Implementation of Particle Swarm Optimization on Sentiment Analysis of Cyberbullying using Random Forest

Helma Herlinda[1*], Muhammad Itqan Mazdadi[2], Muliadi[3], Dwi Kartini[4], Irwan Budiman[5]

[1,2,3,4,5]*Computer Science, Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, Indonesia*

*corr_author: helmahher@gmail.com

Abstract - Social media has exerted a significant influence on the lives of the majority of individuals in the contemporary era. It not only enables communication among people within specific environments but also facilitates user connectivity in the virtual realm. Instagram is a social media platform that plays a pivotal role in the sharing of information and fostering communication among its users through the medium of photos and videos, which can be commented on by other users. The utilization of Instagram is consistently growing each year, thereby potentially yielding both positive and negative consequences. One prevalent negative consequence that frequently arises is cyberbullying. Conducting sentiment analysis on cyberbullying data can provide insights into the effectiveness of the employed methodology. This research was conducted as an experimental research, aiming to compare the performance of Random Forest and Random Forest after applying the Particle Swarm Optimization feature selection technique on three distinct data split compositions, namely 70:30, 80:20, and 90:10. The evaluation results indicate that the highest accuracy scores were achieved in the 90:10 data split configuration. Specifically, the Random Forest model yielded an accuracy of 87.50%, while the Random Forest model, after undergoing feature selection using the Particle Swarm Optimization algorithm, achieved an accuracy of 92.19%. Therefore, the implementation of Particle Swarm Optimization as a feature selection technique demonstrates the potential to enhance the accuracy of the Random Forest method.

Keywords: Rapidminer, social media, data science, text mining, classification

I. INTRODUCTION

Social media has exerted a significant influence on the lives of the majority of individuals in the contemporary era. It enables not only communication among people within specific environments but also facilitates user connectivity in the virtual realm. Through social media, individuals can easily communicate with one another via various applications on digital devices, with Instagram being one of them [1][2]. Instagram, as a social media platform, serves as a means for users to share information and engage in communication through the exchange of photos and videos, which can be commented on by other users. Since its initial launch in October 2010, Instagram has experienced rapid growth and has emerged as a highly popular social media platform. As of now, Instagram boasts a user base of 1 billion active users and has witnessed the sharing of over 40 billion photos since its establishment [3]-[4].

The utilization of Instagram continues to grow annually, giving rise to both positive and negative consequences. One prevalent negative consequence that frequently occurs is cyberbullying [5]. Cyberbullying is a phenomenon characterized by repeated aggressive and intentional behaviors conducted by individuals or groups through electronic devices, aiming to create divisions and gaps among victims. This is achieved through the dissemination of hateful messages or comments [6]-[7]. The number of cyberbullying victims has been steadily increasing over the years. In 2007, approximately 18% of users reported being targeted by cyberbullying, which increased to 36% in 2019. It is expected that this trend will continue to rise due to the growing popularity of Instagram, the influence of social environments, and the unrestricted use of mobile devices by children and adolescents without adequate supervision [8]. Cyberbullying has the potential to spread rapidly due to Instagram's extensive user base and its accessibility to the public. This modern phenomenon imposes numerous consequences on victims, as they are subjected to actions from perpetrators who harbor a sense of superiority [9]-[10].

In recent years, cyberbullying has garnered increasing attention. Extensive research has been conducted to explore the utilization of machine learning algorithms in addressing comments on Instagram and classifying cyberbullying instances through sentiment analysis. Sentiment analysis has demonstrated its efficacy as a

suitable tool for assessing the accuracy of public opinion, which holds crucial significance across various domains. This analytical approach plays a pivotal role in informing decision-making processes [11]-[12].

Previous research on cyberbullying utilizing the same dataset was conducted by Hanni [13]. The research aimed to investigate the effectiveness of the Support Vector Machine (SVM) method in applying sentiment analysis for cyberbullying detection on Instagram. The research obtained accuracy values of 86% for the 90:10 data split, 78% for the 80:20 data split, 82% for the 70:30 data split, and 83% for the 60:40 data split [14]. Previous research was conducted by Afdhal et al. [15] using the Random Forest algorithm for sentiment analysis of comments on Youtube concerning Islamophobia. The research consisted of three experiments with data splits of 90:10, 80:20, and 70:30. The obtained accuracy values for each experiment were 79%, 74.50%, and 73% respectively. When accuracy values are not sufficiently high, the implementation of feature selection methods becomes necessary to enhance these values and attain satisfactory results. Typically, feature selection is employed to assess the performance of a method before and after the incorporation of feature selection techniques. Particle Swarm Optimization is one of the feature selection methods that can be employed due to its ability to aid in both the classification and optimization processes [16]. Previous research was conducted by Setiawan et al. [17], which involved a comparison of the Naive Bayes and Support Vector Machine methods using Particle Swarm Optimization feature selection in sentiment analysis of the Esemka car. The findings of this research demonstrate that the application of Particle Swarm Optimization feature selection can significantly improve accuracy. Specifically, the Naive Bayes method without feature selection achieved an accuracy value of 75.04%, whereas the Naive Bayes method with feature selection obtained an accuracy value of 83.33%. The Support Vector Machine method without the application of Particle Swarm Optimization feature selection yielded an accuracy value of 78.81%. However, when the Support Vector Machine method was combined with the Particle Swarm Optimization feature selection technique, the accuracy value significantly improved to 88.19%.

Considering the aforementioned issue, we propose a research study to conduct sentiment analysis on cyberbullying data, given that this is a crucial issue that is highly prevalent. Sentiment analysis capabilities have shown accurate results in gauging public opinion. The sentiment analysis process in this study will involve the Random Forest method as a data classification method, as Random Forest possesses several advantages, including the ability to achieve high accuracy, resilience to outliers and noise, and faster performance compared to bagging and boosting methods [15]. Therefore, this method is deemed suitable for this research. However, to attain good accuracy results, apart from utilizing a good classification method, there are other stages that can help improve the accuracy level, such as feature selection. As seen in the previously mentioned research, the addition of Particle Swarm Optimization feature selection applied prior to SVM and Naïve Bayes classification yielded better accuracy results.

Therefore, the researcher also proposes to incorporate the Particle Swarm Optimization feature selection process into the Random Forest classification, aiming to improve accuracy. Based on this rationale, the research process will involve comparing the classification results between the processes that utilize feature selection and those that do not. The objectives of this research are to evaluate the classification performance of Random Forest and to determine the difference in accuracy obtained after incorporating the Particle Swarm Optimization feature selection. Both methods will be applied to cyberbullying case data.

II. METHOD

The research involved several stages, starting with the acquisition of readily available data. The collected data then underwent various processes including pre-processing, feature extraction using TF-IDF, feature selection employing the Particle Swarm Optimization algorithm, and classification utilizing the Random Forest method. This research will later compare the classification result between data that uses feature selection and does not use feature selection to obtain comparative results. An overview of the research methodology is presented in Fig. 1.

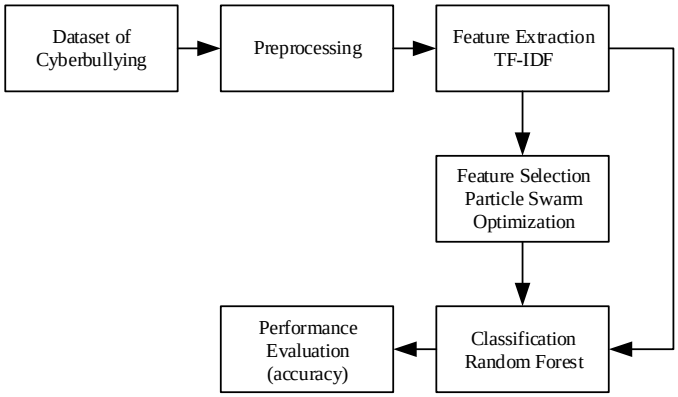

Fig. 1 Research flowchart

A. Data Collection

The dataset used in this research was obtained from the Kaggle website. Following the methodology proposed by Hanni [13], the dataset was manually collected over a duration of two to three months. The collection process involved visiting the Instagram profiles of various Indonesian artists/celebrities, selecting specific uploads from their feeds, and extracting comments provided by netizens on these selected uploads. The research dataset used in this research comprises a total of 650 data points, which are classified into two categories: bullying and non-bullying. Each category contains an equal number of data instances. The dataset includes five features, namely Instagram name, comment, category, post date, and artist/celebgram Instagram account name. Here are the examples of 10 comments provided by netizens, as shown in Table I.

The identity of the comment sender's account can be seen in Table I, along with manual categorization of whether the comment is classified as bullying or non-bullying. From these comments, further processing will be conducted according to the research methodology that has been established.

B. Pre-processing

Pre-processing plays a crucial role in cleansing the data by eliminating irrelevant words that could potentially impact the sentiment analysis process [11]. The pre-processing steps to be undertaken in this research include cleaning, case folding, tokenization, normalization, stopword removal, and stemming. Cleaning involves the elimination of non-essential components in the document, such as special characters, numbers, URL links, hashtags, and mentions, to ensure data cleanliness. Case folding serves the purpose of converting all the text in the dataset into lowercase. Tokenization, on the other hand, plays a crucial role in removing punctuation marks and dividing each sentence in the document into individual words. Normalization involves identifying erroneous or damaged words and subsequently correcting or removing them. This process contributes to enhancing the accuracy of the classification results by converting non-standard or slang words into standard words. Stopword removal is a crucial step in which significant words are retained while irrelevant words are eliminated. The stemming stage plays a pivotal role in identifying base words by removing affixes from each word in the dataset [13]. For this research, the pre-processing stage will be conducted using Google Colaboratory.

TABLE I
RESEARCH DATASET

No.	Instagram Name	Comments	Category	Posting Date	IG Account Name
1	@delliananda	*"Kaka tidur yaa, udah pagi, gaboleh capek2"*	Non-bullying	14 Oct 2019	@isyanasarasvati
2	@fenninbl	*"makan nasi padang aja begini badannya"*	Non-bullying	14 Oct 2019	@isyanasarasvati
3	@abdurahmanshq	*"yang aku suka dari dia adalah selalu cukur jembut sebelum manggung"*	Bullying	14 Oct 2019	@isyanasarasvati
4	@najla.yoo	*"Hai kak Isyana aku ngefans banget sama kak Isyana.aku paling suka lagu kak Isyana itu lagu tetap didalam jiwa"*	Non-bullying	14 Oct 2019	@isyanasarasvati
5	@dessy_____	*"Manusia apa bidadari sih herann deh cantik terus ????"*	Non-bullying	14 Oct 2019	@isyanasarasvati
6	@e.fril	*"@ayu.kinantii isyan skrg berubah ya:(baju nya nakal"*	Bullying	14 Oct 2019	@isyanasarasvati
7	@bahasa.bayi.planet	*"Gemesnya isyan kayak tango, berlapis lapis ciaaaa"*	Non-bullying	16 Sept 2019	@isyanasarasvati
8	@khanayarudinita	*"Makin jelek aja anaknya, padahal ibu ayahnya cakep2"*	Bullying	22 June 2019	@tasyakamila
9	@reniaulia225	*"Kok anaknya kayak udah tua gitu ya mukanya kk tasya"*	Bullying	22 June 2019	@tasyakamila
10	@nurjanah.hani	*"Muka anak nya ko tua banget yaa.. GK ngegemesin GK ada lucu2nya"*	Bullying	22 June 2019	@tasyakamila

C. TF-IDF Feature Extraction

In text mining, TF-IDF is described as a methodology for extracting information, conducting data mining, and uncovering knowledge pertaining to the contents of a database [18]. TF-IDF facilitates the conversion of text into vectors, enabling proper processing of comments by machine learning algorithms. The TF component of TF-IDF is valuable for determining the frequency value of a word appearing in a document. On the other hand, IDF plays a pivotal role in quantifying the significance of a word in distinguishing text classification. TF-IDF is commonly employed in the feature extraction stage, offering superior accuracy compared to the Bag of Words (BOW) approach [19]-[20]. The TF-IDF feature extraction process involves several steps. Firstly, the individual words within each document are collected to construct a set of features for each document. Secondly, the TF-IDF score is calculated for each word within each document. Lastly, all individual words within the document are sorted based on their corresponding TF-IDF scores. Subsequently, a varying percentage of words with the highest TF-IDF scores is selected to form a feature set, also known as a vocabulary, for representation purposes. The feature sets across the entire document are constructed by combining the selected individual words from each document. Lastly, the constructed feature set is utilized to represent each document within the corpus. In this representation, the term weight of each word in a document corresponds to its TF-IDF score within that particular document. Alternatively, if an individual word does not appear in a document, its term weight is assigned as 0. Consequently, a document-term vector is created for each document. The systematic presentation of the TF-IDF architecture can be illustrated through (1) and (2).

$$IDF = log\frac{D}{DF} \quad (1)$$

$$TF - IDF = tf * idf \quad (2)$$

In the aforementioned formula, D represents the total number of documents in the training data, DF indicates the number of documents containing the specific word, tf signifies the term frequency or the frequency of the word within the document, and Idf denotes the inverse document frequency for each term/word.

D. Particle Swarm Optimization Feature Selection

In 1995, Eberhart and Kennedy introduced the Particle Swarm Optimization (PSO) algorithm [17]. PSO involves a swarm of particles that collectively search for the optimal position, enabling the identification of the most suitable solution for solving optimization problems within a virtual search space. Each particle in the Particle Swarm Optimization algorithm maintains knowledge of its personal best position and the distance it has traveled during the movement process. The algorithm's search for the best solution will cease when the optimal solution is discovered or specific conditions are satisfied [16].

In the PSO algorithm, a swarm is comprised of multiple particles, with each particle possessing its own speed and position. In the speed update equation, the particle's velocity is influenced by the current velocity, the particle's individual best position (*pbest*), and the global best position (*gbest*). The position of each particle is determined by the current position and the new update rate. In the t-th generation, the position and velocity of the i-th particle are denoted as $xi(t)$ and $vi(t)$, respectively [21]. The Particle Swarm Optimization algorithm employed in this research incorporates several parameters, such as population size (5), the maximum number of generations (30), the minimum weight (0), the maximum weight (1), the inertia weight (0.6), *gbest* (0.3), and *pbest* (0.6).

In Saputra's research [22], it was elucidated that the Particle Swarm Optimization algorithm is extensively employed across diverse domains due to its straightforward operation and rapid convergence. Assuming the existence of m particles in the solution space, the expressions for the position and velocity of the I-th particle in an n-dimensional space are depicted in (3) and (4).

$$Pi = (P_{i1}, P_{i2}, \ \dots \ , Pin), \quad (3)$$

$$Vi = (V_{i1}, V_{i2}, \ \dots \ , Pin) \quad (4)$$

Eq. (3) and (4) can be enhanced by incorporating inertia. The update of the particle's position and velocity can be performed using (5) and (6).

$$P_i^{e+1} = P_i^e + V_i^{e+1} \quad (5)$$

$$V_i^{e+1} = V_i^e + S_1R_1(P_{best} - P_i^e) + S_1R_1(G_{best} - P_i^e) \quad (6)$$

E. Random Forest Classification

Random Forest algorithm has been developed to address classification and regression problems [23]. This method is based on an ensemble of multiple decision trees, where each tree relies on a randomly sampled vector. Each tree generates its own prediction, which may differ from the others [14], [24]. Random Forest combines multiple decision tree models to address the issue of overfitting in individual decision trees. Key aspects of the Random Forest method include employing integrated sampling techniques to construct prediction trees and aggregating the results of each tree through

majority voting. The advantages of the Random Forest method are its ability to achieve high accuracy, robustness against outliers and noise, and faster performance compared to bagging and boosting methods [15]. In this research, the default Random Forest parameters in the RapidMiner application will be used, including a total number of trees equal to 100, a criterion of gain_ratio, and a maximum depth of 10.

F. Assesment Index

The evaluation of the results in this research aims to obtain an accuracy value for the classification experiment in cyberbullying sentiment analysis. The inclusion of an index rating serves as a measure of the system's success and facilitates comparisons with other algorithms in similar studies. Accuracy is a metric that indicates the success rate of all results in comparison to the total number of negative results across all transactions [25][26]. Eq. (7) illustrates the calculation formula for accuracy.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

True Positive (TP) refers to the number of positive records in the dataset that are correctly classified as positive. True Negative (TN) indicates the number of negative records in the dataset that are accurately classified as negative. False Positive (FP) represents the number of negative records in the dataset that are incorrectly classified as positive. Lastly, False Negative (FN) denotes the number of positive records in the dataset that are erroneously classified as negative [17].

III. RESULT AND DISCUSSION

A. Pre-processing

Once the data has undergone cleaning, case folding, tokenization, normalization, stopword removal, and stemming stages, the resulting data becomes clean and suitable for classification. Table II provides an illustrative example of this process.

B. Feature Extraction TF-IDF

Upon successful completion of all pre-processing steps, the documents will undergo feature extraction using the TF-IDF algorithm to obtain their corresponding weighting values. The term frequency (TF) can be employed to compute the frequency value of a word within a document. Meanwhile, IDF plays a crucial role in assessing the significance of a word in distinguishing text classifications. Table III presents the illustrative results of feature extraction on 650 pre-processed data instances, each containing 325 labels.

Table III displays the weighting results obtained from feature extraction using the TF-IDF method. To obtain these TF-IDF feature extraction results, we refer back to equations (1) and (2), with the first step being to calculate the Term Frequency (TF) for each word based on its occurrence in the data. The TF value is obtained by counting the number of occurrences of the desired word and dividing it by the total number of words in the document.

Once the TF values are obtained, we proceed as in equation (1) to find the Inverse Document Frequency (IDF) value, which serves to reduce the weight of a word (term) if it appears in almost all of the data. The IDF value is calculated as in equation (1) by taking the logarithm of the total number of documents in the corpus divided by the number of documents that contain the word obtained in the TF calculation.

TABLE II
WEIGHTED BY TF-IDF

Category	*aktif*	*anak*	…	*zinah*
bullying	0	0.4159	…	0
bullying	0	0.3888	…	0
bullying	0.5950	0.3636	…	0
…	…	…	…	…
non-bullying	0	0.1814	…	0

TABLE III
DATA PRE-PROCESSING

No	Category	Comments	Stemming
1	non-bullying	*"Kaka tidur yaa, udah pagi, gaboleh capek2"*	*kakak tidur pagi capai*
2	non-bullying	*"makan nasi padang aja begini badannya"*	*makan nasi padang badan*
3	bullying	*"yang aku suka dari dia adalah selalu cukur jembut sebelum manggung"*	*suka cukur jembut panggung*
..	…	…	…
650	non-bullying	*"Inimah bukan main alat musik lagi. Olahraga jari dan kaki ini mah"*	*main alat musik olahraga jari kaki*

After obtaining the TF and IDF values, we then calculate the TF-IDF value as in equation (2) by multiplying the TF value with the IDF value. The weight values obtained after feature extraction can be useful for subsequent steps such as feature selection and classification.

C. PSO Feature Selection and RF Classification

The feature selection and classification process in this research utilizes the RapidMiner software. RapidMiner is a Java-based analytical tool renowned for its applications in data mining, text mining, predictions, and business analysis. With its popularity in the market, RapidMiner has emerged as a leading tool in its field [27]. RapidMiner offers both free and paid access options, depending on individual needs. It encompasses Open-Source Software (OSS), which makes it advantageous for researchers to explore. One of the applications of RapidMiner is sentiment inference, enabling the comparison of accuracy values with other methods. It provides a wide range of operators that can be utilized as per requirements. Meanwhile, if processed based on the used method, the Particle Swarm Optimization (PSO) feature selection can be calculated by referring to Eq. (3), (4), (5), and (6). Table IV presents the accuracy results of each experiment conducted in this research.

In Table IV, a comparison of accuracy values can be observed for the data split compositions of 70:30, 80:20, and 90:10. The comparison includes the Random Forest and Random Forest methods after feature selection using the Particle Swarm Optimization algorithm. Fig. 2 illustrates a comparison diagram of the accuracy values obtained from the experiments.

Table V presents a comparison of the accuracy results obtained from previous studies for reference and comparison purposes.

Based on the findings from previous research, as presented in Table V, it is evident that the utilization of the Support Vector Machine (SVM) method on the same dataset does not yield higher accuracy values compared to the Random Forest approach employed in this research. The SVM method achieved accuracy rates of only 86% in the 90:10 experiment, 78% in the 80:20 trial, 82% in the 70:30 trial, and 83% in the 60:40 trial. Similarly, the Random Forest experiments conducted in previous studies did not yield superior outcomes compared to the Random Forest experiment conducted in this research. The accuracy values obtained were only 79% in the 90:10 experiment, 74.50% in the 80:20 experiment, and 73% in the 70:30 experiment.

TABLE IV
ACCURACY COMPARISON

Split Composition	Remove Useless Attribute		Without Remove Useless Attribute	
	RF	PSO + Random Forest	RF	PSO + Random Forest
70:30	79,08%	87,24%	79,08%	87,24%
80:20	77,69%	88,46%	77,69%	88,46%
90:10	87,50%	92,19%	87,50%	92,19%

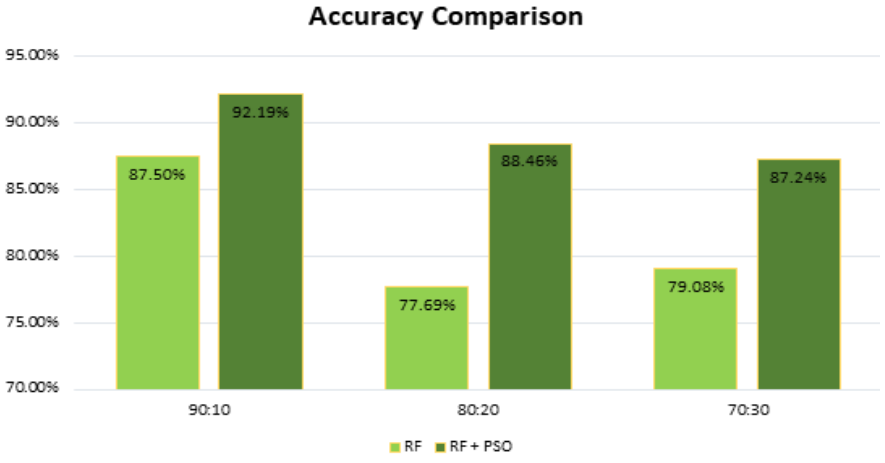

Fig. 2 Diagram of accuracy comparison

TABLE III
ACCURACY RESULTS OF PREVIOUS RESEARCH

Method	90:10	80:20	70:30	60:40	10-Fold Validation
Support Vector Machine [13]	86%	78%	82%	83%	-
Random Forest [15]	79%	74,50%	73%	-	-
Naïve Bayes[17]	-	-	-	-	75,04%
NB + PSO [17]	-	-	-	-	83,33%
Support Vector Machine [17]	-	-	-	-	78,81%
SVM + PSO [17]	-	-	-	-	88,19%
Random Forest [28]	-	-	-	-	97,26%
Support Vector Machine [28]	-	-	-	-	92,15%
Naïve Bayes[28]	-	-	-	-	88,39%

In [28], three experimental methods were employed, namely Random Forest, Support Vector Machine, and Naive Bayes Classifier. The accuracy values obtained through a 10-fold validation were as follows: 97.25% for Random Forest, 92.15% for Support Vector Machine, and 88.39% for Naive Bayes Classifier. These results indicate that the Random Forest method exhibits higher accuracy compared to the other methods. Previous research also demonstrates an improvement in accuracy values following the application of the Particle Swarm Optimization feature selection algorithm. For instance, the Naive Bayes method yielded an accuracy value of 75.04% before applying the Particle Swarm Optimization feature selection, whereas it increased to 83.33% after the feature selection was implemented. Similarly, the experiment conducted on the Support Vector Machine (SVM) method yielded an accuracy value of 78.81% prior to the application of the Particle Swarm Optimization feature selection. However, after implementing the Particle Swarm Optimization feature selection algorithm, the SVM method achieved an improved accuracy value of 88.19%. Thus, previous research has concluded that the implementation of the Particle Swarm Optimization feature selection resulted in an increase in the accuracy value.

Based on the results obtained from the conducted research and previous studies, it can be concluded that incorporating feature selection before the data classification process can lead to better accuracy. This is evident from the research results presented in Table IV of this study. With a composition of 70:30, the accuracy obtained from Random Forest is only 79.08%. However, when Random Forest incorporates the Particle Swarm Optimization (PSO) feature selection process, the accuracy increases to 87.24%. The same pattern is observed with a composition of 80:20, where the accuracy of Random Forest is 77.69%, but it increases to 88.46% when PSO feature selection is applied. Furthermore, even with a composition of 90:10, the accuracy of Random Forest improves to 87.50%, but it further increases to 92.19% after incorporating PSO feature selection.

These research findings successfully address the objectives of the study by demonstrating the performance of Random Forest classification without and with PSO feature selection using Cyberbullying data. The study also proves that incorporating PSO feature selection before applying Random Forest classification leads to improved accuracy. Additionally, the research shows a significant improvement in the performance of Random Forest classification, especially with a composition of 90:10, which achieves the highest accuracy of 92.19%.

IV. CONCLUSION

In conclusion, the utilization of Particle Swarm Optimization feature selection in the Random Forest method, as well as the composition of data splitting, can indeed influence the classification outcomes. This is evident through observed changes in accuracy values. The experiment employed three distinct data comparison scenarios, leading to significant differences in accuracy values. Nevertheless, the accuracy value of the Random Forest method without feature selection can still be considered commendable, even though it does not yield as satisfactory results as when the Particle Swarm Optimization feature selection is applied. In this research, the highest accuracy value was obtained with a 90:10 data distribution. The Random Forest method without the Particle Swarm Optimization feature selection achieved an accuracy value of 87.50%. However, when combined with the Particle Swarm Optimization feature selection, the Random Forest method exhibited an improved accuracy value of 92.19%, indicating an increase of 4.69% in accuracy. This research is limited to utilizing Random Forest as the chosen classification method and Particle Swarm

Optimization as the feature selection technique. Furthermore, the dataset used in this research is focused solely on a single source, collected manually. It is crucial to consider the impact of the chosen methods and datasets on the accuracy of the results. In future studies, exploring alternative classification methods and feature selection algorithms can be worthwhile to uncover performance variations on similar datasets.

REFERENCES

[1] N. Chamidah and R. Sahawaly, "Comparison Support Vector Machine and Naive Bayes Methods for Classifying Cyberbullying in Twitter," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 7, no. 2, p. 338, 2021, doi: 10.26555/jiteki.v7i2.21175.

[2] K. Chemnad, M. Aziz, S. B. Belhaouari, and R. Ali, "The interplay between social media use and problematic internet usage: Four behavioral patterns," *Heliyon*, vol. 9, no. 5, p. e15745, 2023, doi: 10.1016/j.heliyon.2023.e15745.

[3] H. Karayiğit, Ç. İnan Acı, and A. Akdağlı, "Detecting abusive Instagram comments in Turkish using convolutional Neural network and machine learning methods," *Expert Syst. Appl.*, vol. 174, no. January, 2021, doi: 10.1016/j.eswa.2021.114802.

[4] A. Rejeb, K. Rejeb, A. Abdollahi, and H. Treiblmaier, "The big picture on Instagram research: Insights from a bibliometric analysis," *Telemat. Informatics*, vol. 73, no. December 2021, p. 101876, 2022, doi: 10.1016/j.tele.2022.101876.

[5] M. Fortunatus, P. Anthony, and S. Charters, "Combining textual features to detect cyberbullying in social media posts," *Procedia Comput. Sci.*, vol. 176, pp. 612–621, 2020, doi: 10.1016/j.procs.2020.08.063.

[6] N. Yuvaraj, "Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification," *Comput. Electr. Eng.*, vol. 92, no. May, p. 107186, 2021, doi: 10.1016/j.compeleceng.2021.107186.

[7] M. Li, Q. He, J. Zhao, Z. Xu, and H. Yang, "The effects of childhood maltreatment on cyberbullying in college students: The roles of cognitive processes," *Acta Psychol. (Amst).*, vol. 226, no. January, p. 103588, 2022, doi: 10.1016/j.actpsy.2022.103588.

[8] M. F. López-Vizcaíno, F. J. Nóvoa, V. Carneiro, and F. Cacheda, "Early detection of cyberbullying on social media networks," *Futur. Gener. Comput. Syst.*, vol. 118, pp. 219–229, 2021, doi: 10.1016/j.future.2021.01.006.

[9] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Comput. Sci.*, vol. 181, pp. 605–611, 2021, doi: 10.1016/j.procs.2021.01.207.

[10] K. Yokotani and M. Takano, "Social contagion of cyberbullying via online perpetrator and victim networks," *Comput. Human Behav.*, vol. 119, no. January, p. 106719, 2021, doi: 10.1016/j.chb.2021.106719.

[11] Aldinata, A. M. Soesanto, V. C. Chandra, and D. Suhartono, "Sentiments comparison on Twitter about LGBT," *Procedia Comput. Sci.*, vol. 216, pp. 765–773, 2023, doi: 10.1016/j.procs.2022.12.194.

[12] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Syst. Appl.*, vol. 223, no. August 2022, p. 119862, 2023, doi: 10.1016/j.eswa.2023.119862.

[13] C. T. Hanni, "Analisis Sentimen Komentar Cyberbullying pada Media Sosial Instagram Menggunakan Metode Support Vector Machine (SVM)," 2021.

[14] N. N. Amir Sjarif, N. F. Mohd Azmi, S. Chuprat, H. M. Sarkan, Y. Yahya, and S. M. Sam, "SMS spam message detection using term frequency-inverse document frequency and random forest algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 509–515, 2019, doi: 10.1016/j.procs.2019.11.150.

[15] I. Afdhal, R. Kurniawan, I. Iskandar, R. Salambue, E. Budianita, and F. Syafria, "Penerapan Algoritma Random Forest Untuk Analisis Sentimen Komentar Di YouTube Tentang Islamofobia," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 1, pp. 49–54, 2022, [Online]. Available: http://ojs.serambimekkah.ac.id/jnkti/article/view/4004/pdf

[16] R. Azizah Arilya, Y. Azhar, and D. Rizki Chandranegara, "Sentiment Analysis on Work from Home Policy Using Naïve Bayes Method and Particle Swarm Optimization," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 7, no. 3, p. 433, 2021, doi: 10.26555/jiteki.v7i3.22080.

[17] K. Setiawan, R. Beni, Burhanuddin, A. Budi Paryanti, and F. Fauzi, "KOMPARASI METODE NAIVE BAYES DAN SUPPORT VECTOR MACHINE MENGGUNAKAN PARTICLE SWARM OPTIMIZATION UNTUK ANALISIS SENTIMEN MOBIL ESEMKA JISAMAR (Journal of Information System , Applied , Management , Accounting and Researh) p-ISSN : 2598-8700 (Printed) J," *J. Inf. Syst. Applied, Manag. Account. Res.*, vol. 4, no. 3, pp. 102–111, 2020, [Online]. Available: http://journal.stmikjayakarta.ac.id/index.php/jisamarTelp.+62-21-3905050

[18] A. Thakkar and K. Chaudhari, "Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks," *Appl. Soft Comput. J.*, vol. 96, p. 106684, 2020, doi: 10.1016/j.asoc.2020.106684.

[19] M. Liang and T. Niu, "Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs," *Procedia Comput. Sci.*, vol. 208, pp.

460–470, 2022, doi: 10.1016/j.procs.2022.10.064.

[20] P. H. Prastyo, R. Hidayat, and I. Ardiyanto, "Enhancing sentiment classification performance using hybrid Query Expansion Ranking and Binary Particle Swarm Optimization with Adaptive Inertia Weights," *ICT Express*, vol. 8, no. 2, pp. 189–197, 2022, doi: 10.1016/j.icte.2021.04.009.

[21] F. Han, W. T. Chen, Q. H. Ling, and H. Han, "Multi-objective particle swarm optimization with adaptive strategies for feature selection," *Swarm Evol. Comput.*, vol. 62, no. January, p. 100847, 2021, doi: 10.1016/j.swevo.2021.100847.

[22] E. P. Saputra, S. Nurajizah, M. Maulidah, N. Hidayati, and T. Rachman, "KOMPARASI MACHINE LEARNING BERBASIS PSO UNTUK PREDIKSI TINGKAT KEBERHASILAN BELAJAR BERBASIS E-LEARNING COMPARATION OF PSO-BASED LEARNING MACHINE FOR E-LEARNING-BASED," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 2, pp. 321–328, 2023, doi: 10.25126/jtiik.2023106469.

[23] H. Azimi, H. Shiri, and M. Mahdianpari, "Iceberg-seabed interaction analysis in sand by a random forest algorithm," *Polar Sci.*, vol. 34, no. March, p. 100902, 2022, doi: 10.1016/j.polar.2022.100902.

[24] A. K, D. N, D. T, B. R. B B, B. D. N, and N. V, "Effect of multi filters in glucoma detection using random forest classifier," *Meas. Sensors*, vol. 25, no. October 2022, p. 100566, 2023, doi: 10.1016/j.measen.2022.100566.

[25] N. Rtayli and N. Enneya, "Selection features and support vector machine for credit card risk identification," *Procedia Manuf.*, vol. 46, pp. 941–948, 2020, doi: 10.1016/j.promfg.2020.05.012.

[26] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Meas. Sensors*, vol. 25, no. October 2022, p. 100605, 2023, doi: 10.1016/j.measen.2022.100605.

[27] J. Santos-Pereira, L. Gruenwald, and J. Bernardino, "Top data mining tools for the healthcare industry," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 4968–4982, 2022, doi: 10.1016/j.jksuci.2021.06.002.

[28] I. Kurniawan, D. Cahya, P. Buani, W. Apriliah, and R. A. Saputra, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 2, pp. 421–428, 2023, doi: 10.25126/jtiik.202396225.