Journal of Artificial Intelligence and Data Science
Vol. 1, No. 2, 2025, pp. 1-12
DOI:

Integrating Whisper Transcription with Transformer and CNN Models for Indonesian Abusive Speech Detection

Ainandita Riwipapusa, Wiyli Yustanti, Cendra Devayana Putra
Program Study of Information System, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia
e-mail address: wiyliyustanti@unesa.ac.id (corresponding author)

Bartolomeus Priya Perkasa Utama Widada
College of Electrical Engineering and Computer Science, National Cheng Kung University, Taiwan

Received: 1 September 2025 | Revised: 20 October 2025 | Accepted: 10 November 2025

ABSTRACT

The growing popularity of audio social media in Indonesia has fueled the spread of profanity and hate speech, even though there are no regulations specifically covering its use. In this work, we propose an audio-based coarse word detection model by combining one Whisper model for transcription and another IndoBERT in a sentence context classifier to decode it along with a Convolutional Neural Network (CNN) for fine-grained coarse word prediction. The development of IDSS was based on a Knowledge Discovery in Database (KDD) process, consisting of the data selection, preprocessing, transformation, data mining and evaluation phases. The recovered dataset will be pre-processed through cleaning, tokenizing and stemming accordingly prior to labeling. To evaluate the data sharing strategies we used three proportions of shared information (80:20, 70:30 and 60:40) with 5-fold cross-validation. It was found that scenario 1 having data sharing ratio of 80:20 gave highest performance with 100% training accuracy, 98.71% validation accuracy and 99.57% test accuracy. The fusion of IndoBERT and CNN seems to be effective in finding broad sentences and certain words including rudeness. The proposed system would be a first line of defense in reducing the dissemination of hateful language content on auditory social media.

Keywords-Abusive Word Detections; Social Media; IndoBERT; Convolutional Neural Network (CNN); Text Classification

I. INTRODUCTION

Information and communication technology has shown its impact on the way of human interaction, including social media. The use of social media has become more common even in Indonesia. The number of Twitter (X) users increased to 57.5% in January, 2024 [1]. The internet user percentage based on the report of Indonesian Internet Service Providers Association (APJII) in [2], is 79.5 percent of its total population and grows about 78 percent from the previous year use. The complexity of technology on social media when not accompanied by literacy culture and emotional intelligence may lead to ignorance of the users in interacting, one of which is the emergence of a lot of hate speech that can hurt other parties [3]. Hate speech can be directed at individuals or groups according to their faith, ethnicity, gender and sexual orientation [4]. It is reported that in Indonesian, profanity is frequently used to refer to types of animals like “anjing” and “monyet” [5]. Another source, on the other hand indicates there are many more kinds of insulting words, like beastly words from dialects [6]. Online cursing not only serves as an outlet of frustration, but also as a diminutive insult to others, frequently being unrestrained and contagious, which may result in poisonous language environment among people's daily life.

This country has laws related to digital information such as the Information & Electronic Transactions Law (ITE Law) and Broadcasting Law, however, there are no provisions concerning prohibition of the use of insults on social media in audio format. The Indonesian Broadcasting Commission in addition has no power to pre-determine social media contents. This highlights the necessity of a technical component in reducing the dissemination of harmful language, which is particularly noteworthy given that audio-based content is being more widely consumed. This is where tech for identifying abusive words

on audio-based social media comes in. Today, data is not just numerical any more: text, audio, images and videos are also treated as sources of data that can be handled and analyzed [7]. For audio-based abusive word detection, transcribing speech using Speech-To-Text (STT) as a part of Automatic Speech Recognition(ASR) is the first step. We utilize the Whisper model in this study, which can conduct ASR on and transcribe multi-languages such as Indonesian with high performance under noisy conditions or regional accents.

Indo-BERT performed better on detecting verbal violence sentiment in Twitter than BERT as previous research stated, accuracy reached 72% comparing to BERT's 69% [8]. On the other hand, in a different study, the categorization was performed via Convolutional Neural Network (CNN). Another paperS compared by two classification algorithms, Convolutional Neural Network (CNN), and Super Vector Machine (SVM), using application usage sentiment analysis. The best accuracy is the CNN with 86% while the SVM only got below 70% [9]. While Study [10] depicted the highest accuracy for CNN-based model that classify hate speech in Twitter data, 90.85% (training),88.34% (validation), and 82.5% (testing).

So, in this study word classification will be done using CNN Algorithm. The absence of explicit legislation on the use of abusive words in social media has made technological solutions even more relevant. This study provides an approach to mitigate the utilisation of abusive language on voice-oriented social media. As part of the task, this paper attempts to identify abusive words from audio using IndoBERT and CNN. First, stages of processing by turning speech into words and filtering abusive terms, are considered. This model should be a good detector of abusive words.

II. METHOD

The research process that was adopted in this study is as follows. This study applies the Knowledge Discovery in Database (KDD) approach to develop a model of detecting abusive word by using CNN and sentence sentiment analysis by IndoBERT. The concept of KDD (Knowledge Discovery in Database) is a process and approach of applying knowledge discovered from the available database [11]. The process of Knowledge discovery in Database (KDD) is to extract and analyze large volumes of data from various sources with the intent to find some valuable information and knowledge via data mining [12]. Following are the stages of KDD process.

A. Data Selection

The first step of this investigation consists of picking a dataset in an attempt to gather some useful data that the model could be trained on. The formed model seeks to recognize abusive sentences both holistically and more finely, at the word level. These data produced by this selection next are used for the process of data mining, i.e., to set up Convolutional Neural Network (CNN) model object correctly detecting abusive word. The used data is primary data through reaching and extracting from the social media twitter to which it was this crawling hence of a process. The data were collected using several keywords that may have included abusive utterances or negatively worded opinions in Indonesian. The two sets of data were consolidated, for a total number of 1,811 sentences. The extracted tweets are noise, non-structured text [13]. Hence, pre-processing is inevitable to tidy up the data before it is entered in to the stage of mining the data.

B. Data Preprocessing

Data preprocessing is the stage where data are organized to be ready to be processed. The standard stages of preprocessing include case folding, tokenizing, filtering and stemming [14]. At this point, the activities of cleaning up my dataset of any character and punctuation that I didn't need, are implemented. As such, the data is 'cleaned' and prepared to move forward in the analysis or modeling. Data Cleaning is the first stage performed. This is done to remove characters in the data which are not needed in classification. Cleaning is performed by stripping punctuation, emoticons, foreign characters, urls etc. That is because the text format is standardized for processing. The cleaning operation is summarized in Table 1.

TABLE 1. CLEANING

Input Process	Output Process
@geloraco Pantas korupsi ga bisa diberantas hakimnya sdh pada masuk anjing semua.....	Pantas korupsi ga bisa diberantas hakimnya sdh pada masuk anjing semua

Next, case folding stage, it is applied to the sentences or words of the dataset. Words/sentences in which a previous lower case was converted to upper will be converted back [15]. This procedure is aimed at eliminating unnecessary word forms. The case-folding process is visualized in Table 2.

TABLE 2. CASE FOLDING

Input Process	Output Process
Pantas korupsi ga bisa diberantas hakimnya sdh pada masuk anjing semua	pantas korupsi ga bisa diberantas hakimnya sdh pada masuk anjing semua

Then comes tokenization. This is performed by taking sentences, and splitting them into terms or tokens based on detecting space [15]. The goal of this process is to enable statistical analysis over the text (for example, count word occurrences) and also to make the data ready for using it in subsequent processing phases. Table 3 shows process of tokenization.

TABLE 3. TOKENIZING

Input Process	Output Process
pantas korupsi ga bisa diberantas hakimnya sdh pada masuk anjing semua	'pantas', 'korupsi', 'ga', 'bisa', 'diberantas', 'hakimnya', 'sdh', 'pada', 'masuk', 'anjing', 'semua'

The last phase is about stemming. This process is performed to convert the words into their base form for every token [15]. This process aims at normalizing different word forms so that they could be identified as the same object by the model and considered as not-distinct. This stemming can be viewed in Table 4.

TABLE 4. STEMMING

Input Process	Output Process
pantas korupsi ga bisa diberantas hakimnya sdh pada masuk anjing semua	pantas korupsi ga bisa diberantas hakimnya pada masuk anjing semua

C. Data Transformation

Transform In data transform stage, the text goes through a series of transformations into numerical form that can be used by machine learning techniques. The previous preprocessing for the extracted text had undergone into a manual labeling process, that is the step in which sentences are classified as abusive or non-abusive being this task manually verified. After tagging, the TF-IDF (Term Frequency and Inverse Document Frequency) method is invoked [10]. This method computes word's inputs ore based on their occurrence among the corpus, so t hat the distinguishing words have a larger we ight. The output is a vector of numbers features matrix that will be input to SVM classification. Das et al. [16] state that TF-IDF is a statistical model, therefore it can be used to estimate few important words on data. TF-IDF has the biggest number of its applications in text or word classification as well.

D. Data Mining

This study uses two approaches that are IndoBERT for classifying sentence context and CNN for word-abusive classification, hence identification occurs in a multi-stages classification. This is the data reduction step in KDD. By combining IndoBERT and CNN, the system is able to deeply understand a context and can also identify abusive words specifically so that it enhances the accuracy and effectiveness of detecting Indonesian-language audio content. Model like Bidirectional Encoder Representations from Transformers (BERT) is capable of contextual emotion recognition in text and performs well on a wide range of Natural Language Processing (NLP) tasks. BERT was pre-trained on English-language data, but IndoBERT has been introduced and is trained with the Indonesian language instead, and it has shown high performance in Indonesian text classification [17]. IndoBERT is superior in identifying domain-specific entities in the Indonesian context also [18]. IndoBERT is a pretrained BERT-based language model, pre-trained using large corpus of Indonesian text which has ability to understand the relationship between words, their meaning and its context as well as syntactic structure of sentences. The processed-text with the preprocessed and transformed contextual embeddings from IndoBERT as input is the n used to classify at a sentence level (either "abusive" or "non-abusive"), and this gives an indication of whether exploration should continue for specific words.

In addition, CNN is employed to classify the words in the sentences found as abusive by IndoBERT. CNN is the sole arbiter of what qualifies as an abusive word or phrase. The CNN model architecture is shown in Table 5.

TABLE 5. CNN ARCHITECTURE PARAMETERS

Layer	Parameter	Output Shape
Input Layer	shape = (max_length,)	(None, max_length)
Embedding	input_dim = 5000, output_dim = 128	(None, max_length, 128)
Conv1D	filters = 128, kernel_size = 3, activation = 'relu'	(None, new_len, 128) (~ max_length - 2)
GlobalMax Pooling1D	-	(None, 128)
Dense	units = 64, activation = 'relu'	(None, 64)
Dropout	rate = 0.5	(None, 64)
Output (Dense)	units = 1, activation = 'sigmoid'	(None, 1)

E. Evaluation

The model performances are evaluated to examine the performance of the system to detect and classify abusive words or sentences from audio. Complete evaluation is done using multiple standard performance metrics: accuracy, precision, recall and the F1-score which give a holistic view of the model's performance on positive (which are abusive) as well as negative class (non-abusive). Also, a confusion matrix is employed to perform an analysis based on the number of correct and incorrect predictions per class, in order to identify the nature of errors (false positive and false negative) that are committed.

In the proposed method, evaluation is done in a phased manner as per system workflow. Performance of IndoBERT in sentence context classification is first analysed to guarantee that the abusive sentences are identified correctly. We then separately examine the performance of the CNN in identifying abusive lexicons. The union testing of the two models is also taken place to compare the performance with the multi-stage classification method. This tiered evaluations of the system produce results on the individual and combined performance of each model allowing us to clearly identify which components need to be further developed in order to generate more accurate Automation Abusive Words Detection/Filtering system.

III. PROPOSED METHOD

The algorithm presented is to pursue a procedural audio abusive word discovery system. The system overall is a combination of multiple models: Whisper to transcribe audio to text, IndoBERT for classification of sentence context model, and CNN for detection of abusive word. This methodology was crafted such that offending words can be tagged in the context of sentences already flagged as abusive and become part of system's word level analysis instead of classifying at the sentence level.

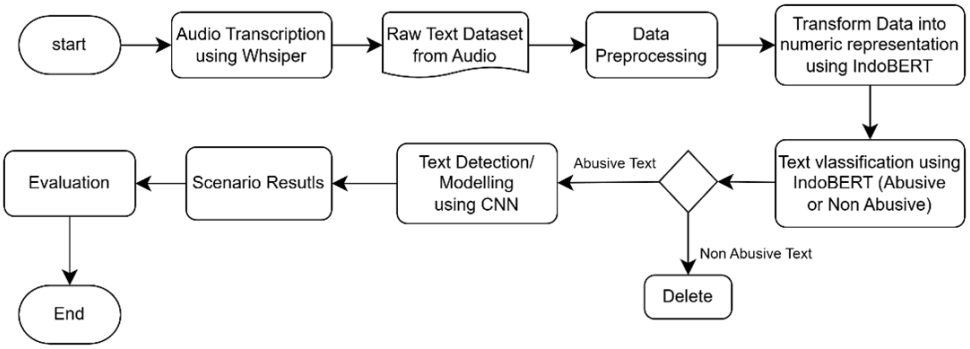


Figure. 1. Research Methods Workflow

Figure 1 shows the main workflow of this system, which will follow by transcribing the audio into text using Whisper to create a raw text representation (or auto-transcription) from an audio database. The preprocessed transcribed text is next converted to numerical features (contextual embeddings) using IndoBERT. Second, IndoBERT is applied to classify the sentences as "abusive" or "non-abusive." Only the detected abusive sentences are passed to the detection of abusive words phase using CNN that looks for specific word or phrases that categorize these as abusive.

Andini et al. [19] that CNN has two main modules including layers of feature extraction (convolution) and layers of data dimensionality reduction (pooling). Additionally, the accuracy of whether or not a sentence classifies as hate speech was 87% using CNN model. A study has been done by Elouali et al. [20], also reported that CNN based model is useful to identify and classify sentences/utterances as hate speech or non–hate speech. Therefore, this method can help the system correctly find out abusive words in subsets of relevant sentences, improving the efficiency and accuracy of the model. Thus, this study assume that our approach of an entirely convoluted network structure will likely utilize audio dataset as the first source of data and infer a sentence context to detect abusive word specifically for research purposes themselves without any audio censorship scheme.

IV. RESULT AND DISCUSSION

The above results indicate that training method (using the CNN to detect abusive words) is very effective. Data was split in three scenarios for training as follows: 80:20, 70:30 and 60:40 using a 5-fold cross-validation approach. The data distribution in the three situations is summarized in Table 6.

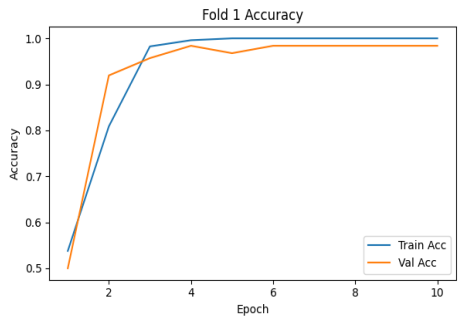
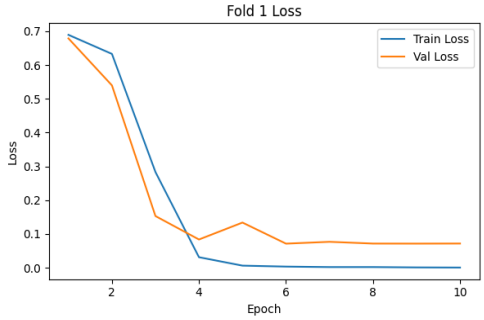
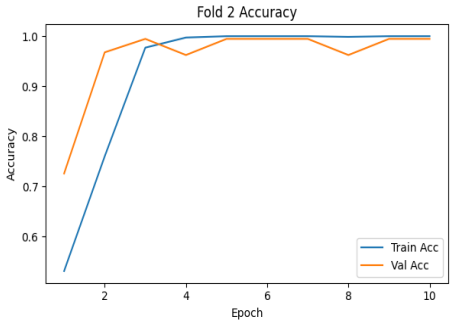
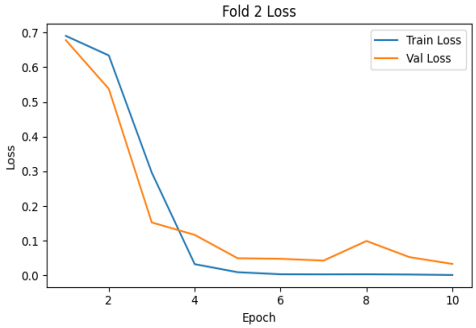
TABLE 6. DATA SPLIT SCENARIOS

Scnario	Ratio	Train Data	Test Data
Scenario 1	80 : 20	928	232
Scenario 2	70 : 30	812	348
Scenario 3	60 : 40	696	464

A. Scenario 1 (80:20)

Modeling was performed in Scenario 1 with a split of the data in an ratio of 80:20. The 80/20 ratio represents the training/validation data which consists of 928 data points, and The test data with a total number of data points as 232. This step generated the accuracy and loss for each fold. See the precision and loss plots comparing this experiment in Table 7.

TABLE 7. ACCURACY & LOSS GRAPH OF SCENARIO 1

Fold	Accuracy Graph of Scenario 1	Loss Graph Scenario 1
1		
2		

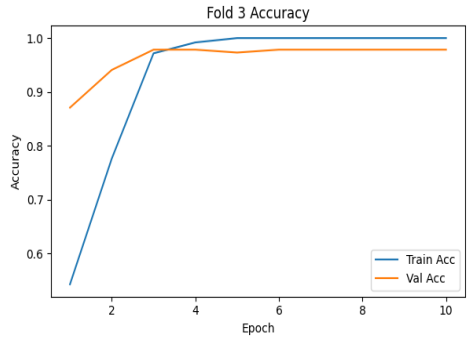
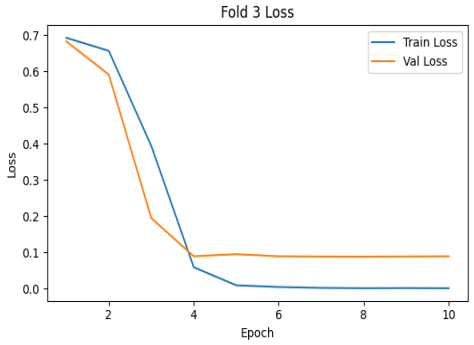
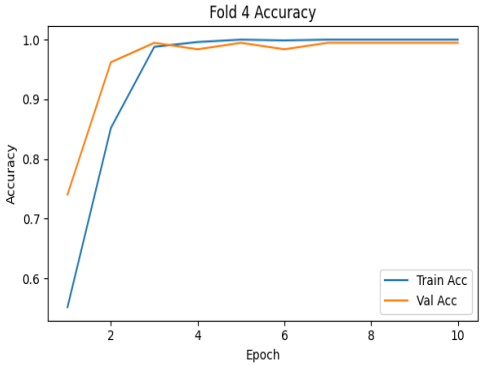
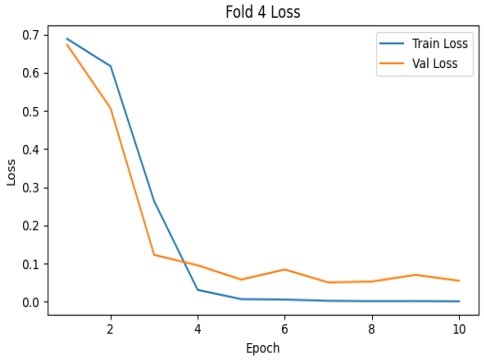
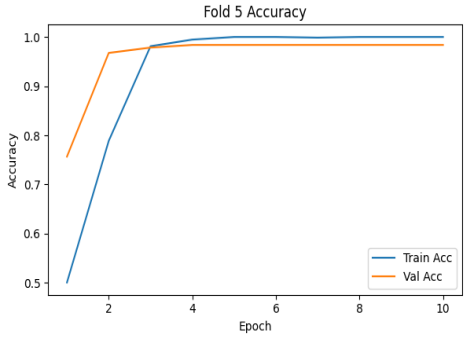
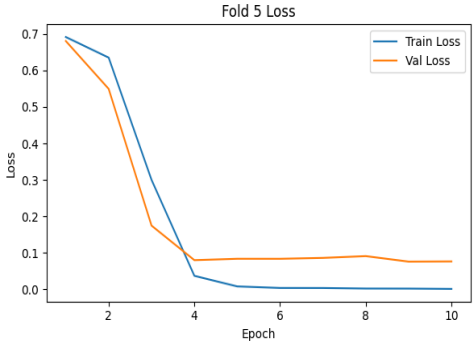
Fold	Accuracy Graph of Scenario 1	Loss Graph Scenario 1
3		
4		
5		

Table 8 is the training results of CNN model using accuracy and loss for training and validation data in each fold. The model performs well, with training accuracy of 100% in all folds and validation accuracy ranging from 0.9785 to 0.9946.

TABLE 8. MODEL TRAINING RESULTS OF SCENARIO 1

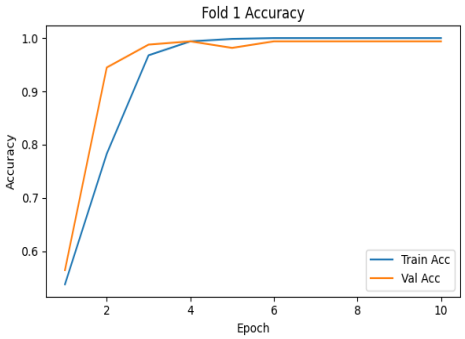
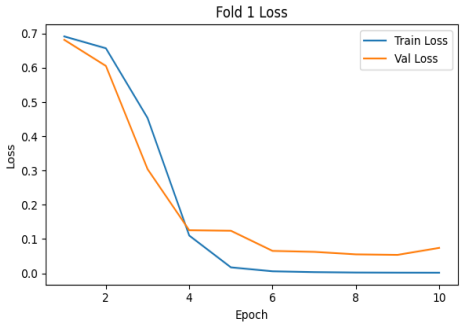
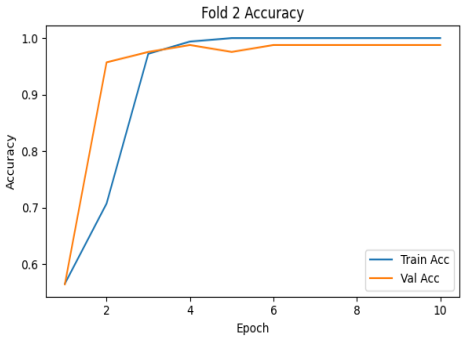
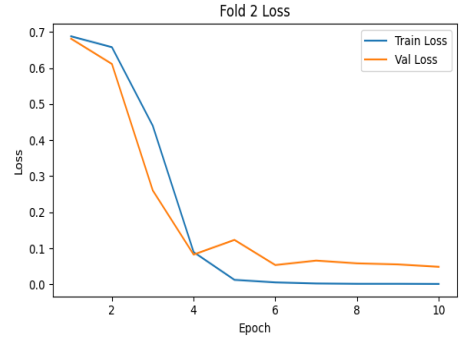
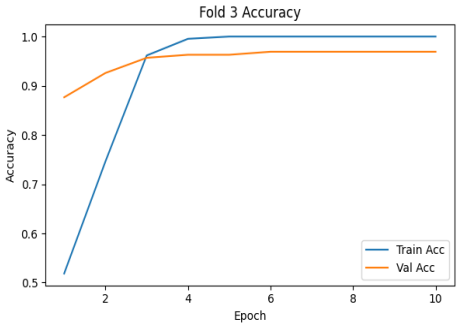
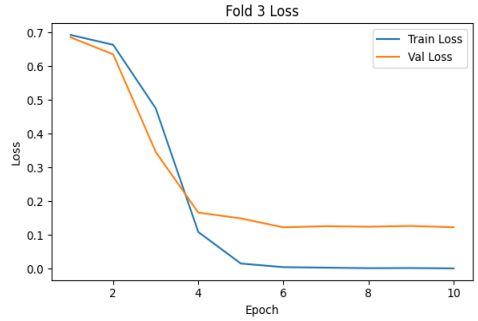
Fold	Accuracy		Loss	
	Train	Validation	Train	Validation
1	1.00	0.9839	0.0006	0.0717
2	1.00	0.9946	0.0011	0.0330
3	1.00	0.9785	0.0013	0.0894
4	1.00	0.9946	0.0012	0.0551
5	1.00	0.9838	0.0009	0.0763
Average	1.00	0.9871	0.0010	0.0651

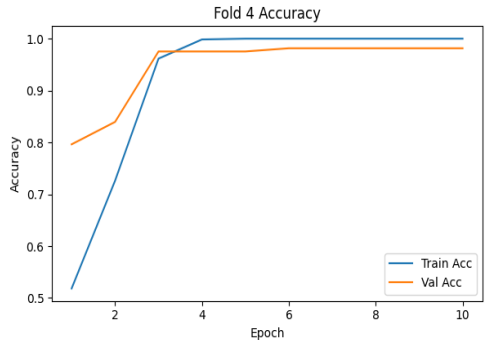
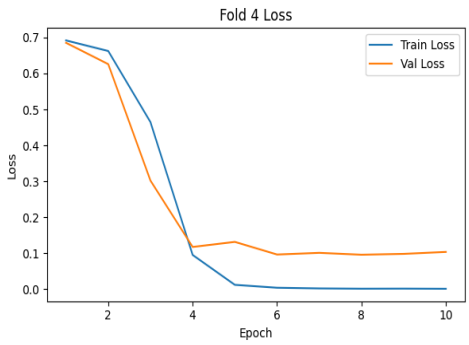
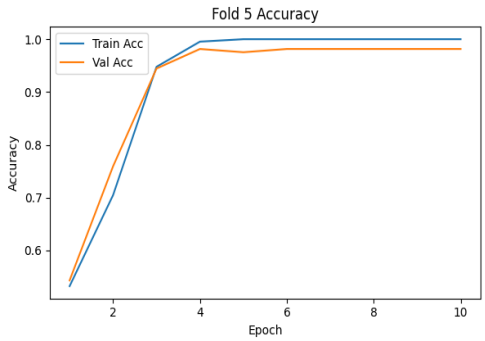
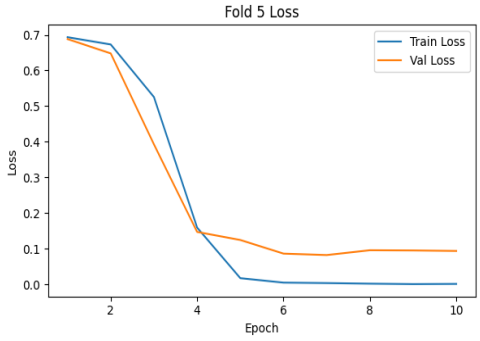
The highest values of validation accuracy (0.9946) and validation loss (0.0330) were also attained on the second fold, which exhibited an outstanding generalization ability. The model from this fold was chosen for the testing set.

B. Scenario 2 (70:30)

Modeling was conducted in Scenario 2 using a 70:30 data split. The ratio of 70 represents the training and validation data consisting of 812 data points, and the ratio of 30 represents the test data consisting of 348 data points. This stage produced accuracy and loss values for each fold. See the comparative accuracy and loss graphs from this experiment in Table 9.

TABLE 9. ACCURACY & LOSS GRAPH OF SCENARIO 2

Fold	Accuracy Graph of Scenario 2	Loss Graph Scenario 2
1		
2		
3		

Fold	Accuracy Graph of Scenario 2	Loss Graph Scenario 2
4		
5		

The learning results of the training and validation data in each fold for the accuracy and loss from the CNN model are shown in Table 10. The model has shown to be very robust and well-performing, as in the 5 folds trained, it was able to achieve a training accuracy of 100% in all cases against a validation accuracy varies from 0.9691 up to 0.9939. Fold 1 obtained the best results with the highest validation accuracy (0.9939) and a validation loss on 0.0742, which proves good generalisation capability. The best model from this fold was then chosen using the test set.

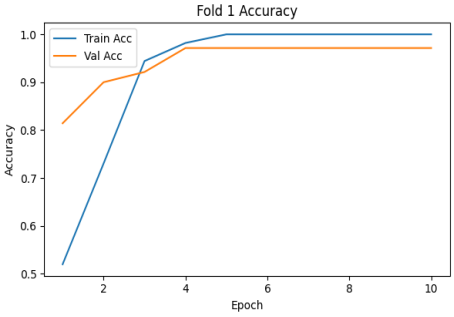
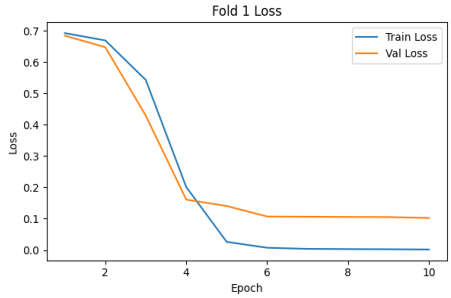
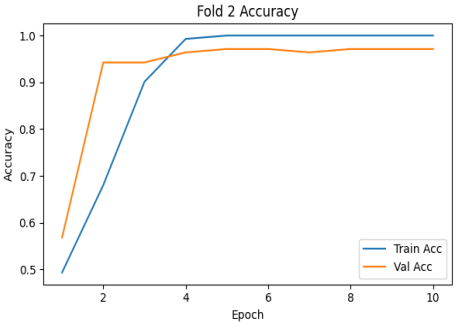
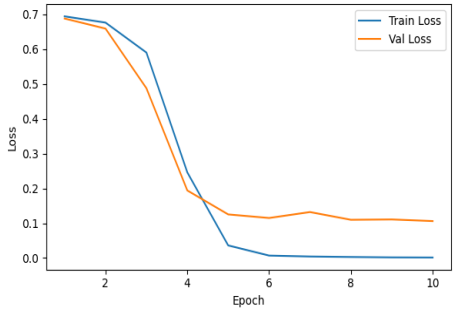
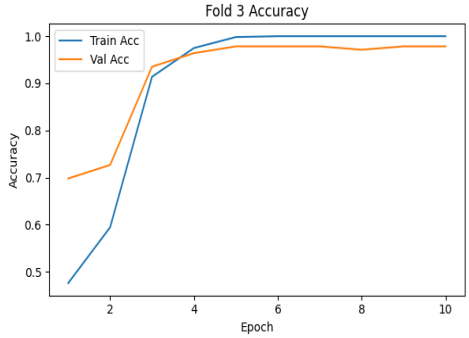
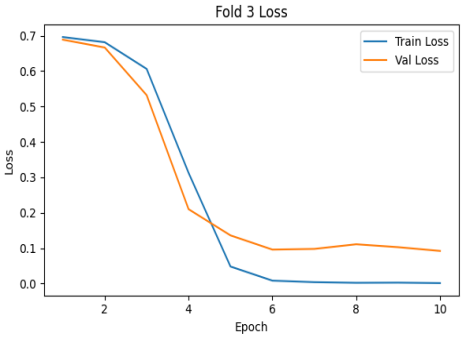
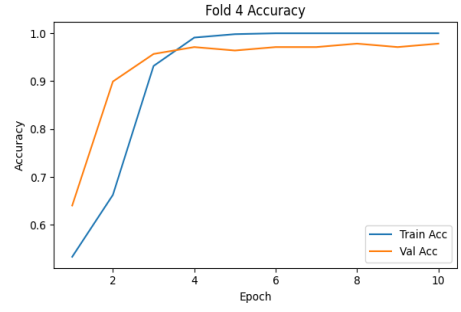
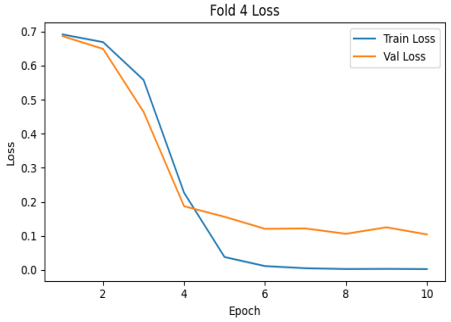
TABLE 10. MODEL TRAINING RESULTS OF SCENARIO 2

Fold	Accuracy		Loss	
	Train	Validation	Train	Validation
1	1.0000	0.9939	0.0018	0.0742
2	1.0000	0.9877	0.0011	0.0486
3	1.0000	0.9691	0.0013	0.1230
4	1.0000	0.9815	0.0010	0.1035
5	1.0000	0.9815	0.0017	0.0941
Average	1.0000	0.9827	0.0014	0.0887

C. Scenario 3 (60:40)

Scenario 3 was modelled with a ratio of 60:40 for the data split. The ratio 60 stands for training and validation data (696 patterns) and the raio 40 stands for test data (464 patterns). On this stage were given accuracy and loss on every fold. Table 11 shows the accuracy and loss graphs from this experiment comparing them to previously created ones.

TABLE 11. ACCURACY & LOSS GRAPH OF SCENARIO 3

Fold	Accuracy Graph of Scenario 3	Loss Graph Scenario 3
1		
2		
3		
4		

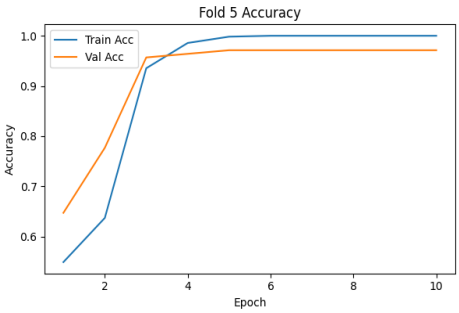
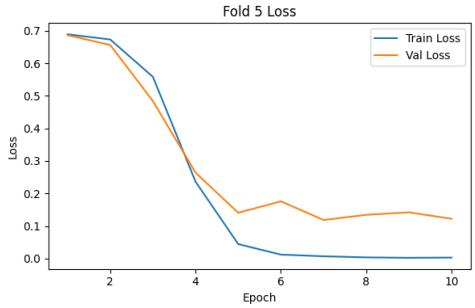
Fold	Accuracy Graph of Scenario 3	Loss Graph Scenario 3
5		

Table 12 shows the training results of CNN model in terms of accuracy and loss for the train and validation data for each fold. The model presents very high and stable performance, with training accuracy of 100% in all folds while validation accuracy ranges from 0.9712 to 0.9784. The best performance was obtained in fold 3, with the highest validation accuracy and lowest validation loss (0.9784 and 0.0923, respectively), which can be considered a good generalization ability. We chosen the model of this fold for testing by test data.

TABLE 12. MODEL TRAINING RESULTS OF SCENARIO 3

Fold	Accuracy		Loss	
	Train	Validation	Train	Validation
1	1.0000	0.9714	0.0018	0.0742
2	1.0000	0.9712	0.0011	0.0486
3	1.0000	0.9784	0.0013	0.1230
4	1.0000	0.9784	0.0010	0.1035
5	1.0000	0.9712	0.0017	0.0941
Average	1.0000	0.9741	0.0014	0.0887

D. Scenario Evaluations

Test data was tested in all three scenarios by the models trained on best-performing folds. For each of the scenarios, the test data utilized corresponded with that of each scenario (80:20 had 232 instances; 70:30 had 348; and 60:40 had 464). The test data used in each setting had different numbers of labels, described by the Table 13.

TABLE 13. TEST DATA LABEL DISTRIBUTIONS

Scenario	Abusive	Non Abusive
Scenario 1	232	232
Scenario 2	174	174
Scenario 3	116	116

The evaluation stage was performed by comparing the accuracy-loss values on the training, validation and test sets as well as using a confusion matrix & classification report of (accuracy, precision, recall, f1-score) to analyze performance for the model.

TABLE 14. ACCURACY & LOSS COMPARISON RESULTS

Scenario	Accuracy			Loss		
	Train	Validation	Test	Train	Validation	Test
Scenario 1	1.0	0.9871	0.9957	0.0010	0.0651	0.0321
Scenario 2	1.0	0.9827	0.9943	0.0014	0.0887	0.0640
Scenario 3	1.0	0.9741	0.9914	0.0020	0.1054	0.0534

We additionally report the average for the comparison between accuracies and losses on training, validation and test data on best model depth 4 in Table 14. In the three experiments, a training accuracy of 100% is obtained that indicates we could learn data feature of the data very well. Both validation and test accuracies increased with the size of training data, where the

best performances were obtained for Scenario 1: the validation accuracy was 0.9871 and test accuracy was 0.9957 and train (0.0010), validate (0.0651) and test (0.0321) losses reached their minimum values respectively.

To get a better understanding of the quality of the model's prediction, additional evaluation was performed via the classification report (which consists of precision, recall and F1-Score for each class). The classification report results for the three settings are given in Table 15.

TABLE 15. CLASSIFICATION REPORT RESULTS

Skenario	Label	Precision	Recall	F1-Score	Support
Scenario 1	Non Abusive	0.99	1.00	1.00	116
	Abusive	1.00	0.99	1.00	116
Scenario 2	Non Abusive	0.99	0.99	0.99	174
	Abusive	0.99	0.99	0.99	174
Scenario 3	Non Abusive	0.96	1.00	0.98	232
	Abusive	1.00	0.96	0.98	232

Performance of the model in each setup is shown in Table 15. It can be seen that the more training data is used such as in Scenario 1 encourages better performance with high balanced accuracy, precision recall and F1-score over both labels. Therefore, the best model is Scenario 1.

V. CONCLUSION

This model was created with the Knowledge Discovery in Database (KDD) approach with a combination of IndoBERT and Convolutional Neural Network (CNN) for abusive-word detection. First, on top of IndoBERT we applied a subsequent classification for abusive or non-abusive sentences and only the abusively marked sentences were passed through the CNN to detect specific words considered as abusive. The experiments were performed for three training and testing data splitting rates (80:20, 70:30, and 60:40) and the performance comparisons are obtained by a 5-fold cross validation. The best performance was observed in Scenario 1 (80:20 data split) with 100% training accuracy, 98.71% validation accuracy, and 99.57% testing accuracy and low training (0.10%), validation (6.51%) and testing (3.21%), thus selected for the abusive-word detection system to be designed.

REFERENCE

[1] A. P. Nugraha, A. Kurnia, P. Irawan, P. Putra, A. Rahman, and D. Dikrurahman, "The Impact of Social Media on Social Interaction and Self-Identity in Indonesian Society," Journal of Social Research, no. 9, 2024, doi: 10.55324/josr.v3i9.2254.

[2] N. Aminudin, N. Hidayat, D. Feriyanto, D. Septasari, and I. Awaliyani, "Digital Landscape and Behavior in Indonesia 2024: A National Survey Analysis of Internet Penetration, Cybersecurity Risks, and User Segmentation Using K-Means Clustering and Logistic Regression," Jurnal Teknik Informatika (Jutif), vol. 6, no. 5, pp. 3336–3351, Oct. 2025, doi: 10.52436/1.jutif.2025.6.5.5117.

[3] S. Dharmawan,) Viny, C. Mawardi,) Novario, and J. Perdana, "Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT)," Jurnal Ilmu Komputer dan Sistem Informasi, vol. 11, no. 1, 2023, doi: 10.24912/jiksi.v11i1.24066.

[4] F. Ihsan, I. Iskandar, N. S. Harahap, and S. Agustian, "Decision tree algorithm for multi-label hate speech and abusive language detection in Indonesian Twitter," Jurnal Teknologi dan Sistem Komputer, vol. 9, no. 4, pp. 199–204, Oct. 2021, doi: 10.14710/jtsiskom.2021.13907.

[5] M. F. Majid and A. Solichin, "Aplikasi Pendeteksi Kalimat Kasar Bahasa Indonesia Pada File Audio Menggunakan Jaccard Similarity Dan N-Gram," Jurnal TICOM: Technology of Information and Communication, vol. 12, no. 1, 2023, doi: 10.70309/ticom.v12i1.100.

[6] M. F. Salim and T. R. Iman, "Penggunaan Bahasa Kasar oleh Remaja Laki-laki BTN Karang Dima Indah Sumbawa dalam Pergaulannya," KAGANGA KOMUNIKA (Journal of Communication Science), vol. 4, no. 2, 2022, doi: 10.36761/kagangakomunika.v4i2.2054.

[7] N. A. Shafira and Irhamah, "Klasifikasi Sentimen Ulasan Film Indonesia dengan Konversi Speech-to-Text (STT) Menggunakan Metode Convolutional Neural Network (CNN)," Jurnal Sains dan Seni ITS, vol. 9, no. 1, 2020, doi: 10.12962/j23373520.v9i1.51825.

[8] Nurjoko and A. Rahardi, "Model Indo-BERT untuk Identifikasi Sentimen Kekerasan Verbal di Twitter," IJCCS, vol. 18, pp. 583–593, 2024, doi: 10.5281/zenodo.12788184.

[9] A. S. Simbolon, N. I. Pangaribuan, and N. M. Aruan, "Sentiment Analysis for E-learning Application using Support Vector Machine and Convolutional Neural Network," SEMINASTIKA, vol. 3, no. 1, pp. 16–25, Nov. 2021, doi: 10.47002/seminastika.v3i1.236.

[10] D. A. N. Taradhita and I. K. G. D. Putra, "Hate Speech Classification in Indonesian Language Tweets by Using Convolutional Neural Network," Journal of ICT Research and Applications, vol. 14, no. 3, pp. 225–239, 2021, doi: 10.5614/itbj.ict.res.appl.2021.14.3.2.

[11] F. Alghifari and D. Juardi, "Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," Jurnal Ilmiah Informatika (JIF), vol. 9, no. 2, 2021, doi: 10.33884/jif.v9i02.3755.

[12] I. K. J. Arta, G. Indrawan,) Gede, and G. R. Dantes, "Data Mining Rekomendasi Calon Mahasiswa Berprestasi di STMIK Denpasar Menggunakan Metode Technique for Others Reference by Similarity to Ideal Solution," Jurnal Ilmu Komputer Indonesia (JIKI), vol. 4, no. 1, 2019, doi: 10.23887/jik.v4i1.2765.

[13] F. Anistya and E. B. Setiawan, "Hate Speech Detection on Twitter in Indonesia with Feature Expansion Using GloVe," Jurnal RESTI, vol. 5, no. 6, pp. 1044–1051, Dec. 2021, doi: 10.29207/resti.v5i6.3521.

[14] H. A. Prathama and I. P. G. H. Suputra, "Evaluasi UI pada Prototype Aplikasi 'WeCare' Menggunakan Metode SUS (System Usability Scale)," JNATIA, vol. 2, no. 1, 2023, doi: 10.24843/JNATIA.2023.v02.i01.p15.

[15] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the Accuracy of Text Classification Using Stemming Method, a Case of Non-formal Indonesian Conversation," J Big Data, vol. 8, 2021, doi: 10.1186/s40537-021-00413-1.

[16] M. Das, S. Kamalanathan, and P. Alphonse, "A Comparative Study on TF-IDF Feature Weighting Method and its Analysis Using Unstructured Dataset," Computational Linguistics and Intelligent Systems, 2023, doi: 10.48550/arXiv.2308.04037.

[17] A. C. Saputra, A. S. Saragih, and D. Ronaldo, "Prediksi Emosi dalam Teks Bahasa Indonesia Menggunakan Model Indobert," Jurnal Teknologi Indormasi: Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika, vol. 19, 2025, doi: 10.47111/jti.v19i1.17617.

[18] N. Istiqomah and F. Novika, "Perbandingan Kinerja Model NER IndoBERT dan IndoLEM dalam Ekstraksi Informasi Kesehatan Pascabencana dari Berita Daring di Indonesia," JOURNAL OF COMPUTER SCIENCE AND INFORMATICS ENGINEERING, vol. 04, no. 3, pp. 158–174, 2025, doi: 10.55537/cosie.v4i3.1173.

[19] N. M. Andini, Y. Findawati, I. R. I. Astutik, and A. Eviyanti, "Implementasi Convolutional Neural Network (CNN) Untuk Mendeteksi Ujaran Kebencian Dan Emosi Di Twitter," SMATIKA JURNAL, vol. 14, no. 02, pp. 314–325, 2024, doi: 10.32664/smatika.v14i02.1346.

[20] A. Elouali, Z. Elberrichi, and N. Elouali, "Hate Speech Detection on Multilingual Twitter Using Convolutional Neural Networks," Revue d'Intelligence Artificielle, vol. 34, no. 1, pp. 81–88, 2020, doi: 10.18280/ria.340111.