JURNAL ECOTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering

Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

Implementation of Natural Language Processing in the Reporting and Handling System of Sexual Violence Cases on Campus

Ilwan Syafrinal[1], Sapta Eka Putra[2], Mahazam Afrad[3]
[1]Universitas Universal, Sungai Panas, Kec. Batam, Batam 29444, Indonesia
[2]Universitas Tamansiswa, Jl. Taman Siswa, Padang 25171, Indonesia
[3]Institut Teknologi Telkom Purwokerto, Jl. DI Panjaitan No.12, Purwokerto 53147, Indonesia

ARTICLE INFO

Article historys:

Received : 02/08/2024
Revised : 22/08/2024
Accepted : 18/09/2024

Keywords:

Natural Language Processing; Reporting Systems; Sexual Violence; Support Vector Machines

ABSTRACT

Sexual violence in the campus environment is a serious problem that requires an effective reporting and handling system. This research aims to develop a Natural Language Processing (NLP)-based system that can improve the process of reporting and handling cases of sexual violence on campus. The methodology used includes the application of NLP techniques such as sentiment analysis and entity recognition to automate the identification and handling of reports. The Support Vector Machines (SVM) algorithm is used for the classification of text in this system. The data is collected from various sources, pre-processed, and used to train NLP models. The results of the study show that the system developed has an accuracy level of 91%, precision of 93%, and recall of 87%, which illustrates its effectiveness in collecting reports of sexual violence anonymously and accurately. Feedback from early adopters shows that the system improves the efficiency and accuracy of the reporting process. The conclusion of this study is that the implementation of NLP can significantly improve the reporting and handling system of sexual violence on campus. Further research is suggested to expand the scope of the system and improve its analysis capabilities.

Corresponding Author:
Ilwan Syafrinal
Universitas Universal, Sungai Panas, Kec. Batam, Batam 29444, Indonesia
Email: ilwansynl@gmail.com

1. INTRODUCTION

This research is motivated by the fact that sexual violence in the educational environment [1-3], especially on campus, has become one of the major sins that threatens the integrity and safety of educational institutions [4]. This problem not only violates human rights but also creates an unsafe environment, hindering the learning process and personal growth of students. Although many cases occur, most go unreported due to the lack of an accessible, anonymous, and sensitive reporting facility to the needs of victims [5]. Therefore, the purpose of this research is to develop a more effective reporting system that addresses these gaps by providing a platform that ensures anonymity and sensitivity to the victim's needs. This research seeks to fill the gap by focusing on the development of reporting mechanisms in educational institutions, particularly in higher education, which have not been adequately addressed in previous studies.

Sexual violence in the campus environment is a serious problem that has a significant impact on the mental and physical health of the victim, and can damage the reputation of educational institutions [6]. According to data from https://databoks.katadata.co.id/, there were many cases of sexual violence, namely 13,156 in 2023. Victims of sexual violence are reluctant to report such incidents due to fear,

JURNAL ECTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

social stigma, and distrust of the existing reporting system [7]. This situation demands a more efficient and reliable reporting system to ensure that all cases of sexual violence are handled appropriately.

Current sexual violence reporting systems are often less effective in identifying and following up on reports [8]. Limitations in analyzing report data quickly and accurately lead to many cases that are not handled properly, leaving victims without adequate support. Therefore, a new approach is needed that can improve accuracy and efficiency in handling reports of sexual violence.

Various studies have been conducted to overcome this problem with various technological approaches, including using text mining techniques to analyze reports of sexual violence, but facing obstacles in handling large volumes of data [9]. Developed a mobile application for reporting sexual violence, but still faced problems in validating report data [10]. Applied the Naive Bayes algorithm for text classification, but the results were less accurate in the context of sexual violence reports [11]. Using Random Forests to improve classification accuracy, but this model still faces challenges in interpreting the results [12]. Lastly, applying deep learning for sentiment analysis, but it requires high computing and is difficult to implement on a large scale [13]. While these approaches make a significant contribution, they still lack accuracy and scalability. This research offers a new approach by combining Natural Language Processing (NLP) and Support Vector Machines (SVM) algorithms to automate the process of reporting and handling cases of sexual violence on campus. SVM was chosen because of its strong ability to handle text classification with high accuracy and good interpretability [14]. This approach is expected to overcome the limitations of previous research and provide a more effective and efficient solution.

The main purpose of this study is to develop and implement a reporting and handling system for sexual violence cases based on NLP and SVM in the campus environment. In addition, the study aims to increase the reporting rate of sexual violence cases, reduce social stigma, and provide better support for victims. This research makes a significant contribution to science by providing a new approach to dealing with sexual violence on campus through NLP and SVM technology. In addition, this research also opens up opportunities for further development in a more sophisticated and user-friendly sexual violence reporting system. Thus, this research is expected to be an important reference for researchers and practitioners in the field of campus security and information technology.

This research seeks to fill the gap by focusing on the development of reporting mechanisms in educational institutions, particularly in higher education, which have not been adequately addressed in previous studies. The problem addressed in this research is the lack of a robust reporting system that can ensure both accessibility and protection for victims, which has led to underreporting and ineffective handling of sexual violence cases. The objective is to create a solution that improves the accuracy and efficiency of reporting, while fostering a safer environment for students.

2. RESEARCH METHOD

Sexual violence on campus is a serious problem that requires special attention. To effectively address this problem, this study uses methodologies based on Natural Language Processing (NLP) and Support Vector Machines (SVM) in developing a system for reporting and handling sexual violence. The diagram below explains the stages of the methodology used in this study. Each stage is designed to ensure that the data is processed correctly and produce an accurate and reliable model.

JURNAL ECOTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

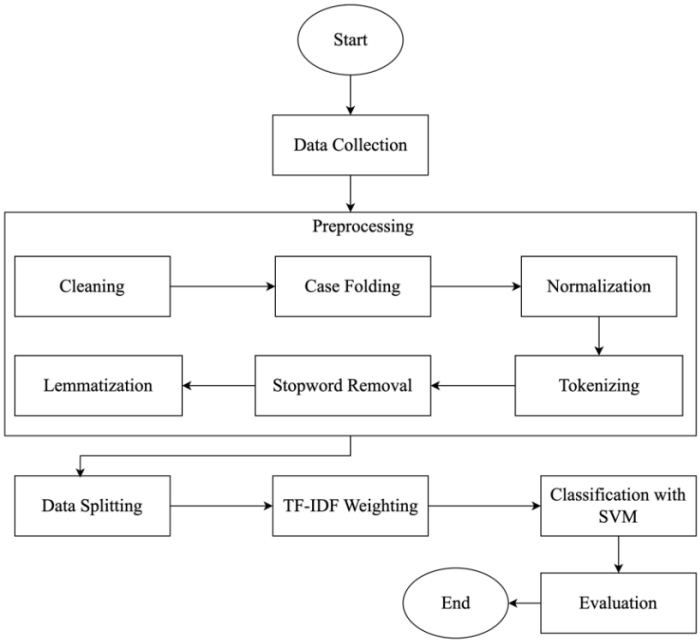

Figure 1. Research methodology

The methodology of this research consists of several systematic stages, ranging from data collection to model evaluation. Each stage has an important role in the process of developing a system for reporting and handling sexual violence. Here is a detailed explanation of each stage:

1. Start, this stage marks the beginning of the research process, where the research objectives and framework are established.
2. Data Collection, Data collection on sexual violence reports is carried out through various sources, such as surveys, anonymous reports, or data from related institutions. The data collected must include the different types of sexual violence that occur on campus to get a comprehensive picture.
3. Preprocessing
    a. Cleaning, the data cleaning process is carried out to remove irrelevant information or noise from the raw data. This includes the removal of unnecessary special characters, numbers, and punctuation.
    b. Case Folding, all text is converted to lowercase letters to ensure consistency in data processing. For example, the words "Violence" and "violence" are considered the same after case folding.
    c. Normalization, Text normalization is done by changing words into standard or standard forms. This can involve replacing abbreviations with full forms or changing non-standard word forms to standard forms.
    d. Tokenizing, the process of tokenization breaks down text into smaller units, such as words or phrases. These tokens are then used in the next stage of analysis.
    e. Stopword Removal, Common words that do not provide significant information (stopwords) such as "and", "which", "with", are removed from the text to reduce noise.
    f. Lemmatization, Lematization transforms words into their basic form or lemma. For example, the words "run," "run," and "run" are all changed to "run."
4. Data Splitting, the processed data was then divided into two sets: a training set of 80% data or 3368 data and a testing set of 20% or 842 data. The training set is used to train the model, while the test set is used to evaluate the model's performance.
5. TF-IDF Weighting, TF-IDF (Term Frequency-Inverse Document Frequency) It is used to give weight to words in text based on their frequency in the document and their relative frequency throughout the document. This weight helps in identifying the most relevant words in the classification.

JURNAL ECOTIPE Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

6. Classification with SVM, Algoritma Support Vector Machines (SVM) used to classify text into appropriate categories. SVM was chosen because of its strong ability to handle text classification problems with high accuracy.

Table 1. Kernel Formula In SVM[15]

Karnel	Formula
Polynomial[16]	k(x, y) = (x. y + c) ^{d}
Sigmoid[17]	$k(x, y)$ = tanh($\gamma x. y + c$)
Linear[18]	$k(x, y) = x. y + c$v
Radial Basis Function (RBF)[19]	$k(x, y)$ = exp ($-\gamma\|x - y\|$ 2)

The SVM, as shown in Table 1, employs various kernel functions (Polynomial, Sigmoid, Linear, and RBF) to convert input data into a higher-dimensional feature space, facilitating the identification of a hyperplane separator. The choice of kernel function and its parameters is influenced by the nature of the data, as each kernel has its own strengths and weaknesses. For instance, the Polynomial kernel elevates a dot product to a specified power and adds a constant, the Sigmoid kernel utilizes a hyperbolic tangent function, the Linear kernel carries out a dot product operation, and the RBF kernel assesses distances using the Gaussian function.

7. Evaluation, the evaluation stage is carried out to assess the performance of the classification model. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure how well the model can identify and classify reports of sexual violence.
   a. Akurasi

$$Accuracy = \frac{\text{TP+FP+TN+FN}}{\text{TP+TN}} \quad (1)$$

In equation (1) it can be explained that, TP (True Positive) is the number of correct positive predictions, TN (True Negative) is the number of correct negative predictions, FP (False Positive) is the number of false positive predictions, FN (False Negative) is the number of false negative predictions. Accuracy provides an overview of how often classification models give correct predictions, but can be misleading if there is a class imbalance[20].

   b. Precision

$$Precision = \frac{TP}{TP+FN} \quad (2)$$

In equation (2) it can be explained that, TP (True Positive) is the number of correct positive predictions, FP (False Positive) is the number of false positive predictions, Precision shows the proportion of correct positive predictions of all positive predictions made. High precision means that the model rarely makes positive mistakes [21].

   c. Recall

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

In equation (3) it can be explained that, TP (True Positive) is the number of correct positive predictions, FN (False Negative) is the number of false negative predictions. The recall shows the proportion of correct positive predictions of all instances that are actually positive. High recall means that the model successfully captures most positive instances [22].

   d. F-1 Score

$$F-1\ Score = 2 * \frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

F1-Score achieves its best score of 1 and worst at 0, providing a balance between precision and recall, especially useful in cases with an unbalanced class distribution [23].

8. End, The final stage of the research process where the results of analysis and evaluation are collected and interpreted to make conclusions and recommendations.

JURNAL ECOTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

With this structured methodology, the research is expected to produce a more effective, efficient, and accurate reporting and handling system for sexual violence in the campus environment.

3. RESULTS AND DISCUSSION

The results of the study will be discussed in depth to understand the implications of the findings, as well as to compare them with previous relevant studies. The analysis will include evaluation metrics such as accuracy, precision, recall, and F1-score to assess the performance of the classification model that has been developed.

3.1. Data Collections

The study used 3000 reports covering different types of sexual and non-sexual violence. The data is categorized into three main types: Non-Sexual Violence Data, Physical Sexual Violence Data, and Non-Physical Sexual Violence Data. Non-Sexual Violence data includes reports such as lost items or non-violent incidents. Physical Sexual Violence data involves reports of physical contact, such as physical harassment and rape. Meanwhile, Non-Physical Sexual Violence Data includes reports such as verbal, visual, and online harassment. Data collection is carried out through surveys, anonymous reports, and data from related institutions. Each report is anonymized to protect the privacy of the reporter and is processed through various stages of preprocessing before being used for training and evaluation of the classification model. The goal is to provide a representative dataset to train an SVM-based classification model that is effective in identifying and handling reports of sexual violence on campus. An example of the data used is shown in Figure 2.

	No	Report	Category
0	1	The individual kept touching my hair and face without permission. I told them to stop, but they did...	Physical Sexual Harassment
1	2	While waiting for the bus, a stranger attempted to grope me. I shouted for help, and they fled the ...	Non-Sexual Harassment
2	3	The perpetrator blocked my path and tried to kiss me. I managed to push them away and run to s...	Non-Sexual Harassment
3	4	The individual followed me to my dorm room and tried to grab me. I managed to escape and lock t...	Physical Sexual Harassment
4	5	The perpetrator touched me without my consent during a campus event. I felt violated and immed...	Physical Sexual Harassment

```
RangeIndex: 4210 entries, 0 to 4209
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   No        4210 non-null   int64
 1   Report    4210 non-null   object
 2   Category  4210 non-null   object
dtypes: int64(1), object(2)
memory usage: 98.8+ KB
```

Figure 2. Data collection results

This DataFrame contains 4,210 entries with three columns, namely "No", "Report", and "Category". The "No" column has an integer data type (int64) that serves as the sequence number or index for each entry. The "Report" column contains a description of the report, while the "Category" column contains the category or label associated with each report. There are no blank values in these three columns, so all entries are complete. The size of the memory used by this DataFrame is about 98.8 KB, indicating that this dataset is light to process.

3.2. Preprocesing

Pre-processing is an important step in data analysis or machine learning workflows, especially when working with text data. The main purpose of pre-processing is to convert the raw data into a clean, structured format that is suitable for analysis. This step involves a series of operations that help improve data quality, reduce noise, and ensure that the resulting features are informative and representative of the underlying patterns in the data.

JURNAL ECTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

Original 'Report' column:

	Report
0	The individual kept touching my hair and face without permission. I told them to stop, but they didn't listen.
1	While waiting for the bus, a stranger attempted to grope me. I shouted for help, and they fled the scene.
2	The perpetrator blocked my path and tried to kiss me. I managed to push them away and run to safety.
3	The individual followed me to my dorm room and tried to grab me. I managed to escape and lock the door.
4	The perpetrator touched me without my consent during a campus event. I felt violated and immediately reported it to security.

Cleaned 'Report' column:

	Cleaned_Report
0	The individual kept touching my hair and face without permission I told them to stop but they didnt listen
1	While waiting for the bus a stranger attempted to grope me I shouted for help and they fled the scene
2	The perpetrator blocked my path and tried to kiss me I managed to push them away and run to safety
3	The individual followed me to my dorm room and tried to grab me I managed to escape and lock the door
4	The perpetrator touched me without my consent during a campus event I felt violated and immediately reported it to security

Figure 3. Data cleaning

Cleaned 'Report' column before casefolding:

	Cleaned_Report
0	The individual kept touching my hair and face without permission I told them to stop but they didnt listen
1	While waiting for the bus a stranger attempted to grope me I shouted for help and they fled the scene
2	The perpetrator blocked my path and tried to kiss me I managed to push them away and run to safety
3	The individual followed me to my dorm room and tried to grab me I managed to escape and lock the door
4	The perpetrator touched me without my consent during a campus event I felt violated and immediately reported it to security

Sesudah Casefolding:

Cleaned 'Report' column after casefolding:

	Casefolded_Report
0	the individual kept touching my hair and face without permission i told them to stop but they didnt listen
1	while waiting for the bus a stranger attempted to grope me i shouted for help and they fled the scene
2	the perpetrator blocked my path and tried to kiss me i managed to push them away and run to safety
3	the individual followed me to my dorm room and tried to grab me i managed to escape and lock the door
4	the perpetrator touched me without my consent during a campus event i felt violated and immediately reported it to security

Figure 4. Case folding

Casefolded 'Report' column after normalisasi:

	Casefolded_Report
0	the individual kept touching my hair and face without permission i told them to stop but they didnt listen
1	while waiting for the bus a stranger attempted to grope me i shouted for help and they fled the scene
2	the perpetrator blocked my path and tried to kiss me i managed to push them away and run to safety
3	the individual followed me to my dorm room and tried to grab me i managed to escape and lock the door
4	the perpetrator touched me without my consent during a campus event i felt violated and immediately reported it to security

Casefolded 'Report' column before normalisasi:

	Normalized_Report
0	the individual kept touching my hair and face without permission i told them to stop but they did not listen
1	while waiting for the bus a stranger attempted to grope me i shouted for help and they fled the scene
2	the perpetrator blocked my path and tried to kiss me i managed to push them away and run to safety
3	the individual followed me to my dorm room and tried to grab me i managed to escape and lock the door
4	the perpetrator touched me without my consent during a campus event i felt violated and i ammediately reported it to security

Figure 5. Normalization

JURNAL ECOTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

Normalized 'Report' column before tokenization:

	Normalized_Report
0	the individual kept touching my hair and face without permission i told them to stop but they did not listen
1	while waiting for the bus a stranger attempted to grope me i shouted for help and they fled the scene
2	the perpetrator blocked my path and tried to kiss me i managed to push them away and run to safety
3	the individual followed me to my dorm room and tried to grab me i managed to escape and lock the door
4	the perpetrator touched me without my consent during a campus event i felt violated and i ammediately reported it to security

After Tokenization:

Normalized 'Report' column after tokenization:

	Tokenized_Report
0	[the, individual, kept, touching, my, hair, and, face, without, permission, i, told, them, to, stop, but, they, did, not, listen]
1	[while, waiting, for, the, bus, a, stranger, attempted, to, grope, me, i, shouted, for, help, and, they, fled, the, scene]
2	[the, perpetrator, blocked, my, path, and, tried, to, kiss, me, i, managed, to, push, them, away, and, run, to, safety]
3	[the, individual, followed, me, to, my, dorm, room, and, tried, to, grab, me, i, managed, to, escape, and, lock, the, door]
4	[the, perpetrator, touched, me, without, my, consent, during, a, campus, event, i, felt, violated, and, i, ammediately, reported, it, to, security]

Figure 6. Tokenizing

Before Stopword Removal:

Tokenized 'Report' column before stopword removal:

	Tokenized_Report
0	[the, individual, kept, touching, my, hair, and, face, without, permission, i, told, them, to, stop, but, they, did, not, listen]
1	[while, waiting, for, the, bus, a, stranger, attempted, to, grope, me, i, shouted, for, help, and, they, fled, the, scene]
2	[the, perpetrator, blocked, my, path, and, tried, to, kiss, me, i, managed, to, push, them, away, and, run, to, safety]
3	[the, individual, followed, me, to, my, dorm, room, and, tried, to, grab, me, i, managed, to, escape, and, lock, the, door]
4	[the, perpetrator, touched, me, without, my, consent, during, a, campus, event, i, felt, violated, and, i, ammediately, reported, it, to, security]

After Stopword Removal:

Tokenized 'Report' column after stopword removal:

	Stopword_Removed_Report
0	[individual, kept, touching, hair, face, without, permission, told, stop, listen]
1	[waiting, bus, stranger, attempted, grope, shouted, help, fled, scene]
2	[perpetrator, blocked, path, tried, kiss, managed, push, away, run, safety]
3	[individual, followed, dorm, room, tried, grab, managed, escape, lock, door]
4	[perpetrator, touched, without, consent, campus, event, felt, violated, ammediately, reported, security]

Figure 7. Stopword Removal

Stopword Removed 'Report' column before lemmatization:

	Stopword_Removed_Report
0	[individual, kept, touching, hair, face, without, permission, told, stop, listen]
1	[waiting, bus, stranger, attempted, grope, shouted, help, fled, scene]
2	[perpetrator, blocked, path, tried, kiss, managed, push, away, run, safety]
3	[individual, followed, dorm, room, tried, grab, managed, escape, lock, door]
4	[perpetrator, touched, without, consent, campus, event, felt, violated, immediately, reported, security]

After Lemmatization:

Stopword Removed 'Report' column after lemmatization:

	Lemmatized_Report
0	[individual, kept, touch, hair, face, without, permission, told, stop, listen]
1	[wait, bus, stranger, attempt, grope, shout, help, flee, scene]
2	[perpetrator, block, path, try, kiss, manage, push, away, run, safety]
3	[individual, follow, dorm, room, try, grab, manage, escape, lock, door]
4	[perpetrator, touch, without, consent, campus, event, felt, violate, immediately, report, security]

Figure 8. Lemmatization

JURNAL ECOTIPE Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

The image shows the steps in the text cleaning and processing process in Natural Language Processing (NLP):

1. Figure 3: Data Cleanup – Remove punctuation, contraction, and lower capital letters to get cleaner text.
2. Figure 4: Case Folding – Lowercase all text for consistency.
3. Figure 5: Normalization - Correcting spelling errors, removing unimportant words, and word duplication.
4. Figure 6: Tokenizing - Breaks down the text into word units (tokens) for more detailed analysis.
5. Figure 7: Stopword Removal - Removes common words that do not have significant informational value.
6. Figure 8: Lemmatization – Converting a word to its basic form for consistency and reducing word variation.

This process ensures that the text becomes cleaner and ready for further analysis in NLP applications.

3.3. Data Splitting

Data splitting is an important step in data analysis and machine learning that aims to separate the dataset into two main parts: training and testing. This data sharing ensures that the built model can be evaluated objectively and has good generalization capabilities against new data. Training sets are used to train models, where the model learns patterns and relationships from that data. The test set is used to evaluate the performance of a model after it has been trained, with the goal of measuring how well the model performs on data that has never been seen before. By dividing the data into training and testing sets, we can avoid overfitting and ensure a more robust and reliable model.

3.4. TF-IDF Weighting

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical method used to evaluate how important a word is in the context of a document relative to the document set (corpus). It helps in identifying keywords that have significant meaning in the text.

	accidentally	account	activity	advance	almost	annoyed	another	anxious
count	3368	3368	3368	3368	3368	3368	3368	3368
mean	0.0094312517	0.0052885365	0.0047927277	0.015868244	0.0049096861	0.0034802996	0.0034802996	0.0039670765
std	0.054041494	0.0454526182	0.0411913616	0.0682104348	0.0453672864	0.0344686643	0.0344686643	0.0387185154
min	0	0	0	0	0	0	0	0
25%	0	0	0	0	0	0	0	0
50%	0	0	0	0	0	0	0	0
75%	0	0	0	0	0	0	0	0
max	0.3429979625	0.395817578	0.3587090384	0.3295720732	0.423995455	0.3447543839	0.3447543839	0.3817461074

Figure 9. TF-IDF Results

Figure 9 shows a basic statistical table for some of the words in the dataset, including mean, standard deviation, minimum, maximum, and specific percentiles for the TF-IDF values of each word. For example, the word "accidentally" has a TF-IDF mean value of 0.0049 and a standard deviation (std) of 0.0540.

JURNAL ECTIPE
Electronic, Control, Telecommunication, Information, and Power Engineering
Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

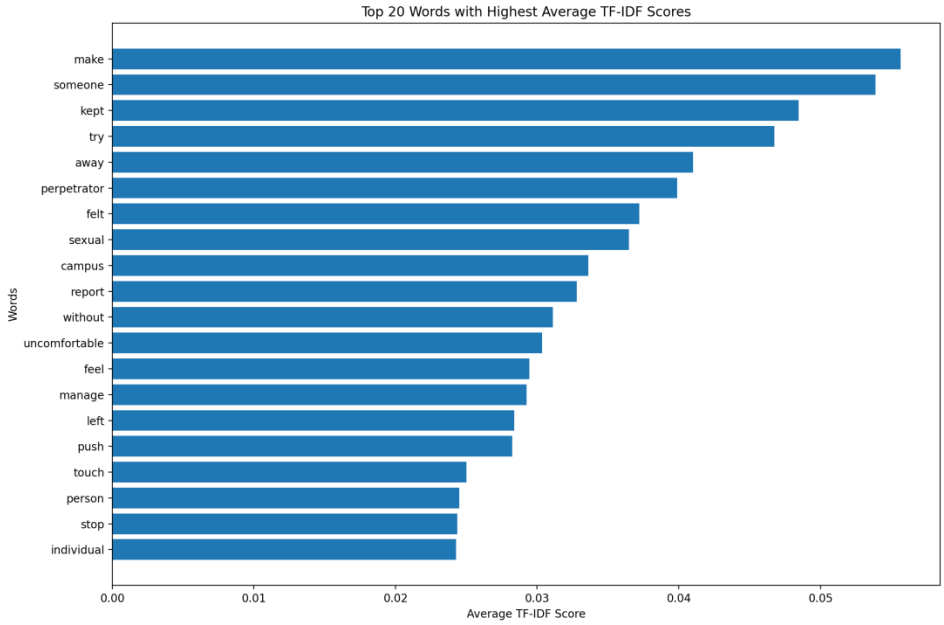

Figure 10. Words with the highest average TF-IDF value in Corpus

Figure 10. shows a bar graph showing 20 words with the highest average TF-IDF value in the corpus. These words include "make", "someone", "kept", "try", and others. The length of each bar represents the average TF-IDF value of the word, with the word "make" having the highest value. These graphs help in identifying the most significant words in the document set and provide insight into the keywords that may be relevant in further analysis.

3.5. Classification with SVM

After the data was divided into train and test data, TFIDF weighting was carried out and continued with classification using the Support Vector Machine (SVM) method. This study uses four SVM kernels: linear, polynomial, sigmoid, and RBF. The four kernels will be tuned to determine the best parameter value on each kernel using the Grid Search method by entering the hyperparameter value as the input. The Grid Search process will generate the best parameter values for each kernel. The parameter input values were processed and tested on the training data using grid search to obtain the optimal combination of parameter values. The best hyperparameters generated from each kernel are presented in Table 2. After determining the optimal parameter values for each kernel, the performance of each kernel is evaluated in terms of accuracy, precision, recall, and F1-Score. The RBF kernel achieved the highest accuracy, precision, recall, and F1-Score, with values of 0.90 0.89 0.88 0.885 respectively.

3.6. Evaluasi

After training the SVM (Support Vector Machine) model with various kernels, the next step is to evaluate the model's performance using several evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics provide a variety of perspectives on the model's performance. Additionally, the confusion matrix is used to provide a more detailed picture of the correct and false predictions made by the model. The confusion matrix helps in identifying the types of errors made by the model, such as false positives and false negatives. Figure 11 will display the results of the confusion matrix for the SVM model.

Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

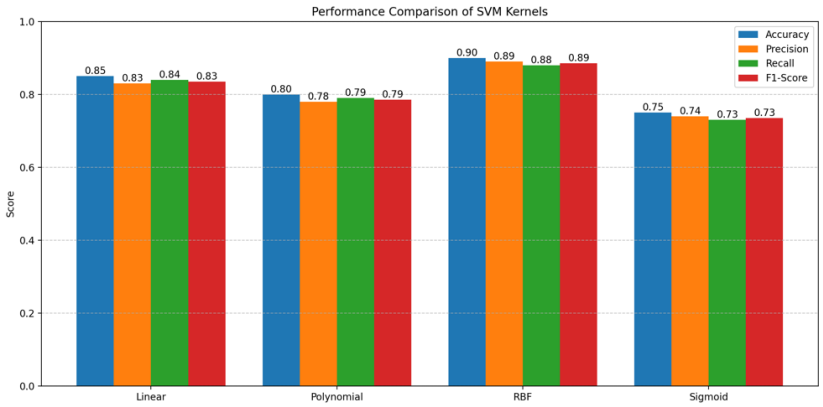

Figure 11. Results of the Confusion Matrix of the Karnel Baningan

Figure 11 shows a comparison of the performance of the SVM (Support Vector Machine) model using different kernel types: Linear, Polynomial, RBF (Radial Base Function), and Sigmoid. The evaluation was carried out based on the main metrics, namely accuracy, precision, recall, and F1-score. The RBF kernel showed the best performance with an accuracy value of 0.90, precision of 0.89, recall of 0.88, and F1-score of 0.89. This kernel is very effective in handling non-linear data. The Linear Kernel also performs quite well with almost balanced metric values, slightly below the RBF. Meanwhile, the Polynomial and Sigmoid kernels showed lower performance with all metrics being around 0.78 to 0.80 for Polynomial and 0.73 to 0.75 for Sigmoid. From the results of this evaluation, it can be seen that the RBF kernel is the best choice for this scenario, as it provides the most accurate and consistent prediction and classification results among all the kernels tested.

The novelty of this research lies in several key aspects that distinguish it from previous studies. First, this research combines Support Vector Machines (SVM) with Natural Language Processing (NLP) to handle sexual violence report data, an approach that has not been widely explored in earlier studies. This integration of technologies offers a more accurate solution for report classification while maintaining the anonymity of victims, which is a sensitive issue in this context. Second, this research conducts a comprehensive analysis of various SVM kernels (Linear, Polynomial, RBF, and Sigmoid) using several evaluation metrics such as accuracy, precision, recall, and F1-score. The findings demonstrate that the RBF kernel provides the best performance in handling non-linear data, with higher accuracy and consistency compared to other kernels. Third, this research offers real-world applications by developing a practical and scalable reporting system for educational institutions. Unlike previous studies, which tend to be theoretical, this study provides a concrete solution that can be implemented to more effectively address sexual violence cases. Finally, this research introduces data-driven improvements, where the use of the RBF kernel in the SVM model has proven to enhance the accuracy and efficiency of classification compared to traditional methods, offering a new contribution in handling sexual violence reports within the campus environment.

4. CONCLUSION

This research aims to develop a Natural Language Processing (NLP)-based system to improve the This research successfully developed a Natural Language Processing (NLP)-based system combined with the Support Vector Machines (SVM) algorithm to improve the reporting and handling of sexual violence cases on campus. The RBF kernel outperformed other kernels with an accuracy of 0.90, precision of 0.89, recall of 0.88, and an F1-score of 0.89, demonstrating its effectiveness in handling non-linear data. The model was trained on 3,000 reports, divided into 80% training and 20% testing data, ensuring that the system could generalize well.

However, this study has several limitations. First, the dataset size was relatively small, and future research should include larger, more diverse data to ensure better generalization. Second, the scope was limited to textual reports, while future studies should consider incorporating multimedia elements such as images or videos for a more comprehensive analysis. Additionally, this research did not specifically address the issue of imbalanced data, which may skew results in categories with fewer occurrences.

Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

Addressing this issue with techniques like oversampling or undersampling could improve the model's performance. Lastly, the computational resources required for deep learning models like this one are considerable, and future research should focus on optimizing the model for greater efficiency and scalability. Despite these limitations, this study provides a solid foundation for improving sexual violence reporting systems on campus, offering better support for victims while ensuring accurate and efficient classification of reports.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the Directorate of Research, Technology, and Community Service (DRTPM) for their invaluable support and guidance during this research. We would also like to thank the Institute for Research and Community Service (LPPM) of Universal University for providing the necessary resources and facilities. The administrative support and enthusiasm provided by LPPM greatly helps the progress and quality of our research. Thank you to all faculty members, staff, and colleagues who have helped directly or indirectly in this research. Our award is also given to participants and respondents whose contributions are critical to the study. Without the support of all the parties mentioned above, this research would not have been possible.

REFERENCES

[1] Raineka Faturani, "Kekerasan Seksual di Lingkungan Perguruan Tinggi," Sep 2022, doi: 10.5281/ZENODO.7052155.

[2] S. Sopyandi dan S. Sujarwo, "Kekerasan Seksual di Lingkungan Pendidikan dan Pencegahannya," *J. Pendidik. Ilmu Pengetah. Sos.*, vol. 15, no. 1, hlm. 19–25, Mei 2023, doi: 10.37304/jpips.v15i1.9448.

[3] A. Y. Susilowati, "Kampus Ramah Mahasiswa dari Kekerasan Seksual: Analisis Tingkat Pengetahuan Mahasiswa Terkait Pencegahan dan Penanganan Kekerasan Seksual," *Empower J. Pengemb. Masy. Islam*, vol. 7, no. 2, hlm. 233, Des 2022, doi: 10.24235/empower.v7i2.11516.

[4] D. S. Yunina *dkk.*, "Sosialisasi 3 Dosa Besar Dalam Pendidikan Untuk Menanamkan Nilai Karakter Peserta Didik di SDN Banjar Kemuning," vol. 05, no. 02, 2023.

[5] Franciscus Xaverius Wartoyo dan Yuni Priskila Ginting, "Kekerasan Seksual Pada Lingkungan Perguruan Tinggi Ditinjau Dari Nilai Pancasila," *J. Lemhannas RI*, vol. 11, no. 1, hlm. 29–46, Mei 2023, doi: 10.55960/jlri.v11i1.423.

[6] F. Bentivegna dan P. Patalay, "The impact of sexual violence in mid-adolescence on mental health: a UK population-based longitudinal study," *Lancet Psychiatry*, vol. 9, no. 11, hlm. 874–883, Nov 2022, doi: 10.1016/S2215-0366(22)00271-1.

[7] L. M. Orchowski, L. Grocott, K. W. Bogen, A. Ilegbusi, A. B. Amstadter, dan N. R. Nugent, "Barriers to Reporting Sexual Violence: A Qualitative Analysis of #WhyIDidntReport," *Violence Women*, vol. 28, no. 14, hlm. 3530–3553, Nov 2022, doi: 10.1177/10778012221092479.

[8] K. Parti dan R. A. Robinson, "What Hinders Victims from Reporting Sexual Violence: A Qualitative Study with Police Officers, Prosecutors, and Judges in Hungary," *Int. J. Crime Justice Soc. Democr.*, vol. 10, no. 2, Jun 2021, doi: 10.5204/ijcjsd.1851.

[9] C. Peersman, M. Edwards, E. Williams, dan A. Rashid, "A Survey of Relevant Text Mining Technology," 2022, *arXiv*. doi: 10.48550/ARXIV.2211.15784.

[10] F. Balahadia, Z. J. Astoveza, G. Jamolin, dan N. E. A. Astoveza, "Development and Implementation of Violence against Women and their Children Report System Mobile Application," *Int. J. Sci. Technol. Eng. Math.*, vol. 2, no. 3, hlm. 17–42, Sep 2022, doi: 10.53378/352906.

Volume 11, Issue 2, October 2024, pp. 193-204
ISSN 2355-5068 ; e-ISSN 2622-4852
DOI: 10.33019/jurnalecotipe.v11i2.4511

[11] Q. Zeng *dkk.*, "Improved Naive Bayes with Mislabeled Data," 2023, *arXiv*. doi: 10.48550/ARXIV.2304.06292.

[12] S. Etzler, F. D. Schönbrodt, F. Pargent, R. Eher, dan M. Rettenberger, "Machine Learning and Risk Assessment: Random Forest Does Not Outperform Logistic Regression in the Prediction of Sexual Recidivism," *Assessment*, vol. 31, no. 2, hlm. 460–481, Mar 2024, doi: 10.1177/10731911231164624.

[13] K. L. Tan, C. P. Lee, K. M. Lim, dan K. S. M. Anbananthen, "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," *IEEE Access*, vol. 10, hlm. 103694–103704, 2022, doi: 10.1109/ACCESS.2022.3210182.

[14] D. Mustafa Abdullah dan A. Mohsin Abdulazeez, "Machine Learning Applications based on SVM Classification A Review," *Qubahan Acad. J.*, vol. 1, no. 2, hlm. 81–90, Apr 2021, doi: 10.48161/qaj.v1n2a50.

[15] A. O. Kuyoro, S. Alimi, dan O. Awodele, "Comparative Analysis of the Performance of Various Support Vector Machine kernels," dalam *2022 5th Information Technology for Education and Development (ITED)*, Abuja, Nigeria: IEEE, Nov 2022, hlm. 1–7. doi: 10.1109/ITED56637.2022.10051564.

[16] B. A. Kindhi, N. Susanto, W. Handayani, S. V. Kurniasari, dan A. P. Pratama, "Prediction of the Tuberculosis Patients Who Can Recover Normally Using a Support Vector Machine with Radial and Polynomial Kernels," dalam *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Surabaya, Indonesia: IEEE, Apr 2021, hlm. 365–368. doi: 10.1109/EIConCIT50028.2021.9431878.

[17] S. Saha, M. Das, B. S. Mondal, S. Sarkar, dan J. Maiti, "D $_i$ PSVM: A Polynomial Kernel-free Support Vector Machine," dalam *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, Sakheer, Bahrain: IEEE, Okt 2021, hlm. 448–452. doi: 10.1109/ICDABI53623.2021.9655976.

[18] R. Chahar, A. K. Dubey, dan S. K. Narang, "A Mental Health Performance Assessment using Support Vector Machine," dalam *2023 3rd International Conference on Intelligent Technologies (CONIT)*, Hubli, India: IEEE, Jun 2023, hlm. 1–7. doi: 10.1109/CONIT59222.2023.10205772.

[19] S. Jueyendah, M. Lezgy-Nazargah, H. Eskandari-Naddaf, dan S. A. Emamian, "Predicting the mechanical properties of cement mortar using the support vector machine approach," *Constr. Build. Mater.*, vol. 291, hlm. 123396, Jul 2021, doi: 10.1016/j.conbuildmat.2021.123396.

[20] L. Fischer dan P. Wollstadt, "Precision and Recall Reject Curves for Classification," 2023, *arXiv*. doi: 10.48550/ARXIV.2308.08381.

[21] R. Yusof, N. Hashim, N. Abdul Rahman, S. Y. Mohd Yunus, dan N. A. Aziz Fadzillah, "Academic Performance Prediction Model Using Classification Algorithms: Exploring the Potential Factors," *Int. J. Acad. Res. Progress. Educ. Dev.*, vol. 11, no. 3, hlm. Pages 706-724, Agu 2022, doi: 10.6007/IJARPED/v11-i3/14753.

[22] Z. Tian, Y. Li, Z. Li, dan S. Li, "Recall Network: A Simple Brain-Inspired Algorithm for Classification," *Comput. Intell. Neurosci.*, vol. 2022, hlm. 1–52, Agu 2022, doi: 10.1155/2022/9374946.

[23] A. Humphrey *dkk.*, "Machine-learning classification of astronomical sources: estimating F1-score in the absence of ground truth," *Mon. Not. R. Astron. Soc. Lett.*, vol. 517, no. 1, hlm. L116–L120, Okt 2022, doi: 10.1093/mnrasl/slac120.