

Classification of Humorous and Serious Communication Styles on X Social Media Using a Hybrid Convolutional Neural Network and Long Short-Term Memory Model

Ade Putra Erwanda ^{1*}, Munirul Ula ^{2*}, Muhammad Daud ^{3*}, Dahlan Abdullah ^{4*}, Muhammad Fikry ^{5*}

* Master's Program in Information Technology, Faculty of Engineering, Malikussaleh University, Lhokseumawe, Aceh, Indonesia.
adeputraerwanda.1998@gmail.com¹, munirulula@unimal.ac.id², mdaud@unimal.ac.id³, dahlan@unimal.ac.id⁴, muh.fikry@unimal.ac.id⁵

Article Info

Article history:

Received 2026-08-07

Revised 2026-09-11

Accepted 2026-10-05

Keyword:

Classification,
CNN-LSTM,
Communication Style,
Natural Language Processing,
X Social Media.

ABSTRACT

Social media platforms like X generate massive volumes of user-generated text containing diverse communication styles, including humorous and serious expressions that are difficult to distinguish due to their reliance on contextual and semantic nuance. This study proposes a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model to classify humorous and serious communication styles in Indonesian social media text, comparing its performance against a single CNN model. The dataset was drawn from the SmSA subset of the IndoNLU benchmark on Hugging Face, comprising 12,679 Indonesian texts. Since the corpus was originally designed for sentiment analysis, the texts were re-labeled into humorous and serious styles using a reproducible rule-based approach. After preprocessing and balancing, the final dataset contained 10,000 samples (5,000 humorous and 5,000 serious) split 80:20 for training and testing. Experimental results show that while the standalone CNN achieved 91.60% accuracy and F1-score, the hybrid CNN-LSTM model demonstrated superior performance, reaching an accuracy and F1-score of 94.10%. These findings confirm that combining CNN for local feature extraction with Bidirectional LSTM for sequential context modeling produces more discriminative text representations. Within the scope of this relabeled dataset, the proposed hybrid model provides an effective framework for automated communication style classification and serves as a strong foundation for future Indonesian social media text analysis.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

The development of information and communication technology has made social media one of the main spaces for people to interact, share information, express opinions, and express various forms of communication digitally [1]. Platform X is one of the social media that is widely used to convey ideas, comments, criticisms, and entertainment in the form of short texts. The latest data shows that the reach of X users in Indonesia will reach around 25.2 million accounts by the beginning of 2025, making the platform an important source of text data for various social media-based researches. The delivery of messages on social media does not only depend on the literal meaning of the sentence, but is also influenced by the context, choice of words, and the

intention of the implied communication so that an upload can be in the form of humor, sarcasm, irony, or serious statements [2], [3].

The characteristics of humorous texts differ from serious texts because humor is often constructed through wordplay, double meanings, hyperbole, or certain contexts that do not always contain explicit humorous words [4]. In contrast, serious texts can use language that is relaxed and resembles everyday conversation, so the two communication styles often have overlapping linguistic characteristics [5]. This condition causes the process of classification of communication styles to be more complex than classifications that only focus on the polarity of positive, negative, or neutral sentiments.

The increasing volume of text data on social media encourages the use of automatic classification methods to obtain information quickly and consistently [6]. The deep learning approach is one of the most widely used methods because it is able to learn data representation automatically without relying on manual feature engineering [7]. The Convolutional Neural Network (CNN) model is known to be effective in extracting local patterns in the form of important words and phrases, while Long Short-Term Memory (LSTM) is able to retain contextual information through word relationships in a sequence of sentences [8]. The combination of the two models allows the learning process of local features and sentence context to take place simultaneously, potentially improving the accuracy of text classification.

Various studies have proven the effectiveness of the CNN-LSTM model on the classification of social media texts. Cliche [9] successfully combined CNN and LSTM in an ensemble system for Twitter sentiment classification in SemEval-2017 and obtained the best ratings across all English sub-tasks. Geethanjali and Valarmathi [10] developed the IChOA-CNN-LSTM hybrid model for multimodal emotion recognition on social media with an accuracy of up to 97.8%. Tabassum and Nunavath [11] showed that CNN-LSTM performed better than a single model on text- and image-based multiclass cyberbullying classifications. The results of the study show that the CNN-LSTM hybrid model is able to take advantage of CNN's advantages in extracting local features as well as LSTM's ability to understand sentence context dependencies.

Research on the classification of Indonesian texts has also been carried out using various machine learning and deep learning approaches. Fasya et al. [12] applied the K-Nearest Neighbors (KNN) algorithm to the sentiment analysis of reviews of the Sirekap application on the Google Play Store. Putri et al. [13] demonstrated the effectiveness of the Local Mean K-Nearest Neighbor algorithm on the classification of cardiac disease medical record data. Erwanda and Khaidar [14] prove that CNN is able to perform well on the classification of multiclass sentiment of Indonesian-language Twitter. Khaidar [15] also used Logistic Regression to analyze the sentiment of Instagram users towards Finance Minister Purbaya Yudhi Sadewa. These studies show that text classification methods have developed rapidly in Indonesia, but most of them are still oriented towards sentiment analysis or classification based on polarity of opinion.

The research focus on the classification of communication styles is still relatively limited, especially on social media texts in Indonesian. Humorous and serious communication styles have different characteristics because they are influenced by sentence structure, relationships between words, the context of conveying the message, and the pragmatic intentions of the author, so it is not enough to distinguish only based on the appearance of certain words [16]. This condition shows that there is a research gap that

opens up opportunities for the application of the CNN-LSTM hybrid model to identify variations in communication styles on social media X.

To make the classification target explicit, this study adopts the following operational definitions. A humorous communication style is a text whose primary communicative intent is to entertain or amuse and which signals that intent through explicit surface markers such as laughter expressions, overt humor evaluations, wordplay, or exaggeration. A serious communication style is a text whose primary communicative intent is to inform, evaluate, complain, or persuade, and which is delivered without such markers. It must be stated at the outset that humor is a multifaceted phenomenon that may also be realized as satire, irony, sarcasm, meme language, or absurd expression, and that a serious message may carry humorous elements at the same time, so that the two categories are not mutually exclusive in the general case. The present study therefore addresses a deliberately narrowed operationalization of the construct: the binary separation of explicitly marked humorous texts from texts that carry no such marker. Implicit humor without a surface marker, most notably deadpan sarcasm and irony, falls outside this label definition and is treated as a boundary of the study rather than as a model error.

The combination of CNN and LSTM is by now a well-established architecture for text classification, and the contribution of this study therefore does not lie in the architecture itself but in the task, the language, and the empirical protocol to which it is applied. Existing Indonesian CNN-LSTM studies are almost exclusively concerned with sentiment polarity, topic, or hate-speech detection, where the label is a property of the content of the message. Communication style, in contrast, is a property of the manner of delivery and is largely orthogonal to polarity: a negative complaint may be phrased humorously, and a positive endorsement may be phrased in a strictly formal register. The specific contributions of this work are therefore: (i) the formulation of humorous versus serious communication style as a supervised classification task for Indonesian user-generated text, together with an explicit operational definition of both categories; (ii) a transparent and fully reproducible rule-based relabeling protocol, comprising the humor-marker lexicon, the decision rule, and worked examples including ambiguous cases; (iii) a controlled comparison between a single CNN and a CNN-LSTM hybrid under identical partitions, preprocessing, and hyperparameters, complemented by per-class metrics that identify which style is the harder one to recognize; and (iv) an explicit delimitation of the conditions under which the reported performance can be expected to hold.

II. METHOD

This study used a quantitative experimental approach to classify humorous and serious communication styles in Indonesian-language social media text X using a single CNN

model and a CNN-LSTM hybrid model. The research stage includes literature study, data collection, data labelling and balancing, *pre-processing* and tokenization, model design and training, and ending with evaluation and comparison of the performance of the two models using *accuracy*, *precision*, *recall*, and *F1-score metrics*.

A. Data Sources and Characteristics

The research data is sourced from the public SmSA (*Sentiment Analysis*) dataset on the IndoNLU benchmark collection accessed through the Hugging Face platform, with the official IndoNLP/indonlu repository on GitHub. This dataset contains 12,679 texts in Indonesian that were initially labeled sentiment (positive, negative, neutral). Because the focus of this research is the classification of communication styles, not sentiments, the original labels are not used and each text is relabeled into two categories, namely humor (a style of communication that is funny, lighthearted, and intended to entertain) and serious (an informative, formal, and objective style of communication without an element of humor).

The choice of SmSA requires explicit justification, because the corpus was originally constructed for sentiment analysis and not for communication-style classification. Three considerations motivated its selection. First, SmSA is one of the largest manually curated collections of Indonesian user-generated text that is publicly and permanently available [21]; its documents preserve the informal register, abbreviations, slang, emoticons, and Indonesian-English code-mixing that also characterize platform X. Second, the use of a public benchmark makes the entire pipeline reproducible: any reader can obtain exactly the same 12,679 documents, apply the labeling rule reported in Section II.B, and reconstruct the partitions and the results. Third, and most important for the validity of the study, SmSA is used only as a source of raw text. The original sentiment labels are discarded in full before labeling begins and play no role whatsoever in the construction of the communication-style labels.

This last point addresses a conceptual objection that must be stated explicitly. Sentiment polarity and communication style are two distinct dimensions of a message: polarity describes what attitude is expressed towards the object of the message, whereas style describes how the message is delivered. A positive sentiment is not equivalent to humor, and a negative sentiment is not equivalent to seriousness; angry complaints are frequently phrased through sarcasm and wordplay, while enthusiastic recommendations are frequently phrased in a plain, formal register. Any relabeling scheme that mapped positive sentiment onto humorous and negative sentiment onto serious would therefore yield labels that do not represent communication style at all. No such mapping is performed here: the style label of a document is determined exclusively by the presence or absence of surface humor markers in the text itself, independently of the original polarity annotation of that document.

It must equally be acknowledged that SmSA is a general benchmark of Indonesian user-generated text rather than a corpus collected purposely from platform X. It is linguistically comparable to X content, but it does not reproduce the distribution of communication styles among X users, which is shaped by the character limit, the reply and quote mechanics, and the trending-topic dynamics of that platform. SmSA is therefore treated in this study as a linguistically appropriate proxy corpus, and the conclusions are scoped accordingly. Validation on a corpus collected directly from X remains necessary before any claim of platform-level generalization can be made.

B. Rule-Based Relabeling Protocol and Class Balancing

Style labels were assigned automatically by a rule-based labeling procedure applied to the raw text of every document. The rule uses a closed lexicon of explicit humor markers encoded as regular expressions and organized into five groups: basic laughter expressions (wkwk, haha, hehe, lol, lmao); Indonesian humor slang (ngakak, kocak, lucu, rekeh, gokil, konyol); humor intensifiers (bikin gakak, ketawa sampai nangis, parah banget); casual or joking expressions (bercanda, just kidding, santuy, lebay); and other expressive reactions that commonly accompany a humorous register in informal Indonesian writing. A function named `auto_label()` processes each document by first converting the text to lowercase, so that matching does not depend on capitalization, and then checking the lowercased text against each pattern in turn.

The decision rule is deterministic and can be stated formally. For a document d with lowercased form $l(d)$ and marker lexicon M , the assigned label is humorous as soon as one pattern m in M matches $l(d)$, and serious if the whole lexicon is exhausted without a match. No threshold, frequency weighting, or positional constraint is applied; a single match is sufficient. The rule is deliberately high-precision and low-recall with respect to the underlying construct: it captures humor that the writer has explicitly signalled and does not attempt to detect unmarked satire, irony, or sarcasm. Documents assigned to the serious class are therefore more precisely described as texts without explicit humor markers, and this is the sense in which the term is used throughout the paper. A rule-based scheme was preferred over a third-party classifier because it is transparent and fully traceable, so that the transformation from the original sentiment categories to communication-style categories can be audited explicitly rather than delegated to a black box.

Table I illustrates the behaviour of the rule on constructed examples, including two ambiguous cases. The first ambiguous case is a document that expresses a clearly negative sentiment yet carries an explicit humor marker; the rule labels it humorous, which demonstrates concretely that the style label is independent of polarity. The second is an ironic complaint that carries no surface marker; the rule

labels it serious, which is a known false negative of the procedure and is discussed further in Section III.D.

TABLE I
ILLUSTRATIVE EXAMPLES OF THE RULE-BASED RELABELING
PROCEDURE

No.	Text (Indonesian)	Marker matched	Assigned label	Basis of the decision
1	wkwk gagal fokus liat komentar di bawah	wkwk	Humorous	Explicit laughter marker; intent to amuse
2	receh banget tapi akhirnya ketawa juga	receh	Humorous	Explicit humor slang
3	pelayanan di kantor ini lambat dan tidak profesional	–	Serious	No marker; evaluative complaint
4	pemerintah menaikkan tarif mulai bulan depan	–	Serious	No marker; informative statement
5*	haha parah sih pelayanannya, nunggu tiga jam baru dilayani	haha	Humorous	Negative sentiment but explicit marker: style is not polarity
6*	mantap sekali ya, cuma tiga jam antre buat fotokopi selembar	–	Serious	Irony, no surface marker: known false negative of the rule

Applying the rule to the 12,679 documents produced a heavily skewed distribution: 604 documents (4.76%) were labeled humorous and 12,075 (95.24%) serious, a ratio of approximately 1:20. A model trained directly on this distribution could reach 95.24% accuracy simply by always predicting the majority class, so the corpus was balanced to 5,000 documents per class. The humorous class was expanded from 604 to 5,000 by random oversampling with replacement. To prevent the models from merely memorizing identical repeated sentences, each duplicated document was then augmented by shuffling the words in the middle of the sentence while keeping the first and the last word fixed, so that the sentence structure is not altered drastically; a subsequent check confirmed that no two documents in the balanced corpus remained textually identical. The serious class was reduced from 12,075 to 5,000 by random sampling without replacement. A fixed random state of 42 was used throughout the resampling so that the balanced corpus is reproducible. No synthetic text was generated, so every document derives verbatim from the original corpus. Table II reports the distribution before and after balancing.

TABLE II
CLASS DISTRIBUTION BEFORE AND AFTER BALANCING

Class	Before (n)	%	Balancing method	After (n)	%
Humorous	604	4.76	Random oversampling with replacement + word-order augmentation	5,000	50.00

Serious	12,075	95.24	Random undersampling without replacement	5,000	50.00
Total	12,679	100.00	–	10,000	100.00

The augmentation removes exact duplication but not the underlying dependence between instances: the 5,000 humorous documents still derive from only 604 unique source texts, so each source text is represented on average by about 8.3 variants. The consequences of this for the interpretation of the reported test scores are examined explicitly in Section II.D.

It must be acknowledged that these automatically generated labels were not validated against independent human annotation. Humor is a strongly context-dependent construct and a lexical rule can only approximate it, so the scores reported in Section III measure the ability of the models to reproduce this operational definition rather than an ability to recognize humor in general; a high accuracy against automatically generated labels is not by itself evidence of humor understanding. A validation study in which a stratified random sample of the corpus is annotated by at least two independent native speakers, and in which the agreement between the annotators and between the rule and the human consensus is quantified with Cohen's Kappa [22], is the necessary next step; it is recorded among the limitations in Section III.D and among the recommendations in Section IV.

C. Data Preprocessing

The *preprocessing* stage is performed to clean and standardize the text before it is used in model training. This stage includes changing lowercase letters, removing URLs/mentions/hashtag symbols, removing non-alphabetic characters, normalizing non-standard words, and removing label trigger words (e.g. "haha", "wkwk") to prevent *label leakage*. A summary of the *preprocessing* stages is presented in Table III.

TABLE III
TAHAPAN DATA PREPROCESSING

No.	Stages	Purpose
1	Lowercasing	Standardize text shapes
2	Removal of URLs & mentions	Removing irrelevant elements
3	Hashtag symbol removal	Retaining the meaning of words without symbols
4	Removal of non-alphabetic characters	Cleans numbers and punctuation
5	Normalization of slang	Standardizing the shape of the standard word
6	Label trigger word removal	Prevent label leakage
7	Elimination of blank lines	Ensure data validity

After preprocessing, 7 documents of the humorous class became empty as a result of the removal of the label trigger words and were discarded, leaving 4,993 humorous and 5,000 serious documents. To preserve the balanced distribution, 7 humorous documents were re-sampled, so that the final corpus used for training and testing consists of 10,000 documents, evenly divided into 5,000 humorous and 5,000 serious documents.

The removal of the marker words in stage 6 is essential rather than cosmetic. Because the markers define the label, leaving them in the model input would allow both architectures to rediscover the labeling rule and reach near-perfect accuracy without learning anything about communication style, which would make the evaluation meaningless. After their removal the models must rely on the remaining lexical, syntactic, and contextual evidence, which is precisely the ability the experiment is intended to measure.

D. Data Splitting, Tokenization, and Leakage Control

The data were divided into training data (80%) and testing data (20%) in a stratified manner so that the class distribution remained balanced, resulting in 8,000 training documents (4,000 humorous, 4,000 serious) and 2,000 testing documents (1,000 humorous, 1,000 serious). The text was then converted into a numerical representation through tokenization using a Tokenizer with a maximum vocabulary of 10,000 words and an out-of-vocabulary (OOV) token. The tokenizer was fitted on the training partition only, not on the complete corpus, so that no vocabulary information from the test partition entered the model; this step identified 10,699 unique word types in the training data, of which the 10,000 most frequent were retained. Sequences were then standardized in length through padding of up to 100 tokens, producing matrices of 8,000 x 100 for training and 2,000 x 100 for testing.

The order in which the pipeline stages are executed determines whether information can leak from the test partition into training, and it is therefore stated explicitly. Preprocessing is deterministic and operates on each document separately, so it cannot transfer information between documents, and the tokenizer vocabulary was built from the training partition alone. Class balancing, however, was applied to the whole corpus before the split. Although the augmentation step guarantees that no two documents are textually identical, all 5,000 humorous instances derive from the same 604 source texts, so near-identical variants of one source sentence can be assigned to the training and to the testing partition simultaneously. The test scores for the humorous class should therefore be read as optimistically biased. The methodologically stricter order, in which the original imbalanced corpus is divided first and oversampling is applied to the training partition only, is recorded as a limitation in Section III.D and as a priority for follow-up work.

Model selection relied on a validation set consisting of 20% of the training partition, whose loss and accuracy are reported in Tables VI and VII and which drove both the EarlyStopping and the ReduceLROnPlateau criteria; the testing partition was reserved for the final evaluation reported in Section III.C. Because both architectures contain several hyperparameters, a stratified k-fold cross-validation over the training partition would nevertheless give a more reliable estimate than a single split, and its absence is noted among the limitations.

The vocabulary size of 10,000 types and the maximum sequence length of 100 tokens can be justified empirically from the corpus itself. Tokenization identified 10,699 unique word types in the training partition, so a vocabulary of 10,000 retains all but the 699 rarest types and the out-of-vocabulary token absorbs only a thin tail of spelling variants. The sequence length of 100 tokens follows common practice for short informal social-media text and keeps the embedding matrix small enough to be trained from scratch on ten thousand documents. Both values nevertheless bound the linguistic information available to the models, and no ablation over them was carried out in this study; this is stated among the limitations.

E. Model Architecture

This study applies two model architectures. The single CNN model begins with an Embedding layer that maps each token to a 128-dimensional vector, followed by two successive Conv1D blocks with 96 and 48 filters respectively, a kernel size of 3, and ReLU activation. Each convolutional block is followed by a Dropout layer with a rate of 0.3, with MaxPooling1D after the first block and GlobalMaxPooling1D after the second. The extracted features are passed to two Dense layers of 64 and 32 neurons, followed by Dropout rates of 0.4 and 0.3 respectively, before a single-neuron output layer with sigmoid activation produces the binary probability. The model contains 1,336,081 trainable parameters.

The CNN-LSTM hybrid model uses the same Embedding layer (10,000-word vocabulary, 128 dimensions, sequence length 100), followed by a SpatialDropout1D layer with a rate of 0.2. A single Conv1D layer with 96 filters, a kernel size of 3, and ReLU activation extracts local features, followed by Dropout of 0.3 and MaxPooling1D. The resulting feature maps are passed to a Bidirectional LSTM layer with 64 units per direction, which produces a 128-dimensional representation capturing word-order relationships in both the forward and the backward direction. Two Dense layers of 64 and 32 neurons, with Dropout rates of 0.4 and 0.3, precede the sigmoid output layer. The model contains 1,409,761 trainable parameters. Both models were compiled with the Adam optimizer, a learning rate of 0.001, and a binary cross-entropy loss, and were trained for a maximum of 25 epochs with a batch size of 32, holding out 20% of the training data as a validation set. Two callbacks were applied identically to both models: EarlyStopping

monitoring validation loss with a patience of 4 epochs and restoration of the best weights, and ReduceLROnPlateau halving the learning rate after 2 epochs without improvement, down to a minimum of 0.000001. The architecture of the CNN-LSTM hybrid model is shown in Figure 1.

The configuration shared by the two architectures, and the points at which they differ, are summarized in Table V. Every setting that is not listed as differing between the two models is identical, so that the presence of the Bidirectional LSTM layer is the only systematic difference between them and the observed performance gap can be attributed to it rather than to divergent training regimes.

TABLE V
ARCHITECTURE AND HYPERPARAMETER CONFIGURATION OF BOTH MODELS

Component / hyperparameter	Single CNN	Hybrid CNN-LSTM
Vocabulary size	10,000	10,000
Input sequence length	100	100
Embedding dimension	128	128
SpatialDropout1D	–	0.2
Convolutional layers	2 x Conv1D (96 and 48 filters)	1 x Conv1D (96 filters)
Kernel size	3	3

Convolution activation	ReLU	ReLU
Dropout after convolution	0.3	0.3
Pooling	MaxPooling1D + GlobalMaxPooling1D	MaxPooling1D
Recurrent layer	–	Bidirectional LSTM, 64 units per direction
Dense layers	64 and 32 neurons	64 and 32 neurons
Dropout after Dense layers	0.4 and 0.3	0.4 and 0.3
Output layer	1 neuron, sigmoid	1 neuron, sigmoid
Loss function	Binary cross-entropy	Binary cross-entropy
Optimizer	Adam	Adam
Learning rate	0.001	0.001
Batch size	32	32
Validation split	20% of training data	20% of training data
Maximum epochs	25	25
EarlyStopping	val_loss, patience 4, restore best weights	val_loss, patience 4, restore best weights
ReduceLROnPlateau	factor 0.5, patience 2, min lr 0.000001	factor 0.5, patience 2, min lr 0.000001
Trainable parameters	1,336,081	1,409,761

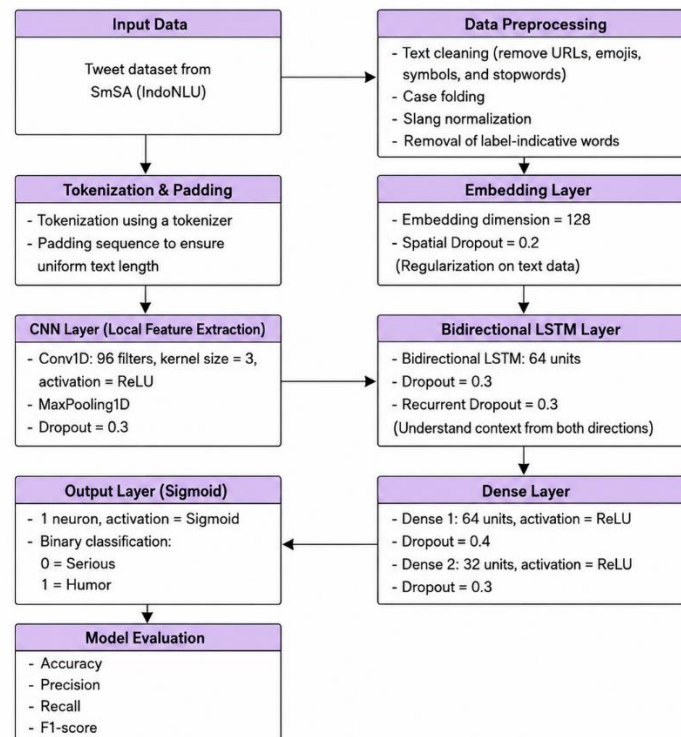


Figure 1. A CNN-LSTM hybrid model architecture that combines layer embedding, local feature extraction by CNN, and word sequence context modeling by a Bidirectional LSTM.

III. RESULTS AND DISCUSSION

A. Data Description and Distribution

The initial dataset amounted to 12,679 texts in Indonesian. After keyword-based autolabeling, the class distribution showed a significant imbalance, namely 604 humor data (4.76%) and 12,075 serious data (95.24%), as

shown in Figure 2. This imbalance has the potential to cause biased models to study the patterns of the majority class, so that the resulting accuracy can be seen as high but the ability to recognize the minority class (humor) is low.

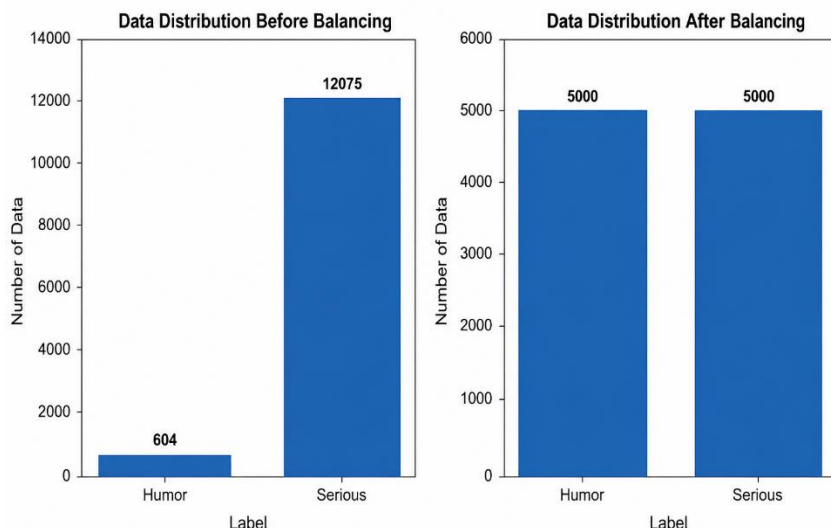


Figure 2. Data distribution before and after the data balancing process.

To overcome this, data balancing was carried out through oversampling of the humorous class and undersampling of the serious class so that each class amounted to 5,000 documents, a total of 10,000. After preprocessing, 7 humorous documents became empty and were removed, and the class was re-balanced back to 5,000, so that the final corpus consists of 10,000 documents (5,000 humorous and 5,000 serious). A summary of the distribution of training and testing data is presented in Table IV.

TABLE IV
SHARING TRAINING AND TESTING DATA

Data Type	Quantity	Humorous	Serious
Total data	10.000	5.000	5.000
Data training (80%)	8.000	4.000	4.000
Data testing (20%)	2.000	1.000	1.000

B. Model Implementation and Training

The single CNN model was trained with a maximum of 25 epochs, but training stopped at the 7th epoch because EarlyStopping detected that the validation loss was no longer improving. The final training accuracy reached 99.75% while the validation accuracy was 90.87%, with the best weights obtained at the 3rd epoch, which produced the lowest validation loss (0.2262) and the highest validation accuracy (91.19%). A summary of the training results is presented in Table VI.

TABLE VI
SINGLE CNN MODEL TRAINING RESULTS

Epoch	Loss	Accuracy	Val Loss	Val Acc
1	0,6523	0,5941	0,5439	0,7538
2	0,3259	0,8662	0,2584	0,9019
3	0,1030	0,9659	0,2262	0,9119
4	0,0535	0,9831	0,2545	0,9112
5	0,0348	0,9894	0,3055	0,9050
6	0,0149	0,9950	0,3728	0,9031
7	0,0075	0,9975	0,3400	0,9087

The rise in validation loss after the 3rd epoch, from 0.2262 to 0.3400 at the point where training stopped, indicates overfitting of the single CNN model, a condition in which the model is over-adjusted to the training data so that its generalization ability decreases. The pattern is confirmed by the training accuracy, which continued to climb to 99.75% while the validation accuracy remained flat in the 90-91% range.

The CNN-LSTM hybrid model was trained with the same configuration, but training stopped at the 8th epoch with a training accuracy of 99.69% and a validation accuracy of 93.69%. The best weights were obtained at the 4th epoch, which produced the lowest validation loss (0.2233); the highest validation accuracy of the run, 94.19%, occurred at the 5th epoch but was accompanied by a higher validation

loss and was therefore not selected. The training log is presented in Table VII.

TABLE VII
CNN-LSTM HYBRID MODEL TRAINING RESULTS

Epoch	Loss	Accuracy	Val Loss	Val Acc
1	0,6230	0,6367	0,4766	0,7837
2	0,3030	0,8875	0,3178	0,8850
3	0,1247	0,9567	0,3714	0,8913
4	0,0719	0,9772	0,2233	0,9275
5	0,0395	0,9895	0,2280	0,9419
6	0,0283	0,9911	0,4981	0,8963
7	0,0165	0,9948	0,5010	0,9131
8	0,0099	0,9969	0,3780	0,9369

Compared to the single CNN model, the CNN-LSTM hybrid model showed a more stable learning process, with a consistent *validation accuracy* in the range of 92–95% and no drastic decline in validation performance as the single CNN model, which showed better generalization ability due to the contribution of the LSTM layer in understanding the context of the word sequence.

C. Model Evaluation and Comparison

On the test set of 2,000 documents, the single CNN model obtained an accuracy of 91.60%, a precision of 91.61%, a recall of 91.60%, and an F1-score of 91.60%. The confusion matrix in Figure 3 shows that the model correctly predicted 908 of 1,000 humorous documents (90.8%) and 924 of 1,000 serious documents (92.4%), with 92 humorous documents (9.2%) misclassified as serious and 76 serious documents (7.6%) misclassified as humorous.

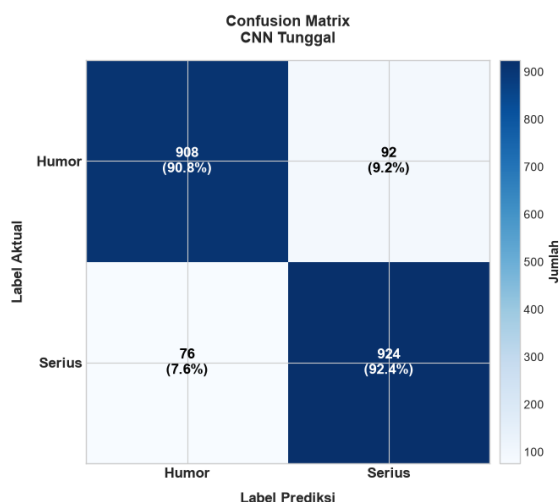


Figure 3. Confusion matrix of a single CNN model on data testing.

The CNN-LSTM hybrid model produced higher performance, with an accuracy of 94.10%, a precision of 94.19%, a recall of 94.10%, and an F1-score of 94.10%. Based on the confusion matrix in Figure 4, the model

correctly predicted 963 of 1,000 humorous documents (96.3%) and 919 of 1,000 serious documents (91.9%), with 37 humorous documents (3.7%) misclassified as serious and 81 serious documents (8.1%) misclassified as humorous. Compared with the single CNN, the hybrid model shows a substantial gain in humorous-class recognition, from 90.8% to 96.3%, accompanied by a small decrease in serious-class recognition, from 92.4% to 91.9%.

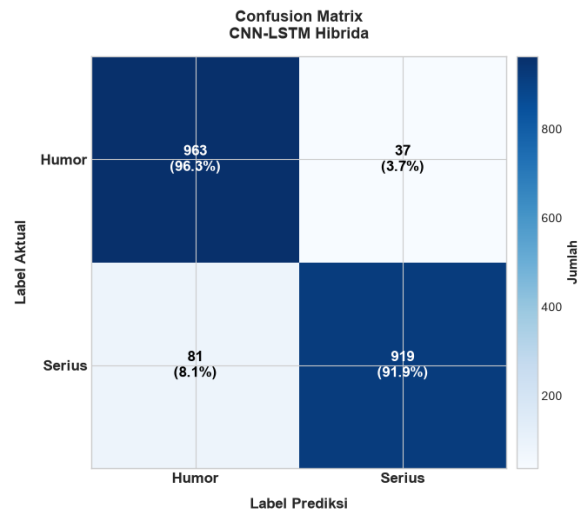


Figure 4. Confusion matrix of the CNN-LSTM hybrid model on data testing.

Aggregate accuracy conceals which of the two styles is the harder one to recognize, so Table VIII reports precision, recall, and F1-score separately for each class, computed from the confusion matrices in Figures 3 and 4. The macro-averages of these per-class values reproduce exactly the aggregate figures in Table IX, which confirms that the confusion matrices and the reported metrics originate from the same evaluation run. The error profile of the single CNN is close to symmetric, with a small deficit on the humorous class. The hybrid model changes the profile rather than merely shifting it upward: humorous-class recall rises by 5.5 percentage points while serious-class recall falls by 0.5, and precision moves in the opposite direction on each class. The two architectures therefore do not only differ in overall accuracy; they distribute their errors differently.

TABLE VIII
PER-CLASS EVALUATION METRICS DERIVED FROM THE
CONFUSION MATRICES

Model	Class	Precision	Recall	F1-score	Support
Single CNN	Humorous	92.28%	90.80%	91.53%	1,000
Single CNN	Serious	90.94%	92.40%	91.67%	1,000
Hybrid CNN-LSTM	Humorous	92.24%	96.30%	94.23%	1,000
Hybrid CNN-LSTM	Serious	96.13%	91.90%	93.97%	1,000

A comparison of the final performance of the two models across the evaluation metrics is presented in Table IX and visualized in Figure 5.

TABLE IX
COMPARISON OF MODEL TEST RESULTS

Model	Accuracy	Precision	Recall	F1-score
CNN Single	91,60%	91,61%	91,60%	91,60%
CNN-LSTM Hybrid	94,10%	94,19%	94,10%	94,10%
Difference	+2,50	+2,58	+2,50	+2,50

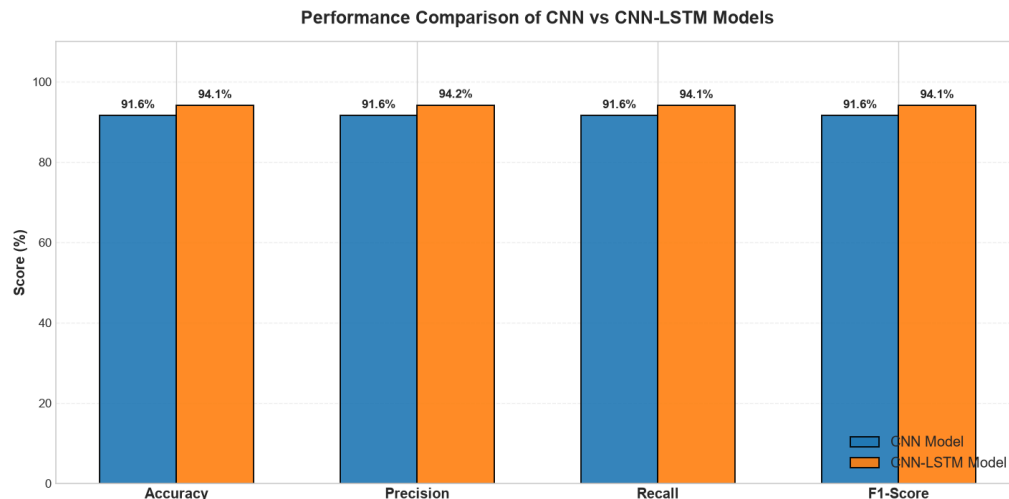


Figure 5. Comparison graph of test results of the single CNN model and the CNN-LSTM hybrid model.

The CNN-LSTM hybrid model obtained a higher evaluation score than the single CNN model across all test metrics, with an increase of 2.50 percentage points in accuracy, recall, and F1-score, and 2.58 percentage points in precision. The gain comes at a modest computational cost: the average inference time per prediction rose from 84.47 ms for the single CNN to 116.79 ms for the hybrid model.

D. Discussion

The improved performance of the CNN-LSTM hybrid model occurs because it combines CNN's ability to extract local features, such as word and phrase patterns, with LSTM's ability to understand the relationship between word sequences and sentence contexts in sequence. These findings are consistent with the results of Cliche [17], Geethanjali and Valarmathi [18], as well as Tabassum and Nunavath [19], which also show that the CNN-LSTM hybrid architecture consistently outperforms a single model on a variety of social media text classification tasks. The accuracy of the hybrid model in this study (94.10%) is indeed lower than the 97.8% accuracy reported by Geethanjali and Valarmathi [18], which integrate IChOA metaheuristic optimization algorithms and multimodal data, so that the application of similar hyperparameter optimization strategies has the potential to be a further development direction for this study.

The hybrid deep learning model in this study shows an advantage in capturing contextual relationships between words that are important for distinguishing humorous and serious communication styles, as the two categories cannot

always be distinguished based on specific keywords alone, but also depend on the phrase structure and the context of the sentence as a whole [20].

The confusion matrices also show where each architecture is weak. The single CNN model was slightly less accurate on the humorous class (90.8% correct) than on the serious class (92.4% correct), because once the explicit humor markers have been removed by the anti-leakage step, part of the humorous text retains a plain structure that closely resembles serious writing, and a model attending only to local features tends to miss it. Adding the Bidirectional LSTM layer corrected this weakness: humorous-class recall rose from 90.8% to 96.3%, while serious-class recall fell only marginally, from 92.4% to 91.9%. The overall gain of 2.50 percentage points is therefore obtained mainly by recovering humorous documents that the single CNN had missed, because the recurrent layer considers sentence context sequentially and in both directions rather than recognizing words in isolation. This reinforces the argument that humorous and serious communication styles carry a higher semantic complexity than conventional binary sentiment classification, and therefore require models capable of understanding context in greater depth [17], [18].

It should be emphasized, however, that accuracy and F1-score alone do not demonstrate how CNN and LSTM capture humor and context differently; they establish only that one configuration reproduces the labels better than the other. The claim that the hybrid model builds a more discriminative representation rests here on the shift in the per-class error profile reported in Table VIII, in which the

Bidirectional LSTM layer specifically repairs the humorous-class recall deficit of the single CNN. A stronger demonstration would require representation-level evidence, for example a comparison of the penultimate-layer embeddings of the two models under dimensionality reduction, an attribution analysis showing which tokens drive each decision, or an ablation in which the recurrent layer is replaced by a pooling layer of equivalent capacity. Such analyses are beyond the scope of the present experiments and are identified as the next step.

The error profile also indicates where the remaining misclassifications are likely to originate. Documents whose humor is carried by irony or sarcasm without a surface marker are assigned to the serious class by the labeling rule, so a model that has learned to recognize a light conversational register is scored as wrong even where its prediction is arguably right against the underlying construct; the same holds for documents that are humorous in register while carrying a strongly negative sentiment. A second group of errors originates in the pipeline rather than in the labels: slang and abbreviations such as gpp or bgt are mapped to the out-of-vocabulary token when they fall outside the 10,000-word vocabulary, Indonesian-English code-mixing places the humorous segment outside the vocabulary learned from the corpus, and emoji and emoticons, which carry a substantial part of the pragmatic signal in informal Indonesian writing, are discarded during preprocessing before the models ever see the text. The first group is a property of the labels and bounds the accuracy attainable by any model trained on them; the second is addressable by design and points towards retaining emoji as tokens, extending the slang normalization dictionary, and using subword tokenization.

Several limitations must be stated explicitly, because they bound the interpretation of the results. First, the labels are produced by a lexical rule and were not validated against independent human annotation, so they encode explicitly marked humor only; unmarked satire, irony, and sarcasm are systematically assigned to the serious class, and the reported scores measure the ability of the models to reproduce this operational definition rather than an understanding of humor as such. Second, the corpus is SmSA, a general benchmark of Indonesian user-generated text that is linguistically comparable to platform X content but was not collected from that platform, so the distribution of communication styles reported here need not match the distribution among X users. Third, class balancing was applied before the train-test split and used oversampling with replacement, so that although augmentation removed exact duplicates, near-identical variants of the same 604 source documents may appear in both partitions and the absolute scores for the humorous class are optimistically biased. Fourth, the classification is binary, so intermediate and mixed styles, in which a serious message carries humorous elements, cannot be represented at all. Fifth, each architecture was trained once on a single data partition and the difference of 2.50

percentage points was not subjected to significance testing, so the superiority of the hybrid model is demonstrated for this run and this partition rather than established as systematic; repeated runs with different random seeds, a paired significance test, and k-fold cross-validation are required for that stronger claim. Sixth, the comparison covers two deep learning architectures only, without classical baselines such as TF-IDF with logistic regression or a support vector machine, without other recurrent variants such as a unidirectional LSTM or a GRU, and without a pretrained Indonesian language model such as IndoBERT, so the reported improvement cannot yet be positioned within the wider landscape of Indonesian NLP methods. Seventh, the vocabulary size and the maximum sequence length were not varied, so their effect on performance is not known.

IV. CONCLUSION

This study succeeded in classifying humorous and serious communication styles in Indonesian-language social media text using the Convolutional Neural Network (CNN) model and the hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model, in which the recurrent component is a Bidirectional LSTM. The test results showed that the hybrid model performed better than the single CNN, with an accuracy and an F1-score of 94.10% against 91.60% and 91.60% for the single CNN. These findings suggest that combining the ability of CNN to extract local features with the ability of the Bidirectional LSTM to model the context of word sequences produces more effective text representations for distinguishing humorous from serious communication styles. These results hold with respect to labels produced by a lexical rule on a relabeled benchmark corpus and for a single training run of each architecture; they should be read as evidence about the separability of explicitly marked humorous texts from texts without such markers, not as evidence that the models understand humor.

This research also contributes a transparent rule-based relabeling protocol for constructing communication-style categories from the SmSA corpus, an explicit operational definition of the humorous and serious categories, and a comparative evaluation of the CNN and CNN-LSTM architectures that reports per-class as well as aggregate performance. The claims are nonetheless bounded by the limitations set out in Section III.D: the labels capture explicitly marked humor only and have not been validated against human annotation, the corpus is a proxy for rather than a sample of platform X, balancing preceded the train-test split, the classification is binary, and the comparison rests on a single training run of two architectures. The proposed model should therefore be regarded as an effective framework for automated communication-style classification on this relabeled corpus rather than as a validated framework for Indonesian content on platform X in general. Further research is accordingly recommended to validate the labels against human annotation using Cohen's Kappa, to repeat the training with several random seeds and test the

difference between the two models for statistical significance, to compare the hybrid model against classical TF-IDF baselines, other recurrent variants, and Transformer-based models such as BERT and IndoBERT, to collect data directly from platform X, and to extend the label set to sarcasm, irony, and mixed styles so that the classification becomes both more accurate and more representative of real communicative behaviour.

REFERENCES

- [1] DataReportal, Digital 2026: Indonesia. We Are Social & Meltwater, 2026.
- [2] M. I. Irwanti and A. P. Putranto, "Satire and Humor in Indonesian Political Culture: Negotiating Government Criticism in Digital Journalism," *Frontiers in Communication*, vol. 11, 2026.
- [3] Y. Hu, C. S. B. Ngai, A. C. K. Lee, and J. A. Banas, "Humor in Social Media Health Communication: A Systematic Review of Strategic Uses and Effects," *Behavioral Sciences*, vol. 16, no. 4, p. 509, 2026.
- [4] S. Nurrachmah, "Gen Z's Self-Expression and Digital Identity Strategies on Social Media X: A Quantitative Study of Communication Styles and Audience Engagement," *Jurnal Multidisiplin Indonesia*, vol. 4, no. 7, 2025.
- [5] A. Hanna Asyifa, E. Sophia, F. A. Rachmah, and S. Hadi, "Analysis of Interpersonal Communication Styles on Social Media Platform X (A Netnographic Study of the MARAH-MARAH Community)," *AMK: Abdi Masyarakat UIKA*, vol. 4, no. 4, 2025.
- [6] E. W. Pamungkas, C. S. Wahyuni, I. Amal, D. Purworini, and B. S. Rintyarna, "Decoding Hate in Memes: Multimodal and Multitask Approaches for Low-Resource Indonesian Social Media," *PeerJ Computer Science*, vol. 12, 2026.
- [7] A. Cliche, "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs," *Proceedings of SemEval-2017*, 2017.
- [8] M. Geethanjali and K. Valarmathi, "An Improved Chimp Optimization Based CNN-LSTM Model for Multimodal Emotion Recognition in Social Media," *Expert Systems with Applications*, 2024.
- [9] S. Tabassum and V. Nunavath, "Hybrid Deep Learning Models for Multi-Class Cyberbullying Detection Using Social Media Text and Images," *Multimedia Tools and Applications*, 2024.
- [10] Z. Fasya, M. Daud, and L. Rosnita, "Analisis Sentimen Ulasan Masyarakat terhadap Aplikasi Sirekap 2024 pada Google Play Store Menggunakan Metode K-Nearest Neighbors (KNN)," *Rabit: Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 11, no. 1, pp. 1608–1627, 2026.
- [11] A. Z. Putri, M. Daud, and H. A. K. Aidilof, "Klasifikasi Penyakit Jantung Berbasis Data Rekam Medis Menggunakan Algoritma Local Mean K-Nearest Neighbor," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 12, no. 1, pp. 134–143, 2026.
- [12] A. P. Erwanda and A. Khaidar, "Convolutional Neural Network Model for Multi-Class Sentiment Analysis on Twitter with Deep Learning Approach," *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 5, no. 1, pp. 1965–1970, 2025.
- [13] A. Khaidar, "Analisis Sentimen di Instagram terhadap Menteri Keuangan Purbaya Yudhi Sadewa Menggunakan Metode Logistic Regression," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3S1, 2025.
- [14] A. AlDayel and W. Magdy, "Stance Detection on Social Media: State of the Art and Trends," *ACM Computing Surveys*, vol. 54, no. 11, 2022.
- [15] M. Bedi, S. Kumar, M. S. Akhtar, and T. Chakraborty, "Multi-modal Sarcasm Detection and Humor Classification in Code-Mixed Conversations," *ACL Findings*, 2021.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL*, 2019.
- [17] M. Cliche, "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs," in *Proc. 11th Int. Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, 2017, pp. 573–580.
- [18] R. Geethanjali and A. Valarmathi, "A Novel Hybrid Deep Learning IChOA-CNN-LSTM Model for Modality-Enriched and Multilingual Emotion Recognition in Social Media," *Sci. Rep.*, vol. 14, art. 22270, 2024.
- [19] I. Tabassum and V. Nunavath, "A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data," *Appl. Sci.*, vol. 14, no. 24, art. 12007, 2024.
- [20] A. TiaraSari and E. Haryatmi, "Penerapan Convolutional Neural Network Deep Learning dalam Pendeteksian Citra Biji Jagung Kering," *J. RESTI (Rekayasa Sist. dan Teknol. Inf.)*, vol. 5, no. 2, pp. 265–271, 2021.
- [21] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proc. 1st Conf. Asia-Pacific Chapter ACL and 10th Int. Joint Conf. NLP (AACL-IJCNLP)*, 2020, pp. 843-857.
- [22] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.