DIJEMSS:
Dinasti International Journal of Education Management and Social Science
E-ISSN: 2686-6331
P-ISSN: 2686-6358
https://dinastipub.org/DIJEMSS dinasti.info@gmail.com +62 811 7404 455
DINASTI PUBLISHER
DOI: https://doi.org/10.38035/dijemss.v6i6

The Use of Multiplayer Perceptron Method to Identify Sexual Harassment on Social Media X (Twitter)

Farhan Wildan Giffary[1], Yudi Prayudi[2]
[1]Universitas Islam Indonesia, Yogyakarta, Indonesia, 22917007@students.uii.ac.id
[2]Universitas Islam Indonesia, Yogyakarta, Indonesia, prayudi@uii.ac.id

Corresponding Author: 22917007@students.uii.ac.id[1]

Abstract: The Multilayer Perceptron (MLP) method, as a type of artificial neural network, is used in this study to classify tweets containing sexual harassment elements. This research begins with data collection from social media platform X (Twitter), where tweets that are considered relevant to the topic of sexual harassment are collected for further analysis. This data collection process was carried out by observing the principles of research ethics and maintaining the confidentiality of user identity. Once the data was collected, a text pre-processing stage was performed to ensure that the data used was clean and ready to be processed by the model. This pre-processing includes several important steps such as cleansing, slangword, and stopword removal. The data that has gone through this stage is then weighted using the TF-IDF method, a technique that helps determine the importance of certain words in a set of tweets. The processed data is then analyzed using the MLP algorithm. MLP was chosen due to its superior ability to handle complex and non-linear data. This algorithm is able to detect patterns that indicate the presence of sexual harassment elements in tweets, by classifying based on certain patterns of words, phrases, or contexts that often appear in cases of sexual harassment on social media. This research also uses the NIST Framework to ensure that the entire process of collecting, processing, and analyzing data is carried out in accordance with applicable digital forensic standards. This is important to maintain the validity and legality of the research results, especially if these results are used to support official investigations by the authorities. With the implementation of the MLP method, it is hoped that social media platforms and authorities can be more effective in detecting, preventing, and overcoming cases of sexual harassment in cyberspace.

Keyword: Digital Evidence, NIST, Multilayer Perceptron, X (Twitter).

INTRODUCTION

In recent years, users of social media X (Twitter) have increased. The number of social media X (Twitter) users increased by 71.2% until 2023 with a total of 25.25 million users, ranking 4th in the world. It is not surprising that this increase has occurred, considering how this platform has changed the way people communicate, collaborate, and disseminate information in the digital era. Tens of millions of people regularly engage in various online interactions. One of the most important sources of digital evidence in social media forensic

investigations, such as background checks with the huge volume of data generated. This encourages research into new methods and technologies to investigate and reduce crime and antisocial behavior.

Sexual Harassment is becoming more common as technology advances and more people use social media. Sexual harassment is an act that often occurs in society and hurts its victims. Someone who has experienced sexual harassment by people around them will have psychological trauma and have a negative impact on the formation of their personality. Sexual harassment can be in the form of sexual content, making jokes that lead to sexuality and insults to someone's body parts, and making physical contact in the form of touching or the like. Therefore, early detection of sexual harassment is important to avoid more severe consequences.

The National Institute of Standards and Technology (NIST) Framework is a widely used standard for cybersecurity risk management. This framework outlines methods for identifying, protecting, detecting, responding to, and recovering from cyber threats. NIST has rules on how to handle electronic evidence. These rules are about how to find, collect, secure, view, and report evidence that is found. These rules are still important for handling digital evidence today. This framework can be used to create a comprehensive detection system for cases of sexual harassment.

From the literature obtained, there are many methods and algorithms used for sexual harassment analysis, including in the study conducted by predicting the number of retweets from tweets by comparing 2 text preprocessing techniques, namely bag of words (TFIDF) and word embeddings (Doc2Vec). The results of the preprocessing will be tested using several machine learning techniques such as logistic regression, support vector machine (SVM), random forest, neural network, and multinomial Naive Bayes. In addition to using text, the model created also uses the tweet creation date and the number of likes as additional features in the model. The results of the study showed that the combination of Doc2Vec and random forest pre-processing produced an accuracy of 62.67%. In this study, the researcher proposed the Multilayer Perceptron and TF-IDF algorithms, the use of the Multilayer Perceptron algorithm because it has faster learning capabilities and performance than other algorithms. The Multilayer Perceptron algorithm can be applied to this process by carrying out several stages, including: data collection, preprocessing process, weighting or tf-idf process, creating a model, and training.

METHOD

In this study, several stages were carried out systematically to become one of the guidelines in completing this study, including implementing NIST (National Institute of Standards and Technology) as a framework in Digital Forensics, starting with the collection stage, the collection stage by taking data from X (Twitter) using scraping/crawling, continued with examination, the examination stage examines and searches for digital evidence in the form of text from the results of scraping/crawling acquisition, continued with analysis, the analysis stage analyzes the text evidence obtained to determine sexual harassment using the Multilayer Perceptron method, and the last stage is reporting, the reporting stage makes reports on the results of the analysis of digital evidence and evaluates digital evidence related to sexual harassment cases.

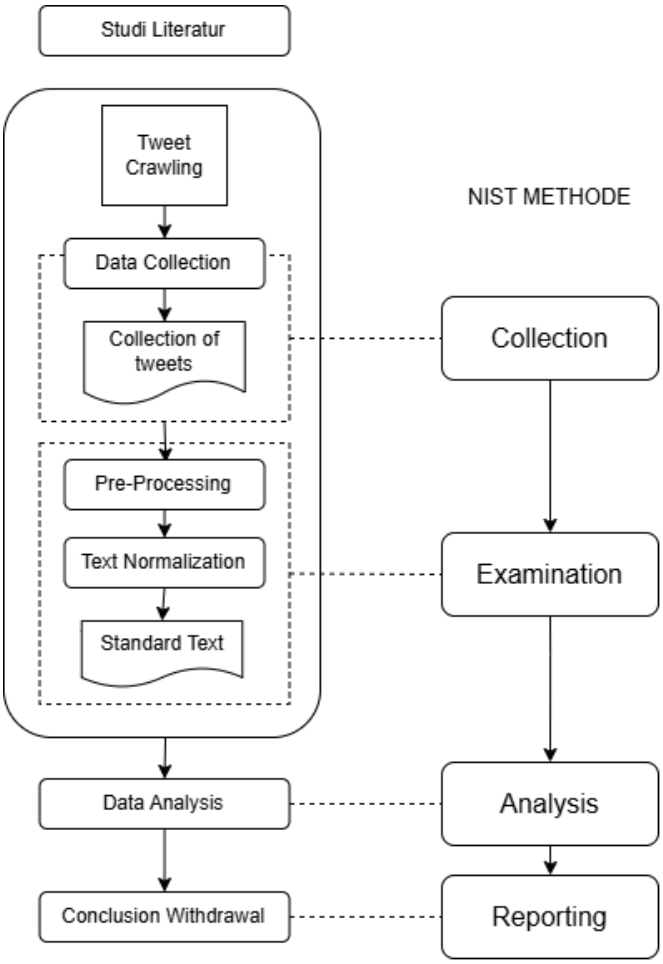

Figure 1. Research Framework

Literature Study

In the literature study stage, researchers conduct exploration and in-depth understanding of literature works that are relevant to the research topic being carried out. The initial step involves searching for literature from academic sources, scientific journals, and related publications. At this stage, attention is focused on theories, concepts, research methodologies, and previous findings that can provide a theoretical basis and context for the research being carried out.

Collection

This stage is part of the identification stage of needs such as collecting data, identifiers, labeling, recording from the source. This data is related to the data integrity maintenance procedure. In this study, data was obtained through a collection of data from X (Twitter) globally using a specific query, namely sexual harassment. The question word is used to obtain data from sentiment about sexual harassment, either through regular tweets or through direct messages on X (Twitter) in Indonesia. Furthermore, observing conversations or text tweets from X (Twitter). At this stage, the researcher uses the crawling method to facilitate the retrieval of text data in X (Twitter) because it can be done automatically without having to copy one by one. The data that has been obtained from the crawling process is then labeled manually, so that it can provide a conclusion of the label on a sentence, in this study the researcher used 1000 tweets from X (Twitter) with a division of 500 given the sexual harassment label and 500 labels not sexual harassment. The data will then be used as a dataset in this study by dividing the data into 80% as training data and 20% as test data. Details of the amount of data in this study can be seen in table 1.

Table 1. Data Details

No	Label	Amount
1	Sexual Harassment	500
2	Not Sexual Harassment	500
	Total	1000

Examination

The examination stage in this study is a step in the examination process, an examination after the digital data is found. Examination is the next stage where an examination is carried out on the social media X (Twitter) which will be extracted and on the crawling data from X (Twitter). The tweet conversation text from X (Twitter) is then subjected to a text pre-processing process with cleansing, slangwords, stopwords removal and stemming. The pre-processing stage in the examination is important where the data that has been collected will be prepared and cleaned to become text that is ready to be analyzed, this makes data that was originally semi-structured or even unstructured become structured data and ready to use.

Analysis

At the analysis stage, it is the stage of analyzing digital evidence that has gone through the examination process and then reprocessed to obtain information related to the case. The analysis is carried out to identify whether there is evidence of sexual harassment in the text evidence of conversations from Twitter direct messages that have been extracted from the previous examination process. This analysis stage is carried out using a text mining process that begins with data labeling, then text processing from the previous process in the examination consisting of cleansing, slangwords, stopwords, and stemming. Then the results are weighted with TF-IDF, and applying Multilayer Perceptron to identify the sexual harassment index in the conversation and tweet data that has been obtained and the accuracy of the Multilayer Perceptron application. The following analysis process is represented in Figure 2.

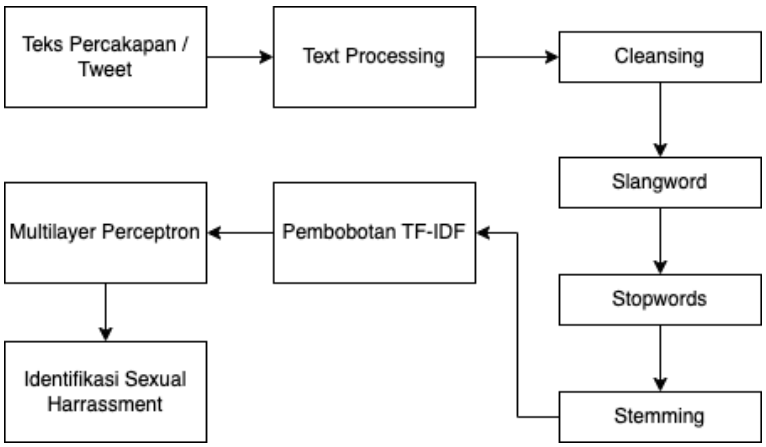

Figure 2. Text Mining Analysis Stages

Reporting

Reporting is the final stage of a series of NIST frameworks after the previous three stages. The results of the analysis of digital evidence and crime cases that have been obtained are then reported at the reporting stage. At this stage, the reporting process and submission of the results of the identification process of conversation texts and tweets suspected of containing elements of sexual harassment on the Twitter social media application are carried out so that the results can be used as valid evidence to be processed in the legal realm, in addition, an evaluation is carried out on what digital evidence is obtained and the performance of the tools used, as well as a discussion in terms of legal review related to cases of sexual harassment on social media.

RESULT AND DISCUSSION

In this study, the results will be discussed from the implementation of the design that has been done previously. The results of this study are divided into according to the application of NIST as the framework used, namely collection, examination, analysis and reporting to process sexual harassment identification.

Collection

At this stage, the researcher carries out the process of identifying important needs such as collecting data, identifying and labeling data. Data collection at the collection stage uses the crawling technique. This technique uses the python programming language. Data from X (Twitter) taken in the form of tweets or tweets that are usually used by someone if they commit sexual harassment on social media or in everyday life, the data in this study collected 1000 tweets or tweets. The following is the crawling result data which can be seen in Figure 3.

text
@tanyakanrl Mainin rambut & kenyot jempol kanan tiap mau tidur sampe sekarang lipetan jempol kanan gue agak mengkerut wkwkw
@kegblgnunfaedh curiga tuh orang bawa sesuatu yang gak bener
@AndyHuskyyy lengan tunggal di bawahnya
@anxiiousz kenyot di kenyot nyot nyot nyot
@acarnanashh Itu makannya di kenyot ya kak ?
@itsMeErniie Bojong soang Bojong kenyot
Aku akan mulai kerjain ntar sore|ø¸ maap lama gusy abis dikenyot kenyot sama realita kehidupan

Figure 3. Data Crawling Results

Then the results of the crawling data will be labeled, labeling is done with initialization 0 and 1, where 0 is a tweet or tweet that does not contain elements of sexual harassment and 1 is a tweet or tweet that contains elements of sexual harassment. The labeling results dataset can be seen in Figure 4

text	label
@tanyakanrl Mainin rambut & kenyot jempol kanan tiap mau tidur sampe sekarang lipetan jempol kanan gue agak mengkerut wkwkw	0
@WindiS64491 Buka semua aj nanti aku kenyot atas bawah	1
@anxiiousz kenyot di kenyot nyot nyot nyot	0
Kuraba tubuhmu dan ku ciummi mmk mu Disaat kamu tertidur kunikmati tubuhmu Dm buat cewe cs #chatsekskasar #chatseks #moancontent #moancewe #desahancewek #desahanenak #hijabsange #sangeberat #SANGE_AAAAAAAAAH	1
@AndyHuskyyy lengan tunggal di bawahnya	0
video str3aming s3x ayam kampus ng3nt0t b0k3p terbaru abg c0lm3k muncrat banyak banget	1

Figure 4. Labeling Result Dataset

Examination

At this stage, the researcher carries out the process of examining the data from the previous stage, where at this stage the researcher pre-processes the previous data. Pre-processing is done using cleansing, slangwords, stopwords and stemming. Pre-processing in the examination is important where the data will be cleaned to become structured text and ready to use. Each stage of pre-processing has complementary functions and justifications.

At the cleansing stage, it is used to remove noise from the dataset to obtain document text with only letter format. The implementation in this study uses a regular expression library or what is commonly called regex which is commonly used to match, search, and manipulate text.

At the slangword stage, the process of changing the writing form of non-standard words into standard ones will be carried out, the use of non-standard abbreviations and contemporary words by overwriting non-standard words into standard words, by defining non-standard words and standard words in a variable with the name replacement into a split token sentence or word and the stopwords stage is the stage of deleting words that do not have important meanings, in its implementation, stopwords data is taken from the Indonesian language corpus in the nltk library and adding several words as a form of adjusting stopwords from the dataset owned. At the stemming stage, the writing form of a word will be changed to its basic form by removing prefix suffix, infix, and konfik affixes, the implementation of stemming uses the Sastrawi library.

Analysis

At this stage, the researcher performs the analysis stage of digital evidence that has been obtained from the previous process, this stage carries out the analysis starting with the TF-IDF weighting process where this process weights words by looking at the level of importance of a word in a document text using the implementation of the countvectorizer and tfidfvectorizer libraries, this process also uses lowercase with a boolean value of false because lowercase has been carried out in the previous stage, the tf-idf process also uses unigrams to convert clean tweets into numeric matrix representations that are ready to be used for the classification analysis stage. After the TF-IDF weighting process, the next step is to carry out the analysis using the Multilayer Perceptron method, where this process uses the utilization of the mlp library from scikit learn which can be used for the identification classification process, the use of mlp in this study uses ReLU activation with the number of hidden layer neurons 100 and a maximum training iteration of 300 and also random state 0 which will be used to create a model from mlp which will then get effective data results and interpretations for identification classification.

Reporting

At this stage, the researcher reports the results of the analysis stage where at that stage tf-idf and multilayer perceptron are applied for identification classification of the data obtained, this method produces good accuracy in the classification process using cross validation and confusion matrix, testing using cross validation in this study uses 5-fold which means the data will be divided into 5 groups and repeated as many as k-fold where the value of k = 5, can be seen in Figure 5 the test results produced using cross validation.

	fit_time	score_time	test_accuracy	test_f1	test_precision	test_recall	fold	features	classifier
0	9.729259	0.018404	0.685	0.695652	0.672897	0.72	1	tf	Multilayer Perceptron
1	11.073628	0.017108	0.700	0.750000	0.642857	0.90	2	tf	Multilayer Perceptron
2	12.537158	0.017251	0.890	0.890000	0.890000	0.89	3	tf	Multilayer Perceptron
3	11.875737	0.016944	0.895	0.897561	0.876190	0.92	4	tf	Multilayer Perceptron
4	11.317334	0.024848	0.745	0.698225	0.855072	0.59	5	tf	Multilayer Perceptron
5	9.796891	0.021464	0.665	0.673171	0.657143	0.69	1	tf-idf	Multilayer Perceptron
6	12.375418	0.016011	0.715	0.753247	0.664122	0.87	2	tf-idf	Multilayer Perceptron
7	12.743796	0.017796	0.845	0.841026	0.863158	0.82	3	tf-idf	Multilayer Perceptron
8	12.654146	0.016455	0.885	0.889952	0.853211	0.93	4	tf-idf	Multilayer Perceptron
9	12.232263	0.017688	0.760	0.717647	0.871429	0.61	5	tf-idf	Multilayer Perceptron

Figure 5. Cross Validation Testing

Testing is also carried out using a confusion matrix to obtain True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) values where these values are obtained

from prediction results using a comparison of 80% training data and 20% random testing data, these results can be seen in Figures 6 and 7.

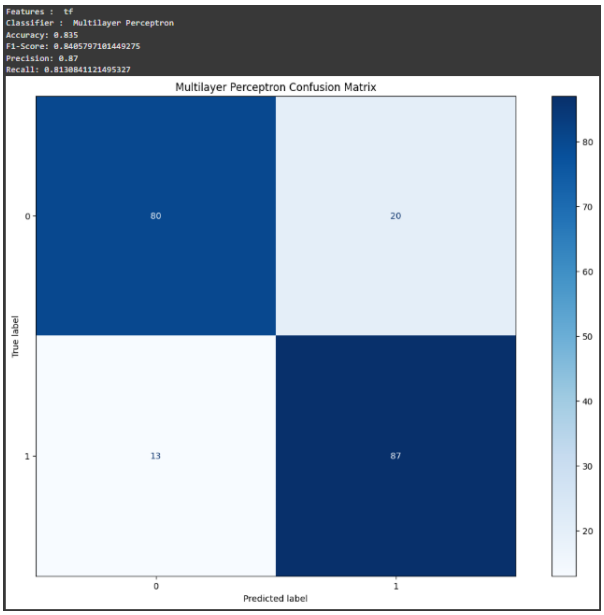

Figure 6. TF – MLP Confusion Matrix Testing

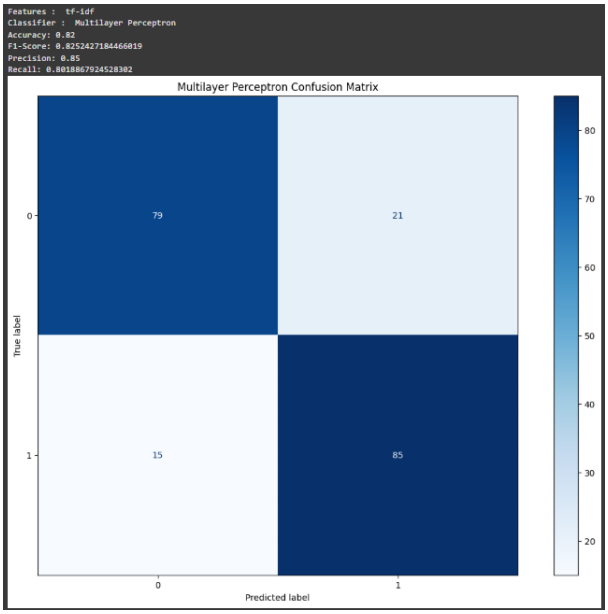

Figure 7. TF-IDF – MLP Confusion Matrix Testing

Through this test, predictions can be made on words or sentences taken from unlabeled data, to find out whether there are indications of elements of sexual harassment in them, this can be used as a reference which is then reported in a digital forensic report, as can be seen in Figure 8 for the prediction results taken from unlabeled data.

	text	prediction
0	@acarnanashh Itu makannya di kenyot ya kak ?	0

Figure 8. Prediction Results

This research can contribute to the development of the investigation process in the world of digital forensics by developing the multilayer perceptron method as an identification of

sexual harassment that occurs on social media and this research can also be a development of science or can be applied in the real world.

Discussion

This study adopts NIST guidelines and approaches in conducting data collection, analysis, data processing and model evaluation for identifying content containing sexual harassment on social media X (Twitter). Referring to NIST standards, the data collection process is carried out systematically to ensure that the data obtained is representative and valid, data is collected using a crawling method with strict filters related to keywords containing elements of sexual harassment while ensuring compliance with research ethics and user privacy, data obtained as much as 1000 data with a division of 500 data containing sexual harassment and 500 data not sexual harassment. The data is then processed to balance the quality and quantity of samples by cleaning text, deleting and normalizing words according to NIST guidelines to ensure consistent data as input to the TF-IDF and Multilayer Perceptron testing models. The testing method to obtain the results in this study uses 5-fold cross validation and confusion matrix to obtain stable and unbiased evaluation results, the evaluation is carried out with the main metrics, namely accuracy, precision recall and f1-score. The cross validation test results show that the tf and multilayer perceptron models get 78% accuracy, 79% precision, 80% recall and 77% f1-score, with the tf-idf and multilayer perceptron models getting 77% accuracy, 78% precision, 78% recall and 77% f1-score, while the test results using confusion matrix show that the tf and multilayer perceptron models get 83% accuracy, 87% precision, 81% recall and 84% f1-score, with the tf-idf and multilayer perceptron models getting 82% accuracy, 85% precision, 80% recall and 82% f1-score. The model created can also predict the presence of sexual harassment elements in tweet data outside the labeled training data, this is in accordance with NIST that the system can operate adaptively to new data for use in the Digital Forensics examination process.

CONCLUSION

Based on the results of the analysis conducted by the researcher, it can be concluded that the use of the multilayer perceptron method in the analysis of digital evidence obtained on social media X (Twitter) for the classification of tweet identification containing elements of sexual harassment with good accuracy and evaluation metrics by combining TF-IDF features in significantly improving performance. This research can also be implemented and can contribute to the development of digital forensic science in the applicable legal process.

Suggestion

Further research can collect data from more tweets than the data used in this study to obtain the maximum level of complexity and in further research can use a combination of other methods in identifying sexual harassment on social media, as well as conducting periodic evaluations to measure performance according to the development of language trends and harassment patterns.

REFERENCE

Mutia, A. (2023, July). *Number of Twitter users in Indonesia ranked 4th in the world as of July 2023*. Databoks. https://databoks.katadata.co.id/datapublish/2023/11/01/jumlah-pengguna-twitter-indonesia-duduki-peringkat-ke-4-dunia-per-juli-2023.

Bokolo, B. G., & Liu, Q. (2024). Artificial intelligence in social media forensics: A comprehensive survey and analysis. *Electronics, 13*(9). https://doi.org/10.3390/electronics13091671.

Budiman, K., Zaatsiyah, N., Niswah, U., Muhanna, F., & Faizi, N. (2020). Analysis of sexual harassment tweet sentiment on Twitter in Indonesia using Naïve Bayes method

through National Institute of Standard and Technology digital forensic acquisition approach. *Journal of Advanced Information Systems and Technology, 2*(2), 21–30. https://journal.unnes.ac.id/sju/index.php/jaist.

Faizal, A., & Luthfi, A. (2024). Comparison study of NIST SP 800-86 and ISO/IEC 27037 standards as a framework for digital forensic evidence analysis. *Journal of Information Systems and Informatics, 6*(2), 701–718. https://doi.org/10.51519/journalisi.v6i2.717.

Daga, I., Gupta, A., Vardhan, R., & Mukherjee, P. (2020). Prediction of likes and retweets using text information retrieval. *Procedia Computer Science, 168*, 123–128. https://doi.org/10.1016/j.procs.2020.02.273.

Prasetya Wibawa, A., Lestar, W., Bella Putra Utama, A., Tri Saputra, I., & Nabila Izdihar, Z. (2020). Multilayer Perceptron untuk prediksi sessions pada sebuah website journal elektronik. *Indonesian Journal of Data Science, 1*(3), 57–67. https://doi.org/10.33096/ijodas.v1i3.15.

Sintia, Defit, S., & Nurcahyo, G. W. (2021). Product codefication accuracy with cosine similarity and weighted term frequency and inverse document frequency (TF-IDF). *Journal of Applied Engineering and Technological Science, 2*(2), 14–21. https://doi.org/10.37385/jaets.v2i2.210.

Umar, R., Riadi, I., & Zamroni, G. M. (2018). Mobile forensic tools evaluation for digital crime investigation. *International Journal of Advanced Science, Engineering and Information Technology, 8*(3), 949–955. https://doi.org/10.18517/ijaseit.8.3.3591.

Nur Faiz, M., Adi Prabowo, W., & Fajar Sidiq, M. (2018). Studi komparasi investigasi digital forensik pada tindak kriminal. *Journal of Informatics, Information System, Software Engineering and Applications, 1*(1), 63–70. https://doi.org/10.20895/INISTA.V1I1.

Bintang, R. A., Umar, R., & Yudhana, A. (2020). Analisis media sosial Facebook Lite dengan tools forensik menggunakan metode NIST. *Techno (Jurnal Fakultas Teknik Universitas Muhammadiyah Purwokerto), 21*(2), 125. https://doi.org/10.30595/techno.v21i2.8494.

Ngejane, C. H., Eloff, J. H. P., Sefara, T. J., & Marivate, V. N. (2021). Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Forensic Science International: Digital Investigation, 36*, 301109. https://doi.org/10.1016/j.fsidi.2021.301109.