Bit (Fakultas Teknologi Informasi Universitas Budi Luhur) p-ISSN: 1693-9166
Vol. 19, No. 1, April 2022, hlm. 29-36 e-ISSN: 2685-127X

PREDIKSI CYBERBULLYING SEBAGAI ALAT KONSELING CYBER DENGAN DATA MINING CLASSIFICATION

Agus Pamuji[1*], Heri Satria Setiawan[2]

[1]Faculty of Usuluddin, Adab and Da'wa, IAIN Syekh Nurjati Cirebon
[2] Faculty of Engineering and Computer Science, University of Indraprasta PGRI
Email: [1*]agus.pamuji@syekhnurjati.ac.id, [2]herisatria@unindra.ac.id

(Naskah masuk: 10 Maret 2022, diterima untuk diterbitkan: 31 Maret 2022)

Abstrak

Pertumbuhan data dan informasi semakin pesat sedangkan pengguna perangkat teknologi informasi terus meningkat. Apalagi kemudahan akses informasi didukung dengan hadirnya teknologi komunikasi bergerak yang dimiliki oleh hampir setiap pengguna. Saat ini terjadi peningkatan jumlah pengguna yang signifikan, yang dinilai berpeluang untuk hadirnya kejahatan dunia maya, terutama dalam kasus bullying. Masalah utama adalah sulitnya memprediksi potensi *cyberbullying* karena dunia maya penuh dengan anonimitas. Kasus *cyberbullying* termasuk kejahatan dunia maya di luar keamanan komputer jika dilihat dari perspektif perilaku. Oleh karena itu, cyberbullying menganalisis perilaku ketika melakukan tindakan negatif. Dalam studi ini, kami telah menyelidiki dan memprediksi kecenderungan bullying menggunakan pendekatan data mining. Analisis studi kasus yang disajikan dan teknis pelaksanaannya melalui pendekatan data mining. Menggunakan data mining sebagai alat, kami menggunakan beberapa metode klasifikasi dengan tujuan perbandingan dan metode mana yang terbaik untuk analisis. Metode klasifikasi dalam data mining adalah K-NN, *Random Forest*, *Decision Tree*, dan *Naive Bayes*. Dengan perbandingan dari empat metode, ada tiga kelas. Ada tiga kelas, yaitu tidak ada potensi *bullying*, kekerasan dan hinaan. Dengan demikian, tiga kelas terdeteksi berdasarkan hasil investigasi dan beberapa teknik dievaluasi dengan menggunakan perbandingan kinerja masing-masing teknik. Hasil akhir akan menunjukkan teknik *decision tree* berkinerja terbaik dikarenakan selama tahap *preprocessing* tidak membutuhkan waktu lama. tambahannya adalah karena kemampuan membreakdown setiap cabang dari data sehingga waktu yang dibutuhkan pada data cleaning lebih cepat dengan akselerasi 7%.

Kata kunci: *Penindasan dunia maya, Penambangan Data, Konseling Siber, Media Sosial, Teknologi Informasi*

CYBERBULLYING PREDICTION AS CYBER COUNSELING TOOLS WITH DATA MINING CLASSIFICATION

Abstract

The growth of data and information is increasing rapidly while the users of information technology devices continue to increase. Moreover, the ease of access to information is supported by the presence of mobile communication technology which is owned by almost every user. Currently, there is a significant increase in the number of users, who are considered to have the opportunity for the presence of cybercrimes, especially in cases of bullying. The main problem is that it is difficult to predict the potential for cyberbullying because cyberspace is full of anonymity. Cases of cyberbullying include cyber crimes outside of computer security when viewed from a behavioral perspective. Therefore, cyberbullying analyzes behavior when carrying out negative actions. In this study, we have investigated and predicted bullying tendencies using a data mining approach. Analysis of the presented case studies and their technical implementation through a data mining approach. Using data mining as a tool, we use several classification methods with the aim of comparison and which method is best for analysis. Classification methods in data mining are K-NN, Random Forest, Decision Tree, and Naive Bayes. By comparison of the four methods, there are three classes. There are three classes, namely no potential for bullying, violence and insults. Thus, three classes were detected based on the results of the investigations and several techniques were evaluated using a comparison of the performance of each technique. The final result will show the best performing decision tree technique because during the preprocessing stage it does not take long. the addition is because of the ability to break down each branch of the data so that the time required for data cleaning is faster with an acceleration of 7%.

Keywords*: Cyberbullying, Data Mining, Cybercounseling, Social Media, Information Technology*

1. INTRODUCTION

Today, communication is being carried out by humans by utilizing internet technology reinforced through cellular technology, computer hardware and software and telephone networks [1]. Mobile technology seems to have the role of directing information that is considered to have power allowing it to be distributed to others through devices behind the scenes. The reality at this time, the amount of information that contains data is increasing as the number of technology users also increases. An increase in both the phenomenon of information and users has the potential for crime or misconduct in the cyber environment. In the beginning, every user is able to study, deepen, every feature of technology to find weaknesses. With many weaknesses found, users of cyber technology begin to intend and act to carry out malicious actions. One of his immoral actions is cyberbullying. From the start, comes harassment, cyberbullying, hate speech, and online trolling. This phenomenon becomes more and more brutal and can result in extreme losses. In the same sense, cyberbullying is described as an act of intimidation that transpires using technologies behind the scenes. In addition, what underlies the concept of cyberbullying is that traditional bullying includes (a) the intention to cause harm, (b) repetition of behavior from time to time, (c) power imbalance between the victim and the bully [2].

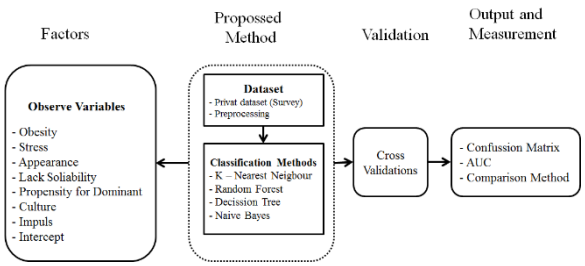
Internet technology is faced with positive impacts and has advantages and is assisted by computer networks spread throughout the world. One of the advantages is that it allows individuals to search, explore all information. The fact must be admitted that bullying is almost done on social media platforms. Social media such as Twitter, Facebook, Instagram and so on. Social media opens up opportunities and freedom for every user to share information ranging from text, photos, videos and so on. Almost all of them are not realized, if there is a possibility or a dangerous action will occur. His concern is that social media is a place and target for finding someone who will be the target of crime, one of which is cyberbullying [3]. Negative actions in cyberspace such as insulting, harassing, ridiculing and even threatening the victim. Since social media is considered trigger-prone, it can connect relationships between individuals, groups and communities.

The increase in cyberbullying actions is associated with the study of data mining concepts. Investigation of data and information about cyberbullying is a strength in addition to the tasks and roles of data mining. One of the advantages of data mining is making predictions based on data. Internet user data and related to bullying as modal and object in this analysis. Through data mining, not only as a predicting concept, but also as a way to see patterns of tendencies towards bullying in cyberspace. Cyberbullying actions can be anticipated with user sentiment analysis. The studies that have been carried out have focused on the issue of sentiment. Sentiment analysis has two types, namely positive sentiment analysis and negative sentiment analysis. Positive sentiment analysis may be harmless compared to negative sentiment analysis albeit in a subtle way. Sentiment analysis is very suitable to be applied to data collection on social media [4]. Thus the concept of data mining with sentiment analysis can be classified into positive and negative sentiments [5].

In this research work, we focus on harassment, where the harasser (bully) sends toxic and harmful communications to the victims. We have presented an automated, data-driven method for identifying harassment. Our approach using data mining can help researchers with the aim of exploring data or extracting knowledge. The results of this knowledge will be a tool in online counseling pursuits or cyber-counselling. Data mining can provide opportunities to effectively predict and detect forms of negative human behavior such as bullying. Because this work deals with data, big data analysis and data mining can reveal knowledge through data mining of raw data. Both data mining and big data analysis can improvise some applications and predict future possibilities[6].

2. METHOD

Almost all forms of cyberbullying are carried out on social media. Bullying in the concept of data mining can be done by prediction, classification, association and so on. The role of data mining in the investigation of negative acts of bullying has also been carried out in the form of modeling. Classification techniques are also widely applied to analyze cyber bullying, one of which is using the Naive Bayes technique. Several analytical techniques in data mining can be compared included in this paper. For example in the classification method, several classification techniques are compared, including K-Nearest Neighbor, Random Forest, Decision Tree, and Naive Bayes. Prediction of cyber bullying is done using the K-Nearest Neighbor technique but focuses on positive comments even though there are potential negative comments [10].



Gambar 1. Proposed Research Framework

Predicting the potential for cyberbullying that occurs in cyberspace requires techniques and data analysis regarding data-based work patterns. In this case, we have identified the factors that can be taken into consideration in addition to the variables. Some

factors are converted to class when using the concept of data mining, namely that there is no potential for bullying, violence and insult (Rosa 2019). Furthermore, there are several features involved in the provided dataset so that feature selection is required. See the picture below [11], a research framework that shows the provision of a specific dataset related to cyberbullying using survey techniques.

Collecting data by observing cyber counseling cases on personal datasets. Although almost all the literature recommends the dataset used is public, especially on social media platforms. Internet users as well as counseling services are given the opportunity to become respondents. We used a quantitative approach in data collection. Likert scale was adopted in completing the needs of data requests and observations. The purpose of the personal dataset is that it is known that the dataset in this cyber counseling study has a high level of sensitivity, which is different from other datasets, only partially based on comments on social media applications. The questionnaire is designed to make it easier for victims of cyberbullying by ensuring a high level of privacy. The identity of the respondents is ensured confidentiality and there is even the option of an anonymous account. This condition becomes more complex, on the other hand, with anonymous data, it will affect the validity of the data. Thus, the survey method is an option to explore information on victims of cyberbullying [12].

Based on the proposed framework, preprocessing data is required prior to the analysis of the case study. There are four tasks in data preprocessing starting at data cleaning. The data in the dataset is cleaned of formats that are not in accordance with the request at the time of observation. The addition is manual data entry if data is found to be incomplete. Next is data reduction, which simplifies complex data into integrated and simple data. Simplifying the meaning of the data in the dataset allows ease of data mining analysis. Another condition is to unify data from various sources because the data is not all in one place and the same container. Thus, this unification process is called data integration. In the data preprocessing phase, it will take a lot of time and resources due to the demands of producing quality data [13].

Because it uses a data mining approach and uses data classification analysis techniques, several steps must be taken. First, entering data and generating a dataset can be used as a pre-processing stage. The selection of this dataset technique is private, although the concept of data mining in general, it uses public datasets. Second, the data analysis process uses algorithms and classification techniques including KNN, RF, DT and NB. Third, the final result of each technique and algorithm will be compared working to the Confusion matrix, and ROC Curve.

2.1 Decision Tree

Classification by extracting decision trees on the movement of data from top to bottom. Labeling with features applies to every node in the decision tree. Classification based on decision trees with expectations produces a model in predicting the value of candidate variables when studying simple decision rules inferred from data features. In addition, the decision tree technique belongs to the supervised machine learning method and the non-parametric category. Iterative Dichotomiser 3 (ID3) algorithm is well known in this type of decision tree algorithm. The ID3 algorithm uses the calculation of the acquisition of entropy information for the selection of attributes to be nodes. The parameterization of an entropy can be done on the magnitude of the value S considered in the item set so that it includes the value, whereas we define n to be how many parts refer to S and pi if the proportion is the case to m in the entity set. Furthermore, the formation of parameter variables on gain can be analyzed through a case set of S. B is the attribute, n is the number of partitions of attribute A, |si| is the case with the notation n as an additional attribute where n denotes the number of parts of attribute A, the proportion of Si to S,|S| is the number of cases in S-i, and entropy(Si) is the entropy of the sample designation that has the -i value.

$$Entropy\ (S) = \sum_{i=i}^{n} -Pi * log\ log\ 2p_i \ \ldots\ldots\ (1)$$

$$Gain\ (S, A) = \ Entropy\ (S) \sum_{i=1}^{n} \frac{|S_i|}{|S|} * \ Entropy\ (S_i) \ \ldots\ (2)$$

2.2 K-Nearest Neighbour

The classification model in machine learning that uses the nearest neighbor point is K-Nearest Neighbor. Machine learning methods as well as data mining techniques that aim to label previously invisible query objects and distinguish two or more destination classes. Such a classifier, in general, requires some training data with a given label. The K-NN method uses a similarity technique between X1 objects and X2 and Xn objects. Thus using the equation X where d(x,y) is the distance between data x to data y. Xi is the i-th test data and yi is the i-th training data.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \ \ldots\ldots\ldots\ldots\ldots\ (3)$$

2.3 Naïve Bayes

The classification technique in the comparative study of cyberbullying predictions is Naive Bayes. With NB performance through bayesian theory and data-based processing. Bayesian overload predicts how likely the probability of a set of items in a group is. Probability results with posterior P(H|X) using the Bayesian proposed concept. We determined the value of X by describing the data on the group test when it was unknown. The hypothesis based on the data flow generated from X becomes more specific as part of

the value of H. The probability of the hypothesis on x is assumed by selecting the value of H so that it is considered the likelihood of the value of P(X|H). The next consideration is prior probability as an opportunity to become a hypothesis H with a P(H) value. Predictor with probability generalization is part of X . value.

$$P(X) = \frac{P(H).P(H)}{P(X)} \quad \text{..... (4)}$$

2.4 Random Forest

Data mining analysis techniques such as Random Forest is one of the classification techniques. A large amount of data fits this technique very well with its accuracy. Moreover, the Random Forest is an excellent technique when doing predictions and also easy to learn in a variety of academics and professionals [14].

$$mg(X,Y) = h_k(h_k(X) = Y) - max_{avk} I(h_k(X) = j) \ldots\ldots (5)$$

2.5 Theoretical Review

Cyberbullying may remain for some time and involve many users of cyberspace. Many researchers do their work with data retrieval platforms. Bullying can result in conditions where participants can be classified as perpetrators or bully-victim even though it is done in cyberspace. For a more point by point, the study of cyberbullying invites a lot of attention from researchers, academics, and practitioners in conducting investigations, analyzing the causative factors, the pattern of delivering messages when bullied by bullies. Data mining is used to extract all data related to the activities of using social media. The implementation of data mining techniques has functionally predicted the presence of intimidation. Data mining techniques can predict with algorithms, namely K-Nearest Neighbor, Naive Bayes, and Decision Tree. In addition to data mining algorithms, a combination of quantitative and qualitative methods is proposed by Elaheh Raisi which uses the concept of machine learning with a weakly supervised approach [7]. Detection of potential cyberbullying actions can still be done using a soft computing approach, although almost all of them are based on data mining. Data mining has also been proposed to predict careers related to counseling services by Yeasin Arafat. The consequences of his research provide predictions about the weaknesses and strengths that students have for a career. There are 9 features proposed to recognise temporary career predictions using classification algorithms . Thus the model obtained is CART with the highest accuracy value. Rizki proposes the use of the support vector machine method when analyzing sentiment related to cyberbullying. In the test results, using as much as 100 data on Twitter. The measurement results show that it is quite effective with the value of the confusion matrix with an accuracy rate of 70% valid [8].

Data mining studies have challenges, which are also applied to predict the factors that cause cyberbullying in Korea [9]. According to the results, victims occupy the most cases compared to perpetrators and observers. The Decision tree method independently succeeded in finding the pattern of factors that influence it. Data mining analysis was also proposed by Bashir with the concept of analyzing different Arabic texts previously in the Korean context. Arabic has a high complexity [10]. The concept that is built is only a protection not to predict. Sultan proposed theoretical methods, namely analysis, synthesis, and empirical. Not only in data mining but adopting machine learning and deep learning. Data mining is different from machine learning, machine learning focuses on strengthening training data so that it can predict the incoming data is considered as testing[11].

The objectivity of this study is to classify the types of cyberbullying, namely perpetrators and victims through sentiment analysis. In this study, we will have examined a dataset related to cyberbullying activities using four data mining classification analysis techniques. In this way, we utilize data mining techniques such as Naive Bayesian, Decision Tree, K-Nearest Neighbor (KNN), and Random Forest to confirm the prediction model on the type of cyberbullying in social media as a cyber counseling tool [9].

3. RESULT AND DISCUSSION

3.1 Preparation of Data

The preparation stage includes the use and processing of datasets. Datasets serve as objects and materials for analysis when there is a potential for excessive users of various files. Furthermore, there are two types of data mining studies when using datasets, namely private datasets and public datasets. Considerations on the dataset there are two types of variants through data mining including private datasets and public datasets [15]. Both datasets are types of data sources that are managed during the analysis. The use of private datasets is a rare activity due to almost all discussion on public datasets. General datasets are readily available on the website, where all users are free to download, extract and experiment. Personal dataset is a collection of data collected specifically by survey methods in addition to quantitative methods. As such, personal datasets are generally applied to and extracted from those of a privacy and institutional nature. In addition, private datasets have characteristics adapted to specific case studies. In contrast to public datasets, data that is already on the website is provided by a certain institution for experimental purposes. In concept, public datasets can be analyzed in terms of comparability.

Data preparation stage, identify the data cleaning. The dataset used shows incomplete (missing) data where data is not available on some records. When data is found to be incomplete, it can

be resolved by filling in the attribute values manually or automatically. The next problem is noise data which contains inappropriate or incorrect attributes. In addition, another problem is the duplication of data as much as 5% of the total dataset. Noisy data can be overcome by grouping and deleting incorrect or invalid values [16].

Data preparation is the first step in generating data mining processing. The dataset will meet the criteria with regularity, not experiencing duplication and so on. In our dataset, we have identified 4% containing duplicate data, 1% empty data and so on. According to the identification results, there are 4% of the data containing duplication, the remaining 2% of the data is empty. Furthermore, the validation process is implemented if after passing the data preparation stage it goes through 10 iterations. The method used is cross-validation as a recommendation by proving reliability in testing the validity of the data. In the concept discussed, the dataset consists of training data and test data in step n as many as 10. The results of the 10-step validation process can be shown in the figure below.

Gambar 2. Correct and Incorrect Classification

3.2 Validation and Model Testing

In order to obtain accurate results for data modeling, the model needs to be validated. The use of the validation technique was tested 10 times. Thus the measurement result is the average value of the 10 tests. This stage will be divided into two parts, namely training data and test data. Tests were carried out 10 times, namely 10 Fold Cross Validation which can be used as recommendations in selecting the best model.

3.3 Comparison of Model

There are several methods with classification capabilities on the concept of data mining. The classification method includes a discussion of excessive users in the case of file sharing. Classification methods including decision trees, K-Nearest Neighbor, Naive Bayes, and Random Forest are applied to the dataset by conducting experiments to determine which method is considered the best based on the analysis. The selection of this method becomes the basis of the tool when detecting excessive users in managing file sharing. The next explanation shows the comparison according to the quality test of the proposed model. Further backed by private datasets such as redundant file sharing data. Furthermore, the Confusion Matrix is carried out in analyzing the level of accuracy of several classification methods in data mining. In addition, the results achieved according to measurements and analysis with the Confusion Matrix were analyzed using the ROC curve with information expectations of differences between several methods or models that have been tested in the previous stage.

Table 1. Cross validation test results.

Testing	Initial Data	Final Data	Valid	Invalid	Acc True (%)	Acc False (%)
1	256	30	27	3	85,12	14,88
2	256	30	27	3	88,26	11,74
3	256	30	24	6	83,12	16,88
4	256	30	28	2	87,06	12,94
5	256	30	24	6	86,63	13,37
6	256	30	25	5	88,37	11,63
7	256	30	24	6	88,12	11,88
8	256	30	27	3	87,67	12,33
9	256	30	26	4	86,63	13,37
10	256	30	28	2	88,37	11,63

The TP level can be shown on the Y axis while the FP level is determined on the X axis which generalizes the ROC curve. Implementation using the ROC curve method by analyzing the coverage area under the normal value of the curve that refers to the AUC (Area Under the ROC Curve). Furthermore, it is determined by graphically representing the output of determining the best performance when classifying. Next, how to measure performance through predictive outcomes on the probability side of the discriminatory. This is obtained by means of a random sample of the population mean. The main indication is that the AUC value with continuous addition will determine the strength of the classification. The value range between 0.0 and 1.0 is the range of values in the AUC area to be the best determining criterion. Table 3 refers to how to analyze the classification accuracy level through the AUC method.

The data mining work presented in the table, through decision tree techniques and algorithms can be considered the best. The results of the discovery and identification were found in 89.42%. Moreover,

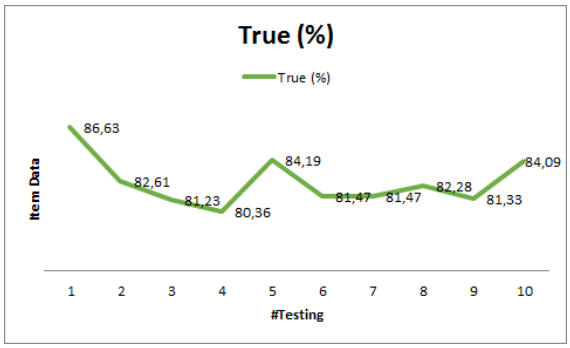
if it is based on another algorithm. The best performance is neighbor detection with performance at 88.07%, then performance based on probability is 85.13% and the last is 72.17%. The results of this measurement use a confusion matrix in the level of model accuracy. The ROC curve test shows that the Naive Bayes Algorithm is the best algorithm with an AUC value of 0.853, K-Nearest Neighbor 0.813, Decision Tree 0.816 and Random Forest 0.811. The measurement results on the level of model accuracy are related to the implementation of the confusion matrix. It must be emphasized that the ROC curve shows and brings the best performance of the Naive Bayes algorithm with an AUC value of 0.853, compared to other algorithms. The main factor is the determination of training data with a large amount but not like a decision tree. even with big data but it can be done quickly because of the breakdown technique.

Table 2. ROC category determination criteria.

Performance	Decision
0,90 – 1,00	Very Strong Classification
0,80 – 0,90	Strong Classification
0,70 – 0,80	Balance Classification
0,60 – 0,70	Less Classification
0,50 – 0,60	Not Identification

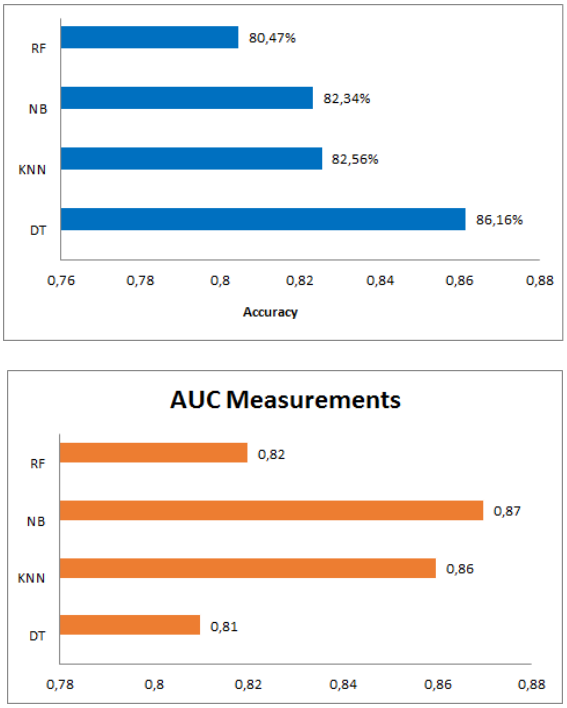
Table 3. Performance comparison on model accuracy and AUC.

Evaluations	Decision Tree	K-Nearest Neighbour	Naive Bayes	Random Forest
Accuracy	89,42 %	88,07 %	85,13 %	72,17 %
AUC	0,816	0,813	0,853	0,811



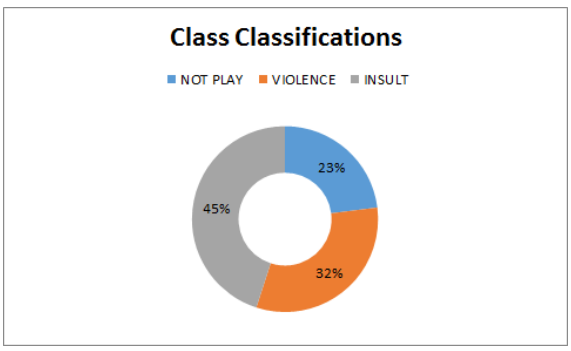
Gambar 3. True Data Testing Cross Validation

In this case, several data mining classification methods have made the results the best performance value. Evaluation of the model or classification method is carried out in order to see the best assessment. The first criterion consideration is accuracy. Accuracy value data for each model is presented in Figure 4 below. The DT algorithm has a better performance value than KNN et al. Thus, an algorithm based on parsing tree-like data can be considered to have a stable value when validation is carried out in the previous stage. Others, such as NB, have smaller values than KNN and DT, so DT can be used as a recommendation.



Gambar 4. Model Evaluations

Although NB is considered to have inadequate performance, when tested using AUC, it is much better with a gain value of 0.87, surpassing KNN with a gain value of 0.86 and the DT algorithm with a value of 0.81.



Gambar 5. Class Classifications

The whole series of processes for data mining processes when predicting cyberbullying actions can be presented in the three classes that are candidates for cyberbullying predictions. In general, cyberbullying can be caused by insults and even almost half of the dataset is filled. The potential for insults often appears and becomes a tool to launch action against the victim. Increased insults when bullying is also followed by violence. Most researchers consider it the peak of wanting to do bullying when a sense of dissatisfaction is present in

the perpetrator towards the victim. The Violence and insults when they dominate, then based on the dataset identified the potential for not bullying, even though there are some features that have attribute values, they are said to be of medium and low rank. Thus, every attribute that has been identified and analyzed that has a low attribute value cannot be considered as an act of bullying

4. CONCLUSION

The role of data mining is very large, especially in extracting data for the purposes of conducting analysis. Not only that, data mining which is so super in identifying each data in detail provides an overview of the data that is processed and has its own criteria. For more details, raising the case of cyberbullying in this study is a tool for researchers other than counselors in providing face-to-face or online counseling services. In this way, counselors who carry out activities in cyber counseling need to be provided with additional tools with predictive data mining results. Given that cyberbullying is extreme, data mining has a strong role in supporting cyber counseling services.

Data mining becomes a reliable tool when it is intended for counselors or researchers. Working with data makes results more objective. The advantages of data mining make predictions more accurate than other methods or approaches. However, it should be noted that the dataset should be prioritized at the data preparation stage. The main reason is that the dataset determines the results before they are analyzed and needs to be processed and ready.

REFERENCES

[1] A. Kumar and N. Sachdeva, "Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data," *Multimed. Syst.*, vol. 2, no. 0123456789, 2020, doi: 10.1007/s00530-020-00672-7.

[2] D. Soni and V. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," in *Proceedings of the ACM on Human-Computer Interaction*, 2018, vol. 2, no. CSCW, pp. 1–25, doi: 10.1145/3274433.

[3] A. Akhter, K. A. Uzzal, and M. M. A. Polash, "Cyber Bullying Detection and Classification using Multinomial Naïve Bayes and Fuzzy Logic," *Int. J. Math. Sci. Comput.*, vol. 5, no. 4, pp. 1–12, 2019, doi: 10.5815/ijmsc.2019.04.01.

[4] H. Rosa *et al.*, "Automatic cyberbullying detection: A systematic review," *Comput. Human Behav.*, vol. 93, no. 3, pp. 333–345, 2019, doi: 10.1016/j.chb.2018.12.021.

[5] H. Gaffney, D. P. Farrington, D. L. Espelage, and M. M. Ttofi, "Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review," *Aggress. Violent Behav.*, vol. 45, no. June, pp. 134–153, 2019, doi: 10.1016/j.avb.2018.07.002.

[6] R. M. Kowalski, S. P. Limber, and A. McCord, "A developmental approach to cyberbullying: Prevalence and protective factors," *Aggress. Violent Behav.*, vol. 45, no. 2017, pp. 20–32, 2019, doi: 10.1016/j.avb.2018.02.009.

[7] J. Wang, K. Fu, and C. T. Lu, "SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection," *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, vol. 3, no. 12, pp. 1699–1708, 2020, doi: 10.1109/BigData50022.2020.9378065.

[8] A. Bozyiğit, S. Utku, and E. Nasibov, "Cyberbullying detection: Utilizing social media features," *Expert Syst. Appl.*, vol. 179, no. April, pp. 1–15, 2021, doi: 10.1016/j.eswa.2021.115001.

[9] Song, T. M., & Song, J., ". Prediction of risk factors of cyberbullying-related words in Korea: Application of data mining using social big data," *Telematics and Informatics.*, vol. 3, no. 1, pp. 12–28, 2021,

[10] Bashir, E., & Bouguessa, M. Data Mining for Cyberbullying and Harassment Detection in Arabic Texts," *International Journal of Information Technology and Computer Science.*, vol. 13, no. 5, pp. 41–50, 2021,

[11] Sultan, D., Suliman, A., Toktarova, A., Omarov, B., Mamikov, S., & Beissenova, G., "Cyberbullying Detection and Prevention: Data Mining in Social Media," In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2021, pp. 338–342.

[12] D. Sultan, A. Suliman, A. Toktarova, B. Omarov, S. Mamikov, and G. Beissenova, "Cyberbullying detection and prevention: Data mining in social media," in *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, 2021, pp. 338–342, doi: 10.1109/Confluence51648.2021.9377077.

[13] M. A. Al-Ajlan and M. Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning," in *21st Saudi Computer Society National Computer Conference, NCC 2018*, 2018, pp. 1–5, doi: 10.1109/NCG.2018.8593146.

[14] P. S. Raj and G.Silambarasan, "Consumer Behaviour Marketing Analysis Using Data mining Concepts," *Int. J. Comput. Tech.*, vol. 5, no. 2, pp. 40–43, 2018, [Online]. Available: http://www.ijctjournal.org.

[15] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," in *SIAM International Conference on Data Mining, SDM 2019*, 2019, pp. 235–243, doi: 10.1137/1.9781611975673.27.

[16] C. Iwendi, G. Srivastava, S. Khan, and P. K. R.

Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimed. Syst.*, vol. 3, no. 23, pp. 1–15, 2020, doi: 10.1007/s00530-020-00701-5.

[17] R. Zhu, W. Guo, and X. Gong, "Short-term photovoltaic power output prediction based on k-fold cross-validation and an ensemble model," *Energies*, vol. 12, no. 7, 2019, doi: 10.3390/en12071220.

[18] Z. Y. Algamal, "Shrinkage parameter selection via modified cross-validation approach for ridge regression model," *Commun. Stat. Simul. Comput.*, vol. 49, no. 7, pp. 1922–1930, 2020, doi: 10.1080/03610918.2018.1508704.

[19] Q. C. Song, C. Tang, and S. Wee, "Making Sense of Model Generalizability: A Tutorial on Cross-Validation in R and Shiny," *Adv. Methods Pract. Psychol. Sci.*, vol. 4, no. 1, 2021, doi: 10.1177/2515245920947067.