Shikali throughout this study, and extends appreciation to colleagues and family for their support during the research process.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sleem, A. (2022) A Comprehensive Study of Cybersecurity Threats and Countermeasures: Strategies for Mitigating Risks in the Digital Age. *Journal of Cybersecurity and Information Management*, **10**, 35-46. https://doi.org/10.54216/jcim.100204

[2] Finnemore, M. and Hollis, D.B. (2016) Constructing Norms for Global Cybersecurity. *American Journal of International Law*, **110**, 425-479. https://doi.org/10.1017/s0002930000016894

[3] Abushark, Y.B., Irshad Khan, A., Alsolami, F., Almalawi, A., Mottahir Alam, M.,

Agrawal, A., *et al.* (2022) Cyber Security Analysis and Evaluation for Intrusion Detection Systems. *Computers*, *Materials & Continua*, **72**, 1765-1783. https://doi.org/10.32604/cmc.2022.025604

[4] Meland, P., Tokas, S., Erdogan, G., Bernsmed, K. and Omerovic, A. (2021) A Systematic Mapping Study on Cyber Security Indicator Data. *Electronics*, **10**, Article 1092. https://doi.org/10.3390/electronics10091092

[5] Achuthan, K., Khobragade, S. and Kowalski, R. (2025) Public Sentiment and Engagement on Cybersecurity: Insights from Reddit Discussions. *Computers in Human Behavior Reports*, **17**, Article ID: 100573. https://doi.org/10.1016/j.chbr.2024.100573

[6] Budimir, S., Fontaine, J.R., Huijts, N.M., Haans, A., Loukas, G. and Roesch, E.B. (2021) Emotional Reactions to Cybersecurity Breach Situations: Scenario-Based Survey Study. *Journal of Medical Internet Research*, **23**, e24879. https://doi.org/10.2196/24879

[7] Koc-Michalska, K., Klinger, U., Bennett, L. and Rommele, A. (2024) Dissonant Public Spheres. Routledge. https://doi.org/10.4324/9781003479598

[8] Al-Khater, W.A., Al-Maadeed, S., Ahmed, A.A., Sadiq, A.S. and Khan, M.K. (2020) Comprehensive Review of Cybercrime Detection Techniques. *IEEE Access*, **8**, 137293-137311. https://doi.org/10.1109/access.2020.3011259

[9] Albalawi, M., Aloufi, R., Alamrani, N., Albalawi, N., Aljaedi, A. and Alharbi, A.R. (2022) Website Defacement Detection and Monitoring Methods: A Review. *Electronics*, **11**, Article 3573. https://doi.org/10.3390/electronics11213573

[10] Almomani, O., Alsaaidah, A., Abu-Shareha, A.A., Alzaqebah, A., Amin Almaiah, M. and Shambour, Q. (2025) Enhance URL Defacement Attack Detection Using Particle Swarm Optimization and Machine Learning. *Journal of Computational and Cognitive Engineering*, **4**, 296-308. https://doi.org/10.47852/bonviewjcce52024668

[11] Dau, H.X., Trang, N.T.T. and Hung, N.T. (2022) A Survey of Tools and Techniques for Web Attack Detection. *Journal of Science and Technology on Information security*, **1**, 109-118. https://doi.org/10.54654/isj.v1i15.852

[12] Harinahalli Lokesh, G. and BoreGowda, G. (2020) Phishing Website Detection Based on Effective Machine Learning Approach. *Journal of Cyber Security Technology*, **5**, 1-14. https://doi.org/10.1080/23742917.2020.1813396

[13] Mao, B. and Bagolibe, K.D. (2019) A Contribution to Detect and Prevent a Website Defacement. 2019 *International Conference on Cyberworlds* (*CW*), Kyoto, 2-4 October 2019, 344-347. https://doi.org/10.1109/cw.2019.00062

[14] Hou, W., Cui, B. and Li, R. (2021) A Survey on Blockchain Data Analysis. 2021 IEEE 45*th Annual Computers*, *Software*, *and Applications Conference* (*COMPSAC*), Madrid, 12-16 July 2021, 357-365. https://doi.org/10.1109/compsac51774.2021.00058

[15] Costa, M. (2025) Exploring Hacktivism: The Role and Impact of Social Media. *ARIS*2—*Advanced Research on Information Systems Security*, **5**, 99-111. https://doi.org/10.56394/aris2.v5i1.56

[16] Surya, D., Setiawan, D., Anni Aryani, Y. and Arifin, T. (2024) Cyberattacks on the Accounting Profession: A Literatur Review. *Media Riset Akuntansi*, *Auditing & Informasi*, **24**, 255-272. https://doi.org/10.25105/v24i2.19953

[17] Jony, A.I. and Hamim, S.A. (2023) Navigating the Cyber Threat Landscape: A Comprehensive Analysis of Attacks and Security in the Digital Age. *Journal of Information Technology and Cyber Security*, **1**, 53-67. https://doi.org/10.30996/jitcs.9715

[18] Hoang, X.D. (2018) A Website Defacement Detection Method Based on Machine Learning. In: Fujita, H., Nguyen, D., Vu, N., Banh, T. and Puta, H., Eds., *Advances in*

Engineering Research and Application, Springer, 116-124. https://doi.org/10.1007/978-3-030-04792-4_17

[19] Madhu, K.S., Reddy, B.C., Damarukanadhan, C., Polireddy, M. and Ravinder, N. (2021) Real Time Sentimental Analysis on Twitter. 2021 6*th International Conference on Inventive Computation Technologies* (*ICICT*), Coimbatore, 20-22 January 2021, 1030-1034. https://doi.org/10.1109/icict50816.2021.9358772

[20] Rizwan, M.M. and Devis, J. (2023) Analyzing Product Reviews to understand Customer Sentiment: A Machine Learning Approach. *Proceedings of the National Conference on Emerging Computer Applications* (*NCECA*), **5**, 255-260. https://doi.org/10.5281/zenodo.7955846

[21] Jouini, M. and Arfa Rabai, L.B. (2020) Towards New Quantitative Cybersecurity Risk Analysis Models for Information Systems: A Cloud Computing Case Study. In: Gupta, B., Perez, G., Agrawal, D. and Gupta, D., Eds., *Handbook of Computer Networks and Cyber Security*, Springer, 63-90. https://doi.org/10.1007/978-3-030-22277-2_3

[22] Stine, R.A. (2019) Sentiment Analysis. *Annual Review of Statistics and Its Application*, **6**, 287-308. https://doi.org/10.1146/annurev-statistics-030718-105242

[23] Othman, R., Abdelsadek, Y., Chelghoum, K., Kacem, I. and Faiz, R. (2019) Improving Sentiment Analysis in Twitter Using Sentiment Specific Word Embeddings. 2019 10*th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems*: *Technology and Applications* (*IDAACS*), Metz, 18-21 September 2019, 854-858. https://doi.org/10.1109/idaacs.2019.8924403

[24] Singh, N. and Jaiswal, U.C. (2023) Sentiment Analysis Using Machine Learning. *ADCAIJ*: *Advances in Distributed Computing and Artificial Intelligence Journal*, **12**, e26785. https://doi.org/10.14201/adcaij.26785

[25] Davanzo, G., Medvet, E. and Bartoli, A. (n.d.) A Comparative Study of Anomaly Detection Techniques in Web Site Defacement Detection. In: Jajodia, S., Samarati, P. and Cimato, S., Eds., *Proceedings of The Ifip Tc* 11 23*rd International Information Security Conference*, Springer, 711-716. https://doi.org/10.1007/978-0-387-09699-5_50

[26] Xu, Q.A., Chang, V. and Jayne, C. (2022) A Systematic Review of Social Media-Based Sentiment Analysis: Emerging Trends and Challenges. *Decision Analytics Journal*, **3**, Article ID: 100073. https://doi.org/10.1016/j.dajour.2022.100073

[27] Jamil, M.L., Pais, S. and Cordeiro, J. (2022) Detection of Dangerous Events on Social Media: A Critical Review. *Social Network Analysis and Mining*, **12**, Article No. 154. https://doi.org/10.1007/s13278-022-00980-y

[28] Zimmer, F., Scheibe, K., Stock, M. and Stock, W.G. (2019) Fake News in Social Media: Bad Algorithms or Biased Users? *Journal of Information Science Theory and Practice*, **7**, 40-53. https://doi.org/10.1633/JISTaP.2019.7.2.4

[29] Yue, L., Chen, W., Li, X., Zuo, W. and Yin, M. (2018) A Survey of Sentiment Analysis in Social Media. *Knowledge and Information Systems*, **60**, 617-663. https://doi.org/10.1007/s10115-018-1236-4

[30] Kumar, A. and Sharma, I. (2023) Performance Evaluation of Machine Learning Algorithms for Website Defacement Attack Detection. 2023 *International Conference on Smart Systems for applications in Electrical Sciences* (*ICSSES*), Tumakuru, 7-8 July 2023, 1-6. https://doi.org/10.1109/icsses58299.2023.10201194

[31] Conteh, N.Y. and Schmick, P.J. (2016) Cybersecurity: Risks, Vulnerabilities and Countermeasures to Prevent Social Engineering Attacks. *International Journal of Advanced Computer Research*, **6**, 31-38. https://doi.org/10.19101/ijacr.2016.623006

[32] Khouzani, M., Liu, Z. and Malacaria, P. (2019) Scalable Min-Max Multi-Objective Cyber-Security Optimisation over Probabilistic Attack Graphs. *European Journal of Operational Research*, **278**, 894-903. https://doi.org/10.1016/j.ejor.2019.04.035

[33] Chen, Y., Cui, M., Wang, D., Cao, Y., Yang, P., Jiang, B., *et al.* (2024) A Survey of Large Language Models for Cyber Threat Detection. *Computers & Security*, **145**, Article ID: 104016. https://doi.org/10.1016/j.cose.2024.104016

[34] Hutchings, A. and Collier, B. (2019) Inside Out: Characterising Cybercrimes Committed Inside and Outside the Workplace. 2019 *IEEE European Symposium on Security and Privacy Workshops* (*EuroS&PW*), Stockholm, 17-19 June 2019, 481-490. https://doi.org/10.1109/eurospw.2019.00060

[35] Shu, K., Sliva, A., Sampson, J. and Liu, H. (2018) Understanding Cyber Attack Behaviors with Sentiment Information on Social Media. In: Thomson, R., Dancy, C., Hyder, A. and Bisgin, H., Eds., *Social, Cultural, and Behavioral Modeling*, Springer, 377-388. https://doi.org/10.1007/978-3-319-93372-6_41

[36] Arief, M. and Samsudin, N.A. (2023) Hybrid Approach with VADER and Multinomial Logistic Regression for Multiclass Sentiment Analysis in Online Customer Review. *International Journal of Advanced Computer Science and Applications*, **14**. https://doi.org/10.14569/ijacsa.2023.0141232

[37] Batailler, C., Brannon, S.M., Teas, P.E. and Gawronski, B. (2021) A Signal Detection Approach to Understanding the Identification of Fake News. *Perspectives on Psychological Science*, **17**, 78-98. https://doi.org/10.1177/1745691620986135

[38] Jacobs, A.M. (2019) Sentiment Analysis for Words and Fiction Characters from the Perspective of Computational (Neuro-)Poetics. *Frontiers in Robotics and AI*, **6**, Article 53. https://doi.org/10.3389/frobt.2019.00053

[39] Stieglitz, S., Mirbabaie, M., Ross, B. and Neuberger, C. (2018) Social Media Analytics—Challenges in Topic Discovery, Data Collection, and Data Preparation. *International Journal of Information Management*, **39**, 156-168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

[40] Jacob, N. (2025) X Sentiment Analysis Dataset. Mendeley Data. https://doi.org/10.17632/jmbr7xmrw7.1

[41] Chang, H.H., Chen, E., Zhang, M., Muric, G. and Ferrara, E. (2021) Social Bots and Social Media Manipulation in 2020. In: Engel, U., Quan-Haase, A., Liu, S. and Lyberg, L.E., Eds., *Handbook of Computational Social Science*, *Volume* 1, Routledge, 304-323. https://doi.org/10.4324/9781003024583-21

[42] Bendovschi, A. (2015) Cyber-Attacks—Trends, Patterns and Security Countermeasures. *Procedia Economics and Finance*, **28**, 24-31. https://doi.org/10.1016/s2212-5671(15)01077-1

[43] Li, Y. and Liu, Q. (2021) A Comprehensive Review Study of Cyber-Attacks and Cyber Security; Emerging Trends and Recent Developments. *Energy Reports*, **7**, 8176-8186. https://doi.org/10.1016/j.egyr.2021.08.126

[44] Yang, Y., Zhang, C., Fan, C., Mostafavi, A. and Hu, X. (2020) Towards Fairness-Aware Disaster Informatics: An Interdisciplinary Perspective. *IEEE Access*, **8**, 201040-201054. https://doi.org/10.1109/access.2020.3035714

[45] Achuthan, K., Khobragade, S. and Kowalski, R. (2025) Cybercrime through the Public Lens: A Longitudinal Analysis. *Humanities and Social Sciences Communications*, **12**, Article No. 282. https://doi.org/10.1057/s41599-025-04459-x

[46] Garcia, J. (2025) Beyond the Headlines: Sentiment Divergence and Financial Distress. *Global Finance Journal*, **66**, Article ID: 101126.

https://doi.org/10.1016/j.gfj.2025.101126

[47] Deb, A., Lerman, K. and Ferrara, E. (2018) Predicting Cyber-Events by Leveraging Hacker Sentiment. *Information*, **9**, Article 280. https://doi.org/10.3390/info9110280