Improving cyberbullying detection through multi-level machine learning

Salsabila, Riyanarto Sarno, Imam Ghozali, Kelly Rossa Sungkono
Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Article Info

Article history:

Received Sep 9, 2023
Revised Oct 11, 2023
Accepted Nov 29, 2023

Keywords:

Cyberbullying
Deep learning
Machine learning
Multi-class multi-level
Text classification

ABSTRACT

Cyberbullying is a known risk factor for mental health issues, demanding immediate attention. This study aims to detect cyberbullying on social media in alignment with the third sustainable development goal (SDG) for health and well-being. Many previous studies employ single-level classification, but this research introduces a multi-class multi-level (MCML) algorithm for a more detailed approach. The MCML approach incorporates two levels of classification: level one for cyberbullying or not cyberbullying, and level two for classifying cyberbullying by type. This study used a dataset of 47,000 tweets from Twitter with six class labels and employed an 80:20 training and testing data split. By integrating bidirectional encoder representations from transformers (BERT) and MCML at level two, we achieved a remarkable 99% accuracy, surpassing BERT-based single-level classification at 94%. In conclusion, the combination of MCML and BERT offers enhanced cyberbullying classification accuracy, contributing to the broader goal of promoting mental health and well-being.

Corresponding Author:

Riyanarto Sarno
Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
Email: riyanarto@if.its.ac.id

1. INTRODUCTION

The impact of cyberbullying can be more dangerous than traditional bullying because access to disseminate information that contains crime can be easier and faster by using technology [1]. Cyberbullying, encompassing online threats, harassment, and humiliation, stands as a significant risk factor for mental health issues such as depression, anxiety, and even suicide [2]–[5]. A study conducted by the Pew Research Center reveals that 40% of internet users have encountered cyberbullying [6]. The proliferation of digital technology has accelerated the spread of harmful content through social media, exacerbating the prevalence of cyberbullying. Furthermore, the lack of a clear-cut definition of cyberbullying has compounded the complexity of the issue. Consequently, the core problem revolves around the accurate identification and mitigation of diverse manifestations of cyberbullying that may surface on social media platforms.

To address this challenge, several previous studies [7]–[13] have endeavored to detect cyberbullying using machine learning (ML) and deep learning (DL) methodologies. However, many of these studies still rely on a single-level classification approach, consolidating various forms of cyberbullying into a single level or category. This research seeks to transcend this limitation by introducing the multi-class multi-level (MCML) approach for classifying cyberbullying at multiple levels. The MCML approach, while typically applied to image data classification, offers a relatively novel application to text data. In this framework, multi-class categories are subdivided into several subclasses, enhancing the precision of classification and bolstering accuracy [14]. Before implementing the MCML algorithm, this study conducts a preliminary single-level classification employing various DL and ML methodologies, including bidirectional long-short term memory (Bi-LSTM), bidirectional encoder representations from transformers (BERT), multinomial naïve Bayes (NB), Bernoulli NB, logistic regression (LR), support vector machine (SVM), and random forest (RF). The accuracy results obtained with DL and ML methods at level one dictates the most suitable approach for multi-level classification. Subsequently, the machine learning algorithm assesses its performance by calculating various metrics, including precision, recall, F1-scores, and accuracy [15].

In summary, the main contribution of this study is the introduction of the MCML approach to enhance the detection of cyberbullying in text data. This innovative method subdivides multi-class into subclasses, improving precision and overall accuracy. Additionally, the research conducts a preliminary single-level classification using ML and DL techniques to determine the optimal approach for multi-level classification, evaluated through key metrics like precision, recall, F1-scores, and accuracy. Ultimately, this study aims to advance cyberbullying detection, thereby supporting mental health in the digital age.

2. METHOD

The proposed method, as depicted in Figure 1, comprises a series of sequential steps. It commences with data preprocessing, which encompasses essential tasks such as data cleaning, normalization, tokenization, and lemmatization. Following this, the target column is encoded to prepare it for classification algorithms. Subsequently, the dataset is split into three distinct sets: the training set, used for model development; the testing set, employed for model evaluation; and the validation set, which aids in model fine-tuning. In the initial phase of the classification process, single-level machine learning techniques are applied to the training set. The primary objective at this stage is to identify the best model from a pool of potential candidates. The best model is employed in the subsequent multi-level classification task. At the first level of multi-level classification, the model's primary aim is to distinguish between instances of cyberbullying and non-cyberbullying, addressing the core concern of identifying harmful online behavior. Moving to the second level, the model further enhances its classification capabilities instances based on aspects such as age, religion, ethnicity, and gender. This hierarchical approach enhances the cyberbullying detection system, providing a better understanding of both cyberbullying incidents and the associated aspects. Finally, the performance of the model is evaluated.

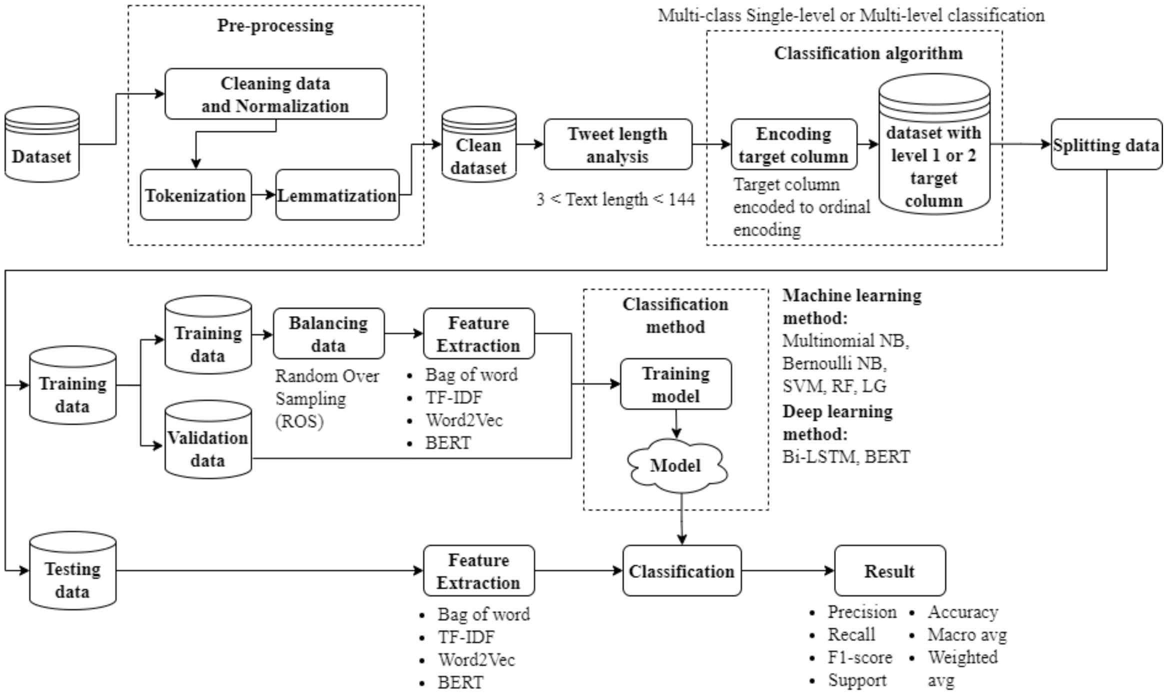


Figure 1. Multi-class multi-level approach

2.1. Dataset

This study utilizes the “cyberbullying_tweets” dataset, curated by Wang *et al.* [7] from Kaggle. This compilation combines six Twitter datasets (Bretschneider, Chatzakou, Waseem, Davidson, WISC, and Hate speech) with a focus on cyberbullying. The dataset comprises 47,000 tweets, featuring “tweet_text” and “cyberbullying type” columns. The dataset is categorized into six classes: age, ethnicity, gender, religion, and not cyberbullying. The dataset demonstrates balance, with approximately 8,000 entries per class, and the “other cyberbullying” class has been excluded to avoid confusion. For model development, an 80:20 split is employed, dividing the dataset into training and testing sets. After thorough data cleaning, the remaining 35,719 data have been utilized for both single-level and multi-level classification in the first level. In the context of multi-level classification, specifically at the second level, 29,869 data labeled as cyberbullying are further categorized into four classes: age, religion, gender, and ethnicity.

2.2. Preprocessing

A significant portion of the text in a dataset typically includes elements such as emojis, hashtags, punctuation marks, stop words, and more, which are often unnecessary. These unnecessary things can negatively impact system performance [16]. Therefore, preprocessing is essential to enhance accuracy [17]. The data preprocessing steps are detailed in Table 1.

Table 1. Steps preprocessing

Preprocessing
Start
1. Take a sentence from dataset
Cleaning data:
2. Removes duplicate data
3. Remove stop words using Python Natural Language Toolkit (NLTK)
4. Remove symbols and non-ASCII letters
5. Remove emojis
6. Convert abbreviations word to original form
Normalization data:
7. Converts all letters to lowercase
Tokenization:
8. Separated sentences into word units using NLTK
Lemmatization:
9. Converts the word into its original version in the dictionary using NLTK
Stop

2.3. Multi-class multi-level (MCML)

The MCML algorithm serves as a method for subdividing a multi-class into various sub-classes. In previous research, datasets featuring multiple labels, such as religion, age, ethnicity, gender, other cyberbullying, and non-cyberbullying, have predominantly been treated as singular categories. Despite this, it is plausible for religion, age, ethnicity, and gender to be regarded as specific forms of cyberbullying, falling within the cyberbullying class. Hence, this study undertakes a multi-level classification approach using the MCML algorithm.

Within this multi-class classification process, subclasses derived from a main class are assigned distinct classification levels. Initially, the classification centers on cyberbullying and not cyberbullying at the first level. Subsequently, at the second level, classifications are performed for subclasses of cyberbullying, specifically religion, age, ethnicity, and gender. Notably, the “other cyberbullying” subclass has been omitted to mitigate confusion. By adopting this multi-level classification algorithm, the classification process becomes more intricate, resulting in enhanced specificity and accuracy [14]. This algorithm is usually used in classifying image data. Figure 2(a) illustrates the application of the single-level, and Figure 2(b) demonstrates the application of multi-level multi-class classification algorithms in this case. At this stage, the algorithm for data classification will be applied. There are two classification algorithms used, namely single-level and multi-level classification. To apply the classification algorithm, the data in the sentiment column is replaced, from what was previously a string to a numeric form. After that, the number of classes is set as desired. Target class coding according to the classification algorithm can be seen in Table 2.

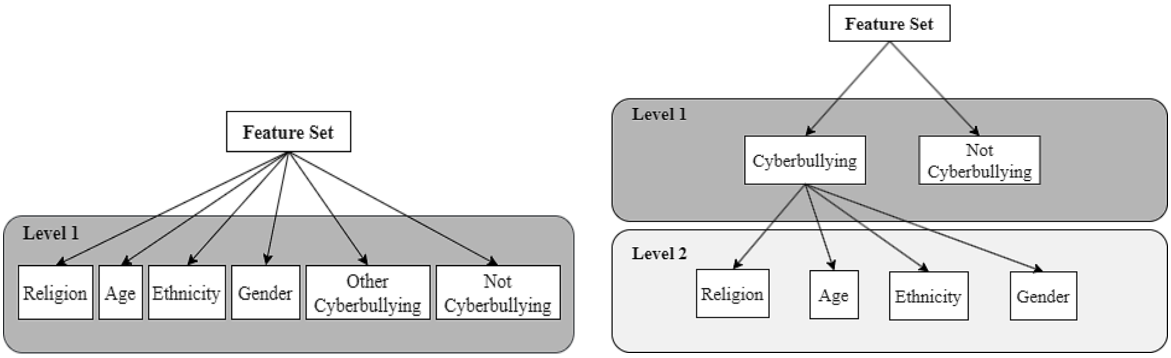
2.4. Balancing data

Balancing imbalanced datasets is crucial in machine learning [18]. There are two commonly used techniques to address this issue: undersampling and oversampling. Oversampling has proven to be highly effective in various machine learning problems [19]. The random over sampling (ROS) method is employed to rectify class imbalance by duplicating samples until an equal number of samples per class is attained [20].

ROS involves replicating minority class instances. The new minority training set is sampled from the original minority class randomly so that the minority class size is the same as the majority class size. The ROS equation can be seen in (1).

$$'|N'_{min}| = |N_{maj}|' \quad (1)$$

Equation (1) states that, to ensure a balanced dataset, the size of the minority class, denoted as $|N'_{min}|$, should be equal to the size of the majority class, represented as $|N_{maj}|$ [21]. After splitting the data, this study found that the training data is still imbalanced. Therefore, this study balanced the data using ROS. The dataset that has been balanced using ROS can be observed in Figure 3.



(a) (b)

Figure 2. Comparing classification using (a) multi-class single-level and (b) multi-class multi-level

Table 2. Encoding target class according to the classification algorithm

Classification	Class target encoding

Single level	Religion = 0
	Age = 1
	Ethnicity = 2
	Gender = 3
	Not cyberbullying = 5

Multi-level 1	Cyberbullying = 1 Not cyberbullying = 0

Multi-level 2	Religion = 0
	Age = 1
	Ethnicity = 2
	Gender = 3

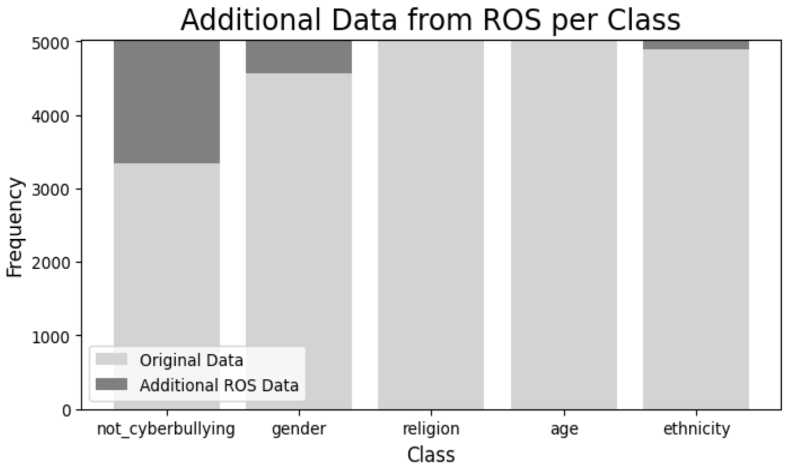


Figure 3. Data before and after balancing with ROS

2.5. Feature extraction

Feature extraction is a critical component in text classification tasks due to its efficient data dimensionality reduction [10]. Several feature extraction techniques are commonly employed, including word embedding, term frequency-inverse document frequency (TF-IDF), bag-of-words (BoW), and natural language processing (NLP)-based features such as word and noun count [22]. This study adopts popular feature extraction methods: TF-IDF, Word2Vec, and BERT. TF-IDF is term frequency (TF) and inverse document frequency (IDF), considering word frequency across all texts. TF-IDF is a widely used technique in text classification, particularly in text detection. In TF-IDF, the term weight W for word t in document d is calculated using (2).

$$W(d,t) = TF(d,t) * log(N/df(t)) \tag{2}$$

In (2), N representing the total corpus size and $df(t)$ being the frequency of word t in the corpus [8]. The initial phase of text classification involves converting data into vectors, a task facilitated by the Word2Vec method. Word2Vec allows for the creation of vector representations for text by training on extensive text data [10]. Once the data is balanced using ROS, the study proceeds with feature extraction. For machine learning methods, the chosen feature extraction technique is TF-IDF, while for deep learning, Word2Vec is employed for Bi-LSTM, and BERT-based feature extraction is utilized for the BERT model. After the feature extraction process, the training and validation data are input into the model training phase, leveraging the selected classification method.

2.6. Text classification

Text classification is the process of determining the class of a text based on a training dataset whose class is known [23]. DL and ML have achieved high accuracy in text classification, but achieving high performance sometimes depends on training data size and quality [24]. Several studies have focused on the detection of cyberbullying using machine learning techniques. Hani *et al.* [9] proposed a method that utilizes TF-IDF feature extraction and sentiment analysis algorithms. The authors employed SVM and neural networks (NN) for classification. In a comparative analysis conducted by [8], various machine learning techniques were evaluated for cyberbullying detection on Twitter. They employed TF-IDF and Word2Vec for feature extraction and applied a range of classification algorithms including multinomial naive Bayes (NB), stochastic gradient descent (SGD), logistic regression (LR), logistic light gradient boosting machine (LGBM), random forest (RF), adaptive boosting (AdaBoost), and support vector machine (SVM). Wang *et al.* [7] introduced SOSNet, a graph convolutional network approach for fine-grained cyberbullying detection. Their feature extraction methods included sentence-BERT (SBERT), BERT, DistilBERT, global vectors (GloVe), Word2Vec, FastText, TF-IDF, and BoW. The authors employed various classification algorithms such as LR, SVM, multi-layer perceptron (MLP), XGBoost (XGB), k-nearest neighbors (KNN), and NB.

Yan and Zheng [10] proposed a text classification model that utilized Word2Vec for multi-level topic feature extraction. They applied recurrent neural networks (RNN), Attention-based bidirectional long short-term memory (Att-Bi-LSTM), region-based convolutional neural networks (RCNN), and regional container lines (RCL) for classification. Paul and Saha [12] developed CyberBERT, a BERT-based model for cyberbullying identification. They employed BERT for feature extraction and applied SVM, LR, convolutional neural networks (CNN), RNN+LSTM, Bi-LSTM, and BERT large for classification. Behzad *et al.* [13] proposed a rapid cyberbullying detection method using compact BERT models. They experimented with different sizes of compact BERT, including BERT-base, medium, small, mini, and tiny. Eronen *et al.* [25] aimed to improve classifier training efficiency for automatic cyberbullying detection using feature density with TF-IDF feature extraction. They employed LR, coordinate gradient descent linear regression (CGD LR), SGD SVM, linear SVM, KNN, NB, RF, AdaBoost, XGBoost, MLP, and CNN as classification algorithms. In contrast to the previous studies, this study proposes a multi-level classification approach to achieve more specific categorization. By considering subcategories within larger classes, this approach enhances the accuracy of cyberbullying detection systems.

This study selected the most used ML methods: multinomial NB, Bernoulli NB, LR, SVM, and RF. For the DL method, this study chose Bi-LSTM and BERT. Compared to other methods, BERT is the newest method. The BERT classification method is a technique created by the Google AI research lab that has been shown to outperform several NLP tasks. The use of the BERT method in recent years has been widely proposed. This method has been used successfully in developing several models for various NLP tasks. BERT is designed to train deep bidirectional representation of text sequences by concurrently conditioning left and right contexts across all layers [12]. Therefore, this study will compare BERT with several old classification methods to prove its superiority and to make the best accuracy in this study. After performing feature extraction, training and validation data are entered into the training model process using the classification method. The model generated from the training stage is used to classify test data in the testing stage. This study compares the accuracy results from the classification of test data using various classification methods to find the best classification method and algorithm that can increase the accuracy value in detecting cyberbullying.

2.6.1. Bidirectional long short-term memory (Bi-LSTM)

Bi-LSTM is a deep learning model consisting of two LSTM units, namely the forward LSTM and the backward LSTM [26]. Long short-term memory (LSTM) replaced RNN because it can overcome the vanishing gradients problem experienced by RNN, using memory cell state and gates to control the flow of information within the network. LSTM enables neural networks to learn long-term patterns in sequential data [27]. Bi-LSTM outperforms the BERT model on small datasets and can be trained faster compared to fine-tuning pre-trained BERT models [28]. This study utilizes a Bi-LSTM model for text classification using PyTorch. The process commences with data preprocessing, which involves tokenization and vocabulary creation. Word2Vec embeddings are employed to represent the words in the text data. The dataset is divided into training, validation, and test sets, with oversampling applied to address class imbalance. The Bi-LSTM architecture comprises a single LSTM layer with 100 hidden units, enabling bi-directionality, and employs a dropout rate of 0.5 for regularization. The training procedure makes use of the AdamW optimizer, employing a learning rate (LR) of 3e-4 and a weight decay of 5e-6. Early stopping is put into effect after 5 epochs, and graphics processing unit (GPU) acceleration is employed to enhance computational efficiency.

2.6.2. Bidirectional encoder representations from transformers (BERT)

BERT is a deep learning model developed by Google in 2018. BERT has the capability to be applied in tasks related to natural language processing, including activities like text classification and text generation [13]. BERT leverages the Transformer to understand the context and relationships between words in sentences. BERT uses a SoftMax function as a classification layer to calculate class probabilities. Through fine-tuning, BERT can be adapted to specific tasks by optimizing its parameters [12]. In this study, BERT is employed for text classification. The selected BERT model is BERT-base-uncased [29], featuring 12 layers and 768-dimensional embeddings. Model training spans 2 epochs using the AdamW optimizer with a learning rate 5e-5 and epsilon set to 1e-8. A dynamic learning rate scheduling strategy is implemented during training. Textual data sourced from a comma-separated values (CSV) file undergoes preprocessing, including tokenization, and is divided into training, validation, and testing sets. Model performance is evaluated using confusion matrices and classification reports, demonstrating the effectiveness of BERT in text classification under optimized training settings.

2.7. Evaluation

In this study, the performance of the classifier method is assessed using recall, precision, and the F1-score. The precision is the exactness between the basic truth information and the answers by the system [30]. The recall is the success rating of the system in rediscovering information [31]. The F1-score is the sync mean of recall and precision [32].

3. RESULTS AND DISCUSSION

This portion discusses the performance outcomes of the study. Firstly, this study compares the classification results in a single-level classification. Then compare the results of the best model in the single level with the results of the multi-level.

3.1. Model performance

Based on the experimental results, this study identified that the BERT classification model achieved the highest accuracy, reaching 94%, during the single-level classification phase. The effectiveness of the BERT method in cyberbullying detection can be attributed to its unique transformer feature known as bidirectionality. This feature enables BERT to grasp the meaning of ambiguous language within the text by considering the context provided by the surrounding text. In contrast to other methods, such as Multinomial NB, Bernoulli NB, LR, RF, SVM, and Bi-LSTM, which can only process text input sequentially (either left-to-right or right-to-left), BERT stands out due to its ability to process text bi-directionally, thus comprehending both directions simultaneously. Consequently, the classification process continued to utilize the BERT method in multi-level classification.

In the context of multi-level classification using BERT, an accuracy of 94% was achieved at level one, and an impressive 99% at level two. The detailed results of single-level classification can be found in Table 3, while those of multi-level classification are presented in Table 4. Notably, the yellow highlights in both Tables 3 and 4 indicate the classification results obtained using the BERT method. It is important to highlight that when comparing single-level and multi-level classifications, the BERT classification method exhibited slightly lower accuracy in the former. These findings strongly support the notion that the integration of the BERT method with the MCML algorithm has a substantial and positive impact on accuracy.

Table 3. Result of classification multi-class with single level

Multi-class single-level classification			
	No.	Experiment	Accuracy
Multi-class single-level classification	1.	TF-IDF+Multinomial NB	0.84
	2.	TF-IDF+Bernoulli NB	0.91
	3.	TF-IDF+LR	0.92
	4.	TF-IDF+RF	0.94
	5.	TF-IDF+SVM	0.93
Deep learning	1.	Word2Vec+Bi-LSTM	0.93
	2.	BERT+BERT	0.94

Table 4. Result of classification multi-class with multi-level

Multi-class single-level classification			
	No.	Experiment	Accuracy
Deep learning	1.	BERT+BERT level 1	0.94
	2.	BERT+BERT level 2	0.99

3.2. Comparison results

This section compares the results of this study with previous studies that used the same dataset. Wang *et al.* [7] are the author of this dataset, and they use it to detect cyberbullying on Twitter using a new classification method, SOSNet. This method proved to be better than other classification methods in their research. The previous study obtained the best accuracy results by combining the SBERT word embedding with the SOSNet classification method. However, in the classification process, previous research carried out a single-level classification, where age, ethnicity, religion, and gender classes are at the same level as the not_cyberbullying class. Even though these classes should be at different levels because age, ethnicity, religion, and gender are types of cyberbullying. Therefore, this study carries out a multi-level classification where level one is carried out to classify cyberbullying and not_cyberbullying and level two is to classify cyberbullying classes into age, ethnicity, religion, and gender classes. This study compares the results of classification using a single-level with the results obtained using a multi-level classification algorithm. Table 5 shows the combination of MCML and BERT produces a much higher accuracy value compared to accuracy from previous studies using single-level classification, SBERT, and SOSNet.

Table 5. Comparison results

Authors	Classification	Experiment	Result	
			F1-score	Accuracy
Wang *et al.* [7]	Single-level	SBERT+SOSNet	0.92	0.92
The proposed method	Multi-level	BERT on MCML (Level 2)	0.99	0.99

4. CONCLUSION

This study proposes to use a multi-level classification algorithm to detect cyberbullying on Twitter. This study compared the multi-class classification using the multi-level and single-level. This study uses several DL and ML methods as classifiers. From this comparison, the best classification method is BERT. This study obtained the best results using BERT at level two in the multi-level classification, with an accuracy value of 99%. The application of MCML algorithms has been proven to increase the accuracy of the classification process. For further work, this study recommends using BERT methods of various shapes, large, small, mini, and tiny, to determine whether the size of the BERT affects classification performance using the MCML algorithm.

ACKNOWLEDGEMENTS

This research was funded by Institut Teknologi Sepuluh Nopember (ITS) under Research Program Department Funds; the Indonesian Ministry of Education and Culture under *Penelitian Terapan Unggulan*

Perguruan Tinggi (PTUPT) Program; and Institut Teknologi Sepuluh Nopember (ITS) under project scheme of the Publication Writing and IPR Incentive Program (PPHKI).

REFERENCES

[1] M. F. López-Vizcaíno, F. J. Nóvoa, V. Carneiro, and F. Cacheda, "Early detection of cyberbullying on social media networks," *Future Generation Computer Systems*, vol. 118, pp. 219–229, May 2021, doi: 10.1016/j.future.2021.01.006.

[2] J. Brailovskaia, T. Teismann, and J. Margraf, "Cyberbullying, positive mental health and suicide ideation/behavior," *Psychiatry Research*, vol. 267, pp. 240–242, Sep. 2018, doi: 10.1016/j.psychres.2018.05.074.

[3] M. C. Martínez-Monteagudo, B. Delgado, Á. Díaz-Herrero, and J. M. García-Fernández, "Relationship between suicidal thinking, anxiety, depression and stress in university students who are victims of cyberbullying," *Psychiatry Research*, vol. 286, Apr. 2020, doi: 10.1016/j.psychres.2020.112856.

[4] T. T. Q. Ho, C. Li, and C. Gu, "Cyberbullying victimization and depressive symptoms in Vietnamese university students: Examining social support as a mediator," *International Journal of Law, Crime and Justice*, vol. 63, Dec. 2020, doi: 10.1016/j.ijlcj.2020.100422.

[5] H.-F. Hu *et al.*, "Cyberbullying victimization and perpetration in adolescents with high-functioning autism spectrum disorder: correlations with depression, anxiety, and suicidality," *Journal of Autism and Developmental Disorders*, vol. 49, no. 10, pp. 4170–4180, Oct. 2019, doi: 10.1007/s10803-019-04060-7.

[6] T. K. H. Chan, C. M. K. Cheung, and R. Y. M. Wong, "Cyberbullying on social networking sites: the crime opportunity and affordance perspectives," *Journal of Management Information Systems*, vol. 36, no. 2, pp. 574–609, Apr. 2019, doi: 10.1080/07421222.2019.1599500.

[7] J. Wang, K. Fu, and C.-T. Lu, "SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection," in *2020 IEEE International Conference on Big Data (Big Data)*, Dec. 2020, pp. 1699–1708, doi: 10.1109/BigData50022.2020.9378065.

[8] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on Twitter," *Future Internet*, vol. 12, no. 11, Oct. 2020, doi: 10.3390/fi12110187.

[9] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, 2019, doi: 10.14569/IJACSA.2019.0100587.

[10] Y. Yan and K. Zheng, "Text classification model based on multi-level topic feature extraction," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, Dec. 2020, pp. 1661–1665, doi: 10.1109/ICCC51575.2020.9344894.

[11] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained BERT model," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Jul. 2020, pp. 1096–1100, doi: 10.1109/ICESC48915.2020.9155700.

[12] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimedia Systems*, vol. 28, no. 6, pp. 1897–1904, Dec. 2022, doi: 10.1007/s00530-020-00710-4.

[13] M. Behzadi, I. G. Harris, and A. Derakhshan, "Rapid Cyber-bullying detection method using compact BERT models," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, Jan. 2021, pp. 199–202, doi: 10.1109/ICSC50631.2021.00042.

[14] N. Hameed, A. M. Shabut, M. K. Ghosh, and M. A. Hossain, "Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques," *Expert Systems with Applications*, vol. 141, Art. no. 112961, Mar. 2020, doi: 10.1016/j.eswa.2019.112961.

[15] K. Arun and A. Srinagesh, "Multilingual twitter sentiment analysis using machine learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 5992–6000, Dec. 2020, doi: 10.11591/ijece.v10i6.pp5992-6000.

[16] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: a survey," *Information*, vol. 10, no. 4, Apr. 2019, doi: 10.3390/info10040150.

[17] L. K. Ramasamy, S. Kadry, Y. Nam, and M. N. Meqdad, "Performance analysis of sentiments in Twitter dataset using SVM models," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 3, pp. 2275–2284, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2275-2284.

[18] M. Fikri and R. Sarno, "A comparative study of sentiment analysis using SVM and SentiWordNet," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 13, no. 3, pp. 902–909, Mar. 2019, doi: 10.11591/ijeecs.v13.i3.pp902-909.

[19] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," *Applied Soft Computing*, vol. 83, Oct. 2019, doi: 10.1016/j.asoc.2019.105662.

[20] A. Viloria, O. B. P. Lezama, and N. Mercado-Caruzo, "Unbalanced data processing using oversampling: Machine learning," *Procedia Computer Science*, vol. 175, pp. 108–113, 2020, doi: 10.1016/j.procs.2020.07.018.

[21] W. Feng *et al.*, "Dynamic synthetic minority over-sampling technique-based rotation forest for the classification of imbalanced hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2159–2169, Jul. 2019, doi: 10.1109/JSTARS.2019.2922297.

[22] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[23] D. Sameh, G. Khoriba, and M. Haggag, "Behaviour analysis voting model using social media data," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 2, pp. 211–221, Apr. 2019, doi: 10.22266/ijies2019.0430.21.

[24] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6381–6387, doi: 10.18653/v1/D19-1670.

[25] J. Eronen, M. Ptaszynski, F. Masui, A. Smywiński-Pohl, G. Leliwa, and M. Wroczynski, "Improving classifier training efficiency for automatic cyberbullying detection with feature density," *Information Processing & Management*, vol. 58, no. 5, Sep. 2021, doi: 10.1016/j.ipm.2021.102616.

[26] J. Deng, L. Cheng, and Z. Wang, "Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification," *Computer Speech & Language*, vol. 68, Jul. 2021, doi: 10.1016/j.csl.2020.101182.

[27] X. Song *et al.*, "Time-series well performance prediction based on long short-term memory (LSTM) neural network model," *Journal of Petroleum Science and Engineering*, vol. 186, Mar. 2020, doi: 10.1016/j.petrol.2019.106682.

[28] A. Ezen-Can, "A comparison of LSTM and BERT for small corpus," *arXiv preprint arXiv:2009.05451*, pp. 1–12, 2020.

[29] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] A. T. Haryono, R. Sarno, and R. Abdullah, "Aspect-based sentiment analysis of financial headlines and microblogs using semantic similarity and bidirectional long short-term memory," *Journal of Intelligent Engineering and Systems*, vol. 15, no. 3, pp. 233–241, 2022.

[31] R. Priyantina and R. Sarno, "Sentiment analysis of hotel reviews using latent Dirichlet allocation, semantic similarity and LSTM," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 4, pp. 142–155, Aug. 2019, doi: 10.22266/ijies2019.0831.14.

[32] I. Ghozali, K. R. Sungkono, R. Sarno, and R. Abdullah, "Synonym based feature expansion for Indonesian hate speech detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 1105–1112, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1105-1112.

BIOGRAPHIES OF AUTHORS

Salsabila was born in Bondowoso in 1999. She received a bachelor's degree in computer science from Universitas Islam Indonesia in 2021. She is currently pursuing a master's degree in informatics engineering from the Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS). Her research interests include Text mining and machine learning. She can be contacted at email: salsabilasabil151199@gmail.com.

Riyanarto Sarno is a professor, Informatics Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. He received the bachelor's degree in electrical engineering from Institut Teknologi Bandung, Bandung, Indonesia in 1983. He received M.Sc. and Ph.D. in computer science from the University of Brunswick Canada in 1988 and 1992, respectively. In 2003 he was promoted to a full professor. His teaching and research interests include internet of things, process aware information systems, intelligent systems and smart grids. He can be contacted at email: riyanarto@if.its.ac.id.

Imam Ghozali was born in Surabaya in 1996. He received a bachelor's degree in computer science from the Department of Informatics, Universitas Brawijaya in 2017. He is currently pursuing a master's degree in informatics engineering from the Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember (ITS). His research interests include text mining and computer vision. He can be contacted at email: imam.ghz17@gmail.com.

Kelly Rossa Sungkono was born in Surabaya, in June 1994. She received the master's degree in computer engineering, in 2016. She is currently a Lecturer at the Institut Teknologi Sepuluh Nopember (ITS). Her research interests include databases, information systems, and machine learning. She can be contacted at email: kelsungkono@gmail.com.