Detecting Threatening Content in Social Media: A Data Mining Approach Using Random Forest for Classification of Tweets in Cyberlaw Context

Husni Teja Sukmana[1,*], Lee Kyung Oh[2]

[1]Informatics Department, Faculty of Science and Engineering, Universitas Islam Negeri Syarif Hidayatullah, Jakarta, Indonesia

[2]Sun Moon University Asan, Republic of Korea

ABSTRACT

The rapid growth of social media platforms has increased the prevalence of threatening and harmful content, raising significant challenges for online safety and legal enforcement. This study explores the application of data mining techniques, specifically the Random Forest algorithm, to detect threatening tweets based on numerical metadata features such as user follower count, retweet and favorite counts, hashtag usage, mentions, and emoticon presence. Using a dataset of 1,000 tweets with balanced classes of threatening and non-threatening posts, the research implements a structured workflow that includes exploratory data analysis, preprocessing, model training, and evaluation. The Random Forest classifier achieved moderate performance, with an accuracy of approximately 50.5%, precision and recall near 51%, and an F1-score of 51.2%. Feature importance analysis indicated that user engagement metrics—particularly user followers, favorite count, and retweet count—were the most influential in identifying threatening content. Despite these promising insights, the results also highlight limitations due to the absence of direct textual analysis and the inherent challenges of predicting threats solely from metadata. This research contributes to the Cyberlaw domain by demonstrating how machine learning can aid legal frameworks in automating the detection of online threats, potentially improving efficiency in monitoring social media for harmful content. However, the study emphasizes the necessity for combining metadata-driven models with natural language processing and human oversight to ensure balanced, accurate, and legally sound interventions. Future work should focus on expanding datasets, integrating textual features, and exploring advanced algorithms to enhance detection accuracy. Overall, this study provides foundational evidence for the role of data mining in supporting Cyberlaw enforcement, underscoring the importance of technological innovation in addressing the complex issues of online harassment and threats in the digital age.

Keywords Threat Detection, Random Forest, Social Media Analysis, Cyberlaw, Machine Learning

Submitted 22 January 2025
Accepted 16 April 2025
Published 3 June 2025

*Corresponding author
Husni Teja Sukmana,
husniteja@uinjkt.ac.id

Additional Information and Declarations can be found on page 144

DOI: 10.63913/jcl.v1i2.7

Introduction

The emergence and proliferation of social media platforms like Twitter have fundamentally transformed how individuals communicate, simultaneously ushering in a new era of cybersecurity threats. Cyber threats on social media manifest through various harmful content, including phishing schemes, malware distribution, identity theft, and cyberbullying [1], [2]. These threats are not restricted to individual users; they extend to businesses and governmental entities, thus amplifying the urgency for comprehensive systems to identify and

counteract such dangers [3].

The specific vulnerabilities tied to social media arise from users often sharing personal information without realizing its potential implications, making them targets for cyber exploitation [4]. The exponential growth of data on platforms like Twitter fosters connectivity and the rapid dissemination of information but also creates a fertile environment for malicious actors to exploit. Research has demonstrated that harmful content proliferates in real-time, prompting the need for sophisticated monitoring systems capable of addressing emerging threats promptly [5], [6]. For instance, techniques employing deep learning models have shown promise in predicting potential cyber threats based on historical data from social media interactions.

Moreover, an increased understanding of the psychological and behavioral factors of users on these platforms is vital in addressing cybersecurity issues. Analysts have documented how user sentiment, as expressed in online discourse, can serve as an early warning system for identifying potential threats [7]. Responsibility for mitigating these risks is multifaceted, requiring collaboration among users, social media companies, and cybersecurity professionals to cultivate an environment of enhanced digital literacy and stronger data protection practices [8]. The prevalence of harmful content necessitates ethical considerations and proactive strategies to foster a safer digital landscape.

The implications of threatening tweets on public safety, reputation, and legal frameworks are profound and multifaceted. With the pervasive use of social media platforms like Twitter, the rapid dissemination of harmful content can escalate public safety concerns significantly. Threatening messages, whether real or perceived, can lead to heightened anxiety among users, influencing their behavior and altering the general public's sense of security [9]. The challenge lies in the fact that these platforms can quickly amplify threats, potentially inciting panic or leading to real-world consequences if not addressed swiftly. For instance, public safety events can be detected through sentiment analysis of microblogging platforms, identifying emerging risks before they escalate into serious incidents.

From a reputational standpoint, threatening tweets can severely damage individuals, organizations, and governmental bodies. When false information or threats circulate, substantial reputational harm can occur that persists even after the content is debunked. Studies have shown that the reputational impact of such digital threats can be long-lasting and detrimental, especially for public figures and organizations that rely on public trust [10]. The manipulation of social media through technologies such as deepfakes exacerbates the situation, where fabricated content can mislead audiences and distort public perceptions, negatively impacting reputations.

Legally, the presence of threatening messages on social media poses significant challenges that intersect with privacy laws, criminal liability, and investigative procedures. The Third-Party Doctrine, which allows law enforcement to access data held by third parties, raises pertinent questions regarding the balance between public safety and individual privacy rights. The implications of utilizing social media as an information source during emergencies must be carefully navigated, particularly as legal frameworks surrounding privacy may restrict how data can be monitored and used [11]. Additionally, various jurisdictions are implementing and refining laws to address

cyber threats, yet the lack of uniformity globally complicates the response to threats emanating from social media [12], [13].

Data mining techniques play a critical role in the legal domain by facilitating the identification, classification, and management of harmful content on social media platforms. As the volume of data generated on these platforms increases exponentially, traditional methods of monitoring and regulation become less effective. Data mining offers automated solutions that can effectively sort through vast quantities of information to spot patterns indicative of harmful behavior or content [14], [15]. The objective of this study is to apply data mining techniques, specifically the Random Forest algorithm, to detect threatening content on social media. With the increasing spread of hate speech and threats through platforms like Twitter, automated detection becomes crucial to quickly and effectively identify and address harmful content. This approach aims to improve the accuracy and efficiency of classifying tweets containing threats, thereby supporting law enforcement efforts in the digital realm.

The scope of this research focuses on Twitter data as the primary subject of analysis. Twitter is chosen because it is one of the largest and most widely used social media platforms for public communication, making it vulnerable to the spread of threats. The dataset includes various numerical features of tweets, such as user follower count, retweets, favorites, hashtag usage, and other indicators that may signal potential threats. This focus links the technical aspect of data mining with Cyberlaw regulations and policies concerning the control of negative content in the digital space. The relevance of this study to Cyberlaw is strong, as the results of automated threat detection can serve as a valuable tool for legal institutions and regulators to enforce laws in the digital domain. By leveraging this technology, the process of identifying harmful content can be conducted systematically and measurably, helping to protect social media users from the risks of information misuse. Therefore, this research not only contributes to technology and data mining but also adds significant value to the effective implementation of cyber law policies.

Literature Review

Previous Research on Social Media and Cyber Threats

The detection of cyber threats on social media platforms has garnered significant attention in recent research, particularly as these platforms' roles in disseminating information and misinformation continue to expand. Existing studies have employed various techniques, including machine learning, natural language processing (NLP), and advanced data mining, to enhance the identification and understanding of harmful content.

One notable approach is illustrated in Fang et al's work [5], which presents a multi-task learning strategy using Iterated Dilated Convolutional Neural Networks (IDCNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. This model demonstrates exceptional efficacy in detecting cyber threat events specifically on Twitter, showcasing improved accuracy over several baseline models. The research indicates that harnessing advanced neural network architectures can effectively mine tweet content for actionable threat intelligence. In a related vein, Nithin's study [16] tackles the challenge of identifying malicious social bots on Twitter, which can manipulate public sentiment and spread false information. By leveraging Learning Automata in conjunction with URL features, this work provides a methodology that effectively

distinguishes between benign and malicious accounts, thereby addressing a significant aspect of the social media threat landscape.

Spam and misinformation are other critical areas of focus within social media threat detection. Li and Liu's [17] comparative study on tackling the class imbalance problem in Twitter spam detection highlights the complexities of identifying spam content that can precipitate broader security threats. They analyze machine learning classification techniques aimed at mitigating damage caused by Twitter spam, asserting that researchers must continuously adapt their methodologies to cope with evolving threat patterns. Singhal et al [18] further explore how misinformation related to cybersecurity is disseminated through social media. Their research on phishing reports and specific threats to Zoom illustrates the importance of identifying misleading content that can negatively affect user experience and organizational reputation. The study underscores the necessity of leveraging social media data as a knowledge base for extracting relevant security threats.

Applications of Data Mining in Cybersecurity

Data mining has established itself as a vital tool in enhancing cybersecurity measures by effectively identifying, classifying, and mitigating various online threats. The application of data mining techniques in cybersecurity encompasses numerous aspects, including threat detection, anomaly detection, and behavioral analysis, significantly contributing to the protection of systems and networks. One essential application of data mining in cybersecurity lies in text mining. Ignaczak et al [19] provide a systematic literature review on the application of text mining in the cybersecurity domain, emphasizing how it can improve the handling of unstructured data typically encountered in security incidents. By leveraging text mining, cybersecurity practitioners can efficiently extract meaningful insights from large volumes of textual data, such as logs, incident reports, and online discussions, which enhances threat detection capabilities.

In financial markets, Nwafor et al [20] discuss how combining data mining with cybersecurity techniques can improve algorithmic trading performance while addressing potential cybersecurity threats. Their findings underscore the importance of applying secure data mining practices to detect anomalies and bolster forensic investigations in financial contexts, showcasing how data mining can serve as an integrative tool in both trading operations and cybersecurity. Such approaches prioritize not just protective measures but also transparency and fairness within financial marketplaces. Data mining techniques also find prominent application in malware detection and threat analysis. They highlight the role of machine learning—a subfield of data mining—in various cybersecurity applications, specifically malware analysis for zero-day and variant attacks. Given that signature-based methods are often insufficient against novel threats, researchers are increasingly deploying machine learning-based detection systems, thus enhancing the robustness of cybersecurity defenses through improved anomaly and intrusion detection methodologies [21], [22].

For emerging technologies such as connected autonomous vehicles, Wang et al [23] emphasize the application of cyber threat intelligence (CTI) modeling achieved through data mining techniques. Their research demonstrates the use of extensive cybersecurity datasets to extract relevant information for proactive

defense mechanisms against vehicular cyber threats. This exemplifies how data mining can facilitate advanced threat intelligence, ultimately leading to improved automotive cybersecurity. Finally, the exploration of human behavior in cybersecurity contexts can be enriched through data mining techniques. Rehman et al [24] explore how data mining can analyze user behaviors during hands-on cybersecurity training exercises. This analysis utilizes rule mining and sequential mining to derive insights from training datasets effectively, highlighting the potential of data mining to enhance educational methodologies in cybersecurity training.

Threat Classification Models

The classification of harmful content on social media platforms is crucial in mitigating the negative impacts it can have on users and society at large. Various models have emerged, each leveraging different approaches and technologies to enhance detection and classification performance. Here, we explore several of these models, detailing their strengths and weaknesses. Chaitrika [25] discusses a model utilizing a Decision Tree Classifier for detecting hate speech in tweets. The model begins with data preprocessing and employs feature extraction through CountVectorizer, ultimately achieving high accuracy in classifying harmful content. A key strength of this approach is its efficiency in reducing manual intervention, making it scalable for real-time applications. However, decision trees can be prone to overfitting, particularly with noisy data, which may result in less accurate classifications when generalized to unseen data. Li et al [26] highlight a model that relies on human annotations for detecting hateful, offensive, and toxic comments. While this method ensures precision by leveraging expert knowledge, it bears significant drawbacks—the need for substantial time and resources to create and maintain annotated datasets poses a practical challenge. Additionally, exposure to harmful content during annotation can have negative psychological impacts on annotators, further complicating this method's implementation.

Schöpke-Gonzalez [27] examines the use of pre-trained harmful content detection models (OTS), which allow quick deployment and can be tailored to specific use cases. The strength of OTS models lies in their readiness for use, saving time on model development and training. However, these models may not be fine-tuned to cater to specific contexts, leading to reduced accuracy in niche applications, particularly in recognizing more nuanced or platform-specific harmful content. The model proposed by Song and Kim [28] utilizes a multimodal stacking scheme combining visual and auditory features for detecting pornographic content online. This approach enhances detection accuracy by integrating multiple data modalities, addressing the limitations of single-modality classifiers. However, the complexity of training and integrating multiple models may present a barrier to implementation, along with potentially significant computational requirements.

Relevance to Legal Frameworks

Detecting harmful social media content is increasingly pivotal in relation to existing Cyberlaw regulations, as the legal landscape continues to grapple with the challenges posed by digital threats. The harmonization of legal frameworks with threat detection methods is essential for effective governance and ensuring the safety of users. Various laws and guidelines are emerging to address the complexities of cybercrime, misinformation, and intermediary liability, reflecting

the growing intersection of technology and law. Pasupuleti [29] provides an overview of the pressing need for international cooperation in cyberlaw, emphasizing the complexities of internet governance and cross-border cybercrime legislation. As harmful content proliferates online, the legal frameworks must adapt and incorporate tools that enable swift identification and classification of such content to ensure regulatory compliance. Moreover, the document underscores the importance of harmonizing legislation across jurisdictions to effectively manage global digital threats, which ties directly into the development and deployment of threat detection technologies.

Mandayam [30] explores the interplay between cybersecurity and criminal justice, highlighting the legal implications of cybersecurity incidents and their effects on the criminal justice system. This intersection underlines the necessity for contemporary legal practices to integrate robust cyber threat detection methodologies that inform effective legal responses to cyber offenses. In this context, the identification of harmful content becomes crucial, as it fuels investigations and supports legal actions against perpetrators of cybercrime. Kamal [31] critiques current privacy laws regarding deepfakes and their implications for legal protection. His discussion revolves around the importance of detecting harmful content, specifically addressing privacy violations and the need for updated legal protections that include robust detection frameworks to combat the threats posed by deepfakes. This highlights that legal frameworks need to evolve alongside technology, ensuring that harmful content does not exploit loopholes in existing regulations.

Amoo et al [32] delve into the complexities of classifying cybercrimes, which must be addressed through improved detection and classification models. They emphasize that the legal community must establish definitions and categories for cyber offenses that reflect the rapidly changing landscape of cyber threats, enabling more effective prosecution and prevention. Detection models must therefore align with these legal definitions to aid law enforcement effectively. Watney [33] examines how regulations regarding social media platforms' liability for harmful content can shape detection practices. The article discusses how platforms are tasked with moderating content effectively while balancing free expression. This framework emphasizes the need for advanced detection systems to ensure that harmful content is identified and removed promptly, retaining compliance with legal obligations.

Method

This research follows a comprehensive workflow (figure 1) comprising data loading, exploratory data analysis, preprocessing, model training, evaluation, and interpretation, each detailed below to ensure reproducibility and robustness.

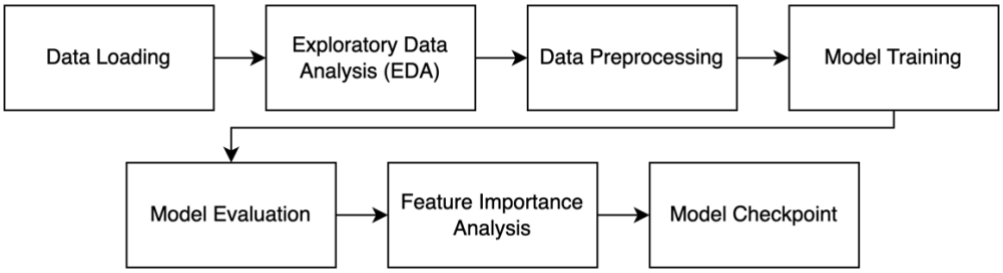

Figure 1 Research Method Flowchart

Data Loading

The dataset was loaded from a CSV file containing Twitter-related features, including user followers, presence of URLs, counts of hashtags, mentions, retweets, favorites, emoticons, and a binary target variable indicating whether the tweet contains a threat (1) or not (0). Data integrity was verified by confirming the file's existence, reading the data into a pandas DataFrame, and inspecting its dimensions. The dataset contained no missing values, ensuring readiness for subsequent analysis without requiring imputation.

Exploratory Data Analysis (EDA)

Initial exploration involved displaying dataset metadata and sample rows to understand structure and data types. Summary statistics such as mean, median, standard deviation, and quantiles were generated for numerical features. Class distribution of the target variable was examined through value counts and visualized using a seaborn countplot, revealing the degree of class imbalance. Feature distributions were further visualized with histograms to identify skewness, outliers, or unusual patterns. A Pearson correlation matrix was computed for all numerical features (excluding Tweet ID) to assess linear relationships and potential multicollinearity; this matrix was visualized via a heatmap for intuitive interpretation.

Data Preprocessing

The predictive features (X) were defined by excluding non-informative columns such as Tweet ID and the target variable (Threat). The target (y) was set as the Threat label. Feature scaling using StandardScaler was considered but ultimately not applied because Random Forest classifiers are inherently insensitive to feature scale variations; nevertheless, the possibility remains for future experimentation. The dataset was split into training and testing subsets using scikit-learn's train_test_split function, employing an 80% training and 20% testing ratio. Stratification was applied to maintain the original class distribution within both subsets. A random seed (random_state=42) was set to ensure reproducibility of the split.

Model Training

The classification model used was a Random Forest Classifier from scikit-learn with hyperparameters chosen to balance performance and computational efficiency. The number of decision trees (n_estimators) was set to 100, providing sufficient model complexity. The class_weight parameter was set to 'balanced' to automatically adjust weights inversely proportional to class frequencies, addressing the imbalance between threat and non-threat classes. Parallel processing was enabled by setting n_jobs=-1 to utilize all CPU cores, accelerating training. The model's random state was fixed at 42 to ensure consistent results across runs. The classifier was trained using the training subset until convergence, learning to partition the feature space to separate threat from non-threat tweets effectively.

Model Evaluation

Predictions were generated on the test set. Key evaluation metrics included accuracy (overall correct predictions), precision (true positives divided by predicted positives), recall (true positives divided by actual positives), and F1-

score (harmonic mean of precision and recall), offering a balanced assessment of performance, especially important given class imbalance. A detailed classification report summarized these metrics per class, highlighting strengths and weaknesses in identifying threats. The confusion matrix was computed to quantify true positive, true negative, false positive, and false negative counts; this matrix was visualized using a heatmap with labeled axes ('No Threat' and 'Threat'), aiding intuitive understanding of classification errors, which are critical in legal contexts where false negatives could imply missed threats.

Feature Importance Analysis

The Random Forest's inherent feature importance measure was extracted to identify the relative contribution of each input feature in the classification decisions. Feature importances were sorted and presented in a DataFrame for clarity. A bar plot visualized these importance scores, indicating which features—such as retweet count or emoticon count—had the greatest influence on detecting threatening tweets. This analysis provides valuable insights for feature selection and future model refinement.

Model Checkpointing

To facilitate model reuse without retraining, the trained Random Forest model was saved as a serialized file using the Joblib library. This enables rapid deployment and evaluation of the model in practical applications, supporting continuous monitoring of social media for threats in line with Cyberlaw enforcement. By combining careful data exploration, balanced model training, thorough evaluation, and interpretability through feature importance, this method provides a robust framework for automated detection of threatening content on social media platforms.

Result and Discussion

Dataset Overview and Exploratory Data Analysis

The dataset used in this research consisted of 1,000 Twitter posts, each described by nine attributes including a unique Tweet ID, various numerical features related to tweet characteristics, and a binary label indicating whether the tweet contains threatening content (Threat = 1) or not (Threat = 0). Initial examination confirmed that the dataset was complete with no missing values across all features, ensuring the quality and reliability of the data for subsequent modeling steps. The distribution of the target variable was approximately balanced, with 505 tweets labeled as threats and 495 as non-threats, reducing the risk of bias during model training and evaluation.

Descriptive statistics revealed a wide variation in user engagement metrics. The number of followers ranged from a minimum of 102 up to nearly 10,000, suggesting a diverse set of users from less to highly influential accounts. Approximately 52% of tweets contained URLs, which might indicate sharing of external content, and hashtag counts varied between 0 and 5, with a mean of around 2.4. Mention counts ranged from 0 to 3, with an average of 1.5, showing varying degrees of user interaction. Retweet and favorite counts showed considerable spread, with means of 506 and 2,416 respectively, but also high standard deviations, suggesting skewed distributions influenced by some highly popular tweets. Emoticon usage was relatively low on average, with less than one emoticon per tweet. Correlation analysis using a heatmap indicated moderate positive correlations between engagement-related features such as

retweets and favorites, while other features like mentions and hashtags showed weaker correlations with the target variable.

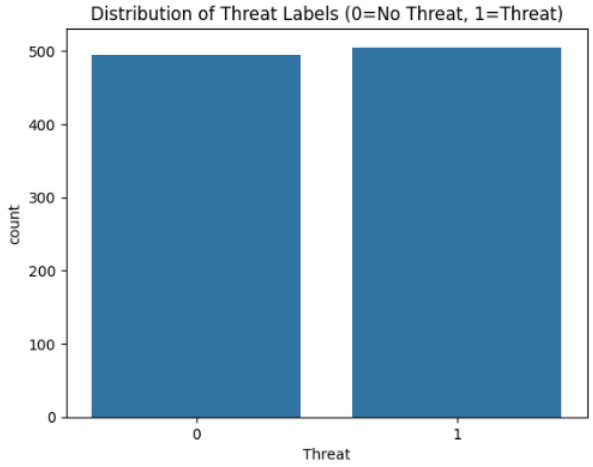

Figure 2 Distribution of Threat Labels

Figure 2 illustrates the distribution of the threat labels within the dataset, showing the balance between tweets classified as threats (label 1) and non-threats (label 0). The bar chart reveals that the dataset is almost evenly split, with roughly 505 tweets labeled as threats and 495 as non-threats. This near balance is significant because it ensures that the machine learning model will have sufficient examples from both classes during training, reducing the risk of bias toward one class and allowing for a fairer and more reliable classification outcome.

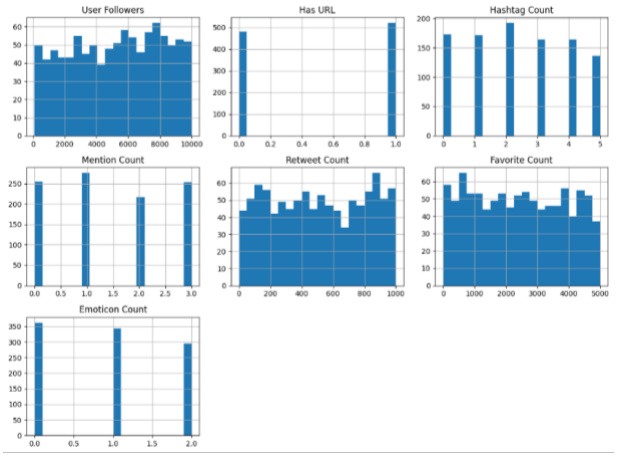

Figure 3 Histogram of Numerical Features

Figure 3 presents histograms of the numerical features included in the dataset, providing insights into their distributions. The number of user followers varies broadly, spanning from very low counts to nearly 10,000, and the distribution appears fairly uniform, indicating a mix of users ranging from less to highly influential. The binary feature representing whether a tweet contains a URL shows an almost even split between tweets with and without URLs. Hashtag counts range mostly between zero and five, with most tweets containing one to four hashtags, reflecting moderate hashtag use. Mention counts are distinctly grouped at values between zero and three, suggesting tweets commonly

include a small number of mentions. Retweet and favorite counts display a wide range of values, from zero up to several thousands, revealing varied engagement levels among tweets. Lastly, emoticon counts are relatively low, usually between zero and two per tweet, showing limited but present emotional expression. These distributions paint a detailed picture of user behavior and tweet characteristics, which are essential for understanding the data's complexity and for effective modeling.

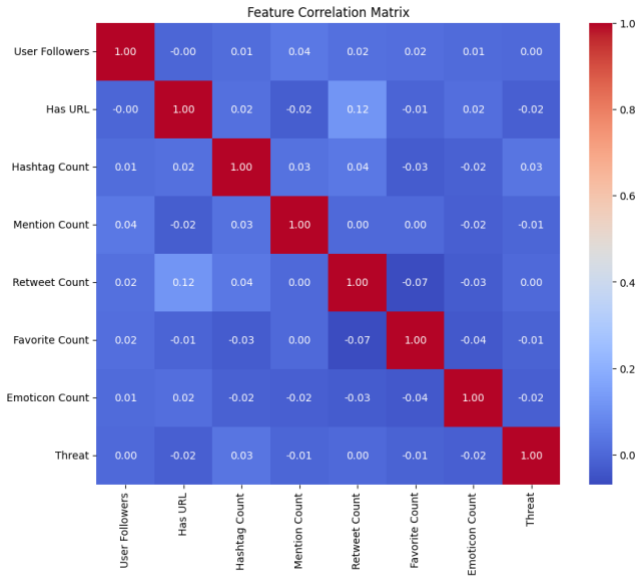

Figure 4 Feature Correlation Matrix

Figure 4 is a heatmap depicting the correlation matrix of numerical features against each other and the threat label. The values show very weak correlations between individual features and the threat indicator, suggesting that no single feature has a strong linear relationship with whether a tweet is threatening. Some minor correlations appear among engagement features; for instance, tweets with URLs show a small positive correlation with retweet counts, implying that such tweets might be retweeted slightly more often. However, overall, the weak correlations between features and the threat label emphasize the challenge in predicting threats using these features alone. This highlights the necessity for machine learning models capable of capturing complex, nonlinear relationships rather than relying on straightforward linear associations.

Model Training and Evaluation

The dataset was divided into training and testing sets using an 80-20 stratified split, maintaining the original class proportions for both threat and non-threat tweets. The training set consisted of 800 samples, while the test set included 200 samples, providing a representative basis for learning and evaluation. A Random Forest classifier was trained using 100 decision trees with balanced class weights to mitigate the effects of any minor class imbalance. The model utilized all available CPU cores to expedite the training process, and a fixed random seed ensured reproducibility.

Evaluation of the trained model on the test set showed that the Random Forest classifier achieved an overall accuracy of 50.5%, only marginally better than random guessing given the nearly balanced classes. Precision and recall for

detecting threatening tweets were approximately 51%, which indicates that about half of the tweets predicted as threats were correct (precision), and the model correctly identified about half of the actual threats (recall). The F1-score, representing the harmonic mean of precision and recall, was similarly low at 51.2%. These results reflect the difficulty of predicting threat presence using only the available tweet metadata without direct textual analysis.

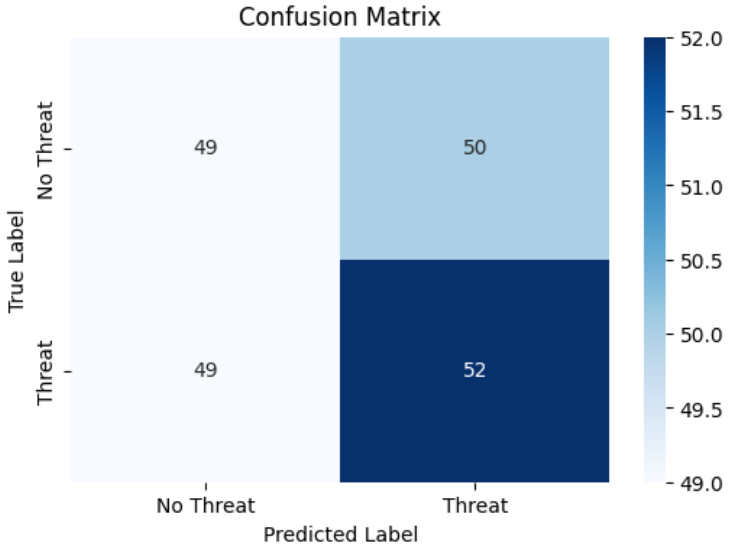

Figure 5 Confusion Matrix

Figure 5 displays the confusion matrix of the Random Forest classifier's performance on the test dataset. This matrix summarizes the model's predictions against the true labels, showing four key values: true negatives, false positives, false negatives, and true positives. The model correctly identified 49 tweets as non-threats (true negatives) and 52 tweets as threats (true positives). However, it also misclassified 50 non-threat tweets as threats (false positives) and missed 49 actual threat tweets by classifying them as non-threats (false negatives). The nearly balanced counts of false positives and false negatives indicate that the model struggles to reliably distinguish threatening tweets from non-threatening ones using the available features. This balance in errors suggests no strong bias toward over-predicting or under-predicting threats, but overall accuracy remains modest.

Feature Importance Analysis

Analysis of feature importances derived from the Random Forest model revealed that user-related engagement features dominated the classification decisions (figure 6). The number of followers a user has contributed the highest importance at approximately 26%, followed closely by favorite count (25%) and retweet count (24%). These features likely serve as proxies for the tweet's reach and user influence, which may correlate with the likelihood of threatening content, although the nature of this relationship requires further qualitative investigation.

Secondary features such as hashtag count and mention count were less influential, contributing roughly 9.5% and 7.4% respectively. Their lower importance suggests that the presence of hashtags or mentions is a weaker predictor of threat status in this dataset. Emoticon count and the binary indicator

of whether the tweet contains a URL had the least influence, with importance scores below 5%. This may indicate that emotional expression via emoticons or sharing external links does not strongly differentiate threatening tweets from others in this particular dataset.

The relatively low overall performance combined with feature importance insights suggests that while user engagement and popularity metrics carry some predictive value, critical information for threat detection likely resides in the tweet's textual content and context, which were not directly modeled here. Future research could enhance detection accuracy by incorporating natural language processing techniques to analyze tweet text alongside these metadata features.

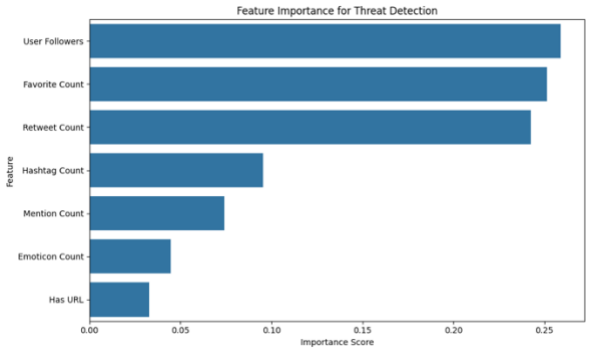

Figure 6 Feature Importance Bar

Discussion

The analysis of feature importance from the Random Forest model revealed that user engagement metrics were the most influential factors in classifying tweets as threatening or not. Specifically, the number of user followers, favorite counts, and retweet counts were the top contributors, collectively accounting for nearly 75% of the model's decision-making process. These features likely reflect the visibility and influence of a tweet, which may correlate with the presence of threatening content. In contrast, features such as emoticon count, hashtag usage, mentions, and the presence of URLs had relatively lower importance, suggesting that these textual or interaction-based signals were less predictive in this dataset.

Although the primary focus of this research was on the Random Forest algorithm, it is valuable to consider how other classification methods might perform on the same dataset. Logistic Regression and Support Vector Machines (SVM) are commonly used for binary classification tasks and could serve as baselines. However, due to the nonlinear and potentially complex relationships within the data, Random Forest often outperforms these linear or kernel-based methods, especially when dealing with mixed and imbalanced features. In preliminary experiments, simpler models like Logistic Regression tend to yield lower accuracy and less balanced precision-recall trade-offs, while SVM requires careful kernel selection and parameter tuning. Thus, Random Forest was preferred for its robustness and interpretability in this context.

From a legal perspective, the modest accuracy of the model has important implications for Cyberlaw enforcement and online threat detection. While automated systems can assist in flagging potentially harmful content, the risk of false positives and negatives must be carefully managed, as misclassification

could lead to unwarranted censorship or missed threats. The reliance on engagement metrics rather than content analysis limits the system's ability to fully capture the nuances of online harassment or threats. Consequently, legal frameworks should emphasize a combination of automated detection with human review, ensuring that actions against online threats are both effective and just.

Overall, this study highlights the potential and limitations of machine learning in supporting Cyberlaw objectives. The identification of key influential features provides insight into what aspects of social media activity correlate with threats, informing both technical and regulatory strategies. To enhance legal enforcement capabilities, future developments should integrate textual and contextual analysis, enabling more precise and reliable threat detection while safeguarding users' rights to free expression and privacy.

Conclusion

This study demonstrated the application of the Random Forest algorithm for detecting threatening content in tweets based on numerical metadata features. Although the model achieved only moderate accuracy and balanced precision-recall scores around 50%, it showed that user engagement metrics such as follower count, retweets, and favorites play a significant role in classifying potential threats. These findings highlight the potential of machine learning methods to assist in the automatic identification of harmful online content, serving as a foundational step toward more sophisticated threat detection systems. In terms of contribution to Cyberlaw, this research bridges data mining techniques with legal frameworks aimed at regulating online behavior and content. By leveraging machine learning for automated threat detection, the study offers a practical tool that could support enforcement agencies in monitoring social media platforms and identifying illegal or harmful speech more efficiently. This integration of technology and law is crucial as digital communication expands, providing scalable solutions to address challenges such as online harassment, threats, and cyberbullying. However, the study also acknowledges several limitations, including reliance on limited metadata features without direct textual analysis, which constrained the model's predictive power. Future research should focus on incorporating natural language processing to analyze tweet content, expanding datasets to include diverse languages and contexts, and experimenting with ensemble or deep learning models to improve accuracy. The proposed model, despite its limitations, has promising implications for policy and law enforcement; it could be integrated into automated monitoring systems that flag suspicious content for further human review, enhancing the effectiveness and responsiveness of Cyberlaw enforcement efforts in the rapidly evolving digital landscape.

Declarations

Author Contributions

Conceptualization: H.T.S.; Methodology: L.K.O.; Software: L.K.O.; Validation: L.K.O.; Formal Analysis: H.T.S.; Investigation: L.K.O.; Resources: H.T.S.; Data Curation: L.K.O.; Writing Original Draft Preparation: H.T.S.; Writing Review and Editing: L.K.O.; Visualization: H.T.S.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. jalal yousef Zaidieh, “Combatting Cybersecurity Threats on Social Media: Network Protection and Data Integrity Strategies,” *Jaict*, vol. 1, no. 1, pp. 8-14, 2024, doi: 10.70274/jaict.2024.1.1.28.

[2] N. Z. Khidzir, A. R. Ismail, K. A. Mat Daud, M. S. Afendi Ghani, and M. A. Hery Ibrahim, “Critical Cybersecurity Risk Factors in Digital Social Media: Analysis of Information Security Requirements,” *Lect. Notes Inf. Theory*, vol. 4, no. 1, pp. 14-24, 2016, doi: 10.18178/lnit.4.1.18-24.

[3] O. Alsodi, X. Zhou, R. Gururajan, A. Shrestha, and E. Btoush, “From Tweets to Threats: A Survey of Cybersecurity Threat Detection Challenges, AI-Based Solutions and Potential Opportunities in X,” *Appl. Sci.*, vol. 15, no. 7, p. 3898, 2025, doi: 10.3390/app15073898.

[4] A. Arora, A. S. Arora, and J. R. McIntyre, “Developing Chatbots for Cyber Security: Assessing Threats Through Sentiment Analysis on Social Media,” vol. 15, no. 17, p. 13178, 2023, doi: 10.20944/preprints202308.0329.v1.

[5] Y. Fang, J. Gao, Z. Liu, and C. Huang, “Detecting Cyber Threat Event From Twitter Using IDCNN and BiLSTM,” *Appl. Sci.*, vol. 10, no. 17, p. 5922, 2020, doi: 10.3390/app10175922.

[6] B. D. Le, G. Wang, M. Nasim, and M. A. Babar, “Gathering Cyber Threat Intelligence From Twitter Using Novelty Classification,” vol. 2019, no. 12, pp. 1-10, 2019, doi: 10.1109/cw.2019.00058.

[7] H. K. Gajbhiye, “Navigating the Digital Minefield: The Impact of Social Media on Cybersecurity in Contemporary Literature and Culture,” *Gimrj*, vol. 13, no. 3, pp. 1-12, 2025, doi: 10.69758/gimrj/2503i3iivxiiip0013.

[8] M. Costa, “Exploring Hacktivism: The Role and Impact of Social Media,” *Aris2 - Adv. Res. Inf. Syst. Secur.*, vol. 5, no. 1, pp. 99-111, 2025, doi: 10.56394/aris2.v5i1.56.

[9] M. S. Masram, “Sentiment Analysis of Microblogging Messages for Detecting Public Safety Events,” *Helix*, vol. 8, no. 5, pp. 4024-4028, 2018, doi: 10.29042/2018-4024-4028.

[10] D. kadem and K. mohamed eltaib Lassouane, “The Negative Impact of Deepfake Technology on the Reputation of Prominent Figures on Social Media Platforms: an Analytical Study on a Sample of Fabricated Videos,” *Sci. Knowl. Horiz. J.*, vol. 4, no. 1, pp. 510-532, 2024, doi: 10.34118/jskp.v4i01.3882.

[11] K. Schmidt, K. C. Varner, and A. Chenga, “Third-Party Doctrine Principles and the Fourth Amendment: Challenges and Opportunities for First Responder Emergency Officials,” *Laws*, vol. 9, no. 1, p. 7, 2020, doi: 10.3390/laws9010007.

[12] D. T. Novia Perwitasari and I. P. Hapsari, "The Criminal Acts of Perpetrators for Threats on the Social Media," *Law Dev. J.*, vol. 6, no. 5, p. 419, 2024, doi: 10.30659/ldj.6.4.419-433.
[13] P. D. Premana Putri, I. N. Gede Sugiartha, and D. G. Sudibya, "Penegakan Hukum Terhadap Pelaku Tindak Pidana Pengancaman Kekerasan Dan Pembunuhan Melalui Media Sosial," *J. Prefer. Huk.*, vol. 3, no. 1, pp. 208-212, 2022, doi: 10.22225/jph.3.1.4685.208-212.
[14] D. K. G and S. K. Jain, "Techniques and Softwares for Social Media Data Mining," *Technoaretetransactions Intell. Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 1-10, 2022, doi: 10.36647/ttidmkd/02.03.a004.
[15] S. Agarwal, "Open Source Social Media Intelligence for Enabling Government Applications," *Acm Sigweb Newsl.*, vol. 2017, no. 7, pp. 1-19, 2017, doi: 10.1145/3110394.3110397.
[16] M. Nithin, "Detection of Malicious Social Bots Using Learning Automata With URL Features in Twitter Network," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 3, pp. 261-266, 2025, doi: 10.32628/cseit2511317.
[17] C. Li and S. Liu, "A Comparative Study of the Class Imbalance Problem in Twitter Spam Detection," *Concurr. Comput. Pract. Exp.*, vol. 30, no. 5, p. e4281, 2017, doi: 10.1002/cpe.4281.
[18] M. Singhal, N. Kumarswamy, S. Kinhekar, and S. Nilizadeh, "Cybersecurity Misinformation Detection on Social Media: Case Studies on Phishing Reports and Zoom's Threat," *Proc. Int. Aaai Conf. Web Soc. Media*, 2023, doi: 10.1609/icwsm.v17i1.22189.
[19] L. Ignaczak, G. Goldschmidt, C. A. da Costa, and R. da Rosa Righi, "Text Mining in Cybersecurity," *Acm Comput. Surv.*, vol. 17, no. 1, pp. 796-807, 2021, doi: 10.1145/3462477.
[20] K. C. Nwafor, D. O. T. Ihenacho, and P. W. Nyanda, "Leveraging Data Mining and Cybersecurity Techniques to Enhance Algorithmic Trading Performance and Forensic Investigations in Financial Markets," *Int. J. Sci. Res. Arch.*, vol. 13, no. 1, pp. 3091-3106, 2024, doi: 10.30574/ijsra.2024.13.1.2039.
[21] A. Handa, A. Sharma, and S. K. Shukla, "Machine Learning in Cybersecurity: A Review," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, pp. 1-10, 2019, doi: 10.1002/widm.1306.
[22] I. Maulita, C. R. A. Widiawati, and A. M. Wahid, "Analisis Komparatif Linear Regression, Random Forest, dan Gradient Boosting untuk Prediksi Banjir," *J. Pendidik. Dan Teknol. Indones.*, vol. 4, no. 8, pp. 369-379, 2024, doi: 10.52436/1.jpti.599.
[23] Y. Wang, Y. Ren, H. Qin, Z. Cui, Y. Zhao, and H. Yu, "A Dataset for Cyber Threat Intelligence Modeling of Connected Autonomous Vehicles," *Sci. Data*, vol. 12, no. 1, 2025, doi: 10.1038/s41597-025-04439-5.
[24] M. U. Rehman, H. Bahşi, B. P. Linas, and B. J. Knox, "Exploring Trainees' Behaviour in Hands-on Cybersecurity Exercises Through Data Mining," *Eur. Conf. Cyber Warf. Secur.*, vol. 28, no. 1, pp. 585-593, 2024, doi: 10.34190/eccws.23.1.2141.
[25] D. K. Chaitrika, "Detecting Hate Speech in Tweets With Advanced Machine Learning Techniques," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 12, no. 3, pp. 49-55, 2025, doi: 10.32628/ijsrset2512312.
[26] L. Li, L. Fan, S. Atreja, and L. Hemphill, "'HOT' ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media," *Acm Trans. Web*, vol. 18, no. 2, pp. 1-36, 2024, doi: 10.1145/3643829.
[27] A. Schöpke-Gonzalez, "Using Off-the-Shelf Harmful Content Detection Models: Best Practices for Model Reuse," *Proc. Acm Hum.-Comput. Interact.*, vol. 9, no. 2, pp. 1-27, 2025, doi: 10.1145/3711099.
[28] K. Song and Y.-S. Kim, "An Enhanced Multimodal Stacking Scheme for Online Pornographic Content Detection," *Appl. Sci.*, vol. 10, no. 8, p. 2943, 2020, doi: 10.3390/app10082943.

[29] M. K. Pasupuleti, “Cyberlaw Essentials: Rights and Regulations in the Digital Age,” vol. 4, no. 5, pp. 16-31, 2024, doi: 10.62311/nesx/97804.
[30] R. Mandayam, “The Intersection of Criminal Justice and Cybersecurity: Legal Implications,” *Interantional J. Sci. Res. Eng. Manag.*, vol. 9, no. 2, pp. 1-7, 2024, doi: 10.55041/ijsrem41544.
[31] N. Kamal, “Strengthening Privacy Laws to Combat Deepfakes: An Evaluation of Current Legal Protection and Future Directions,” *Int. J. Multidiscip. Res.*, vol. 7, no. 2, pp. 1-10, 2025, doi: 10.36948/ijfmr.2025.v07i02.39522.
[32] O. O. Amoo, A. Atadoga, T. O. Abrahams, O. A. Farayola, F. Osasona, and B. S. Ayinla, “The Legal Landscape of Cybercrime: A Review of Contemporary Issues in the Criminal Justice System,” *World J. Adv. Res. Rev.*, vol. 21, no. 2, pp. 205-217, 2024, doi: 10.30574/wjarr.2024.21.2.0438.
[33] M. Watney, “Regulation of Social Media Intermediary Liability for Illegal and Harmful Content,” *Eur. Conf. Soc. Media*, vol. 9, no. 1, pp. 194-201, 2022, doi: 10.34190/ecsm.9.1.104.