Emotion-Based Emoji Classification Using a Pretrained Indonesian RoBERTa-Base Model

Nur Hayatin1,a), Briansyah Setio Wiyono1), Renita Astri2)

1Informatics Department, Engineering Faculty, University of Muhammadiyah Malang, Indonesia

2Prodi Sistem Informasi, Fakultas Farmasi Sains dan Teknologi, Universitas Dharma Andalas, Indonesia

a) Corresponding author: [noorhayatin@umm.ac.id](mailto:noorhayatin@umm.ac.id)

**Abstract.**  Emojis play a crucial role in conveying emotions in social media communication and have become an integral component of natural language processing (NLP) tasks involving sentiment and emotion analysis. Traditionally, emojis are often discarded during text preprocessing, as they are considered noise that may hinder model performance. However, recent research indicates that emojis can enhance classification accuracy when properly utilized. Despite their potential, publicly available emoji dictionaries are typically limited to textual descriptions, lacking emotional categorization. Manual grouping of emojis into emotional classes is time-consuming and inefficient. This study proposes an automatic emoji classification method based on emotional categories using a deep learning approach. We employ a fine-tuned version of the pre-trained IndoBERTa model, a variant of RoBERTa for the Indonesian language, to classify emojis based on emotional context. Experimental results demonstrate that variations in input features significantly impact classification performance. The model configuration using Feature 6 yields the highest accuracy, achieving a precision score of 70%, indicating its effectiveness in enhancing emotion classification tasks involving emojis.

**Keywords:** Emojis, emotion analysis, deep learning, sentiment analysis.

# INTRODUCTION

Artificial intelligence (AI) has been widely applied across various domains, including healthcare. One of the most pressing issues in the healthcare sector today is mental health, which is also recognized as a key priority within the health-related goals of the Sustainable Development Goals (SDGs). Natural Language Processing (NLP), a subfield of AI, focuses on processing and analyzing data through human language. Several studies have applied NLP techniques to mental health, such as analyzing the mental state of Twitter users through their microblog statuses [1][2], detecting abusive language on social media platforms [3], and classifying emotions in Twitter data [4].

In the field of artificial intelligence, data is a crucial component that must be processed into valuable insights. Specifically in NLP, particularly in analyzing social media data, emojis play a significant role as representations of users’ feelings and emotions. Emojis are pictorial characters commonly used in social media to express emotions in textual messages [5]. However, in many studies, emojis are often removed during the preprocessing stage, as they are considered noise that may interfere with model performance. Contrary to this common practice, research by [6] has demonstrated that incorporating emoji data in classification tasks can significantly enhance model performance.

Therefore, there is a need for a dedicated emoji dictionary that can serve as a reference for researchers in the NLP domain. Our analysis reveals that publicly available emoji dictionaries, such as [7], primarily function as collections of emojis and their corresponding names. In order to use them in modeling tasks, researchers must manually group the emojis into specific categories, such as sentiment or emotion classes. This manual grouping process is time-consuming and labor-intensive, and it often results in limited emoji coverage due to practical constraints.

As the development of Indonesian-language NLP continues to progress, emotion analysis becomes increasingly relevant across multiple sectors, including business, social research, and technological applications. Motivated by this background, this study aims to automatically classify emoji data based on emotion categories using a deep learning classification approach. The expected outcome of this research is the creation of an emotion-labelled emoji dataset, which can contribute to the enrichment of NLP resources and be utilized by researchers in the broader AI community.

The research questions guiding this study are: 1) How can emoji data that reflects Indonesian language usage be obtained?, 2) What classification approach can be implemented to categorize emojis into emotion classes?, 3) How can the performance of emotion classification on emoji data be evaluated using deep learning with various feature configurations?. The primary objective of this research is to develop an automatic classification system for emojis based on emotion labels using a pre-trained RoBERTa model tailored for Indonesian emotion analysis. The study adopts a deep learning-based classification approach to achieve this goal.

# literature review

Emotion analysis is a technique within Natural Language Processing (NLP) used to identify and categorize emotions expressed in textual data. Its primary goal is to detect emotional states conveyed in various forms of written content such as reviews, social media posts, emails, and articles.

Emotion analysis typically involves three main stages: The first is text Processing, where the text is segmented into smaller units such as words or phrases; The second is Emotion Detection, where NLP models are applied to detect emotional states such as happiness, sadness, anger, fear, surprise, or neutrality. These emotions are often categorized using frameworks such as Plutchik’s Wheel of Emotions or Ekman’s Six Basic Emotions; and the third is Emotion Classification, where Machine learning techniques are employed to classify emotions with high accuracy, such as rule-based systems or deep learning.

Common applications of emotion analysis include marketing, where Companies analyze customer reviews or feedback to gauge emotional reactions to products or services; Reputation Management, Monitoring emotional tones on social media helps organizations respond to negative comments or amplify positive sentiments; and Chatbot Interaction, Emotion-aware chatbots can respond more empathetically by adjusting tone and content based on user emotions.

Emotion analysis for Indonesian language data operates similarly to analysis in other languages but requires sensitivity to the linguistic structure and usage patterns specific to Bahasa Indonesia. While sentiment analysis focuses on identifying the polarity of a text (positive, negative, or neutral), emotion analysis goes deeper by identifying more nuanced emotional categories such as anger, joy, love, fear, and sadness.

Emojis have become a ubiquitous component of digital communication, conveying emotions, tone, and intent in social media, messaging apps, and emails. As such, researchers have explored various approaches to classify emojis using natural language processing (NLP), sentiment analysis, and deep learning methods. Initial work on emoji usage focused on understanding their semantics and correlation with sentiment. Novak et al. (2015) conducted a large-scale sentiment analysis using over 1.6 million emoji-labeled tweets, showing that emojis can reliably signal user sentiment and thus be used as noisy labels for training classifiers [8]. Similarly, Eisner et al. (2016) examined how emoji semantics vary across users and cultures, emphasizing the ambiguity and context-dependence of emoji interpretation [9]. With the rise of deep learning, emoji prediction and classification tasks have shifted toward neural models. Meanwhile, Arifiyanti and Wahyuni (2020) converted emoji and emoticon into its lexicon sentiment which aims to make classification model performance more stable [10].

**Deep Learning for Emoji Classification**

Deep learning for emotion analysis involves using neural network architectures to automatically detect and classify emotions from text, audio, or images with high accuracy. These models, particularly those based on deep neural networks, are capable of capturing complex patterns and semantic relationships in data.

Advantages of deep learning in emotion analysis include: Complex Pattern Recognition, Deep learning models can uncover subtle and intricate relationships in text that traditional models often miss; High Generalizability, Given sufficiently large and diverse training datasets, these models can generalize well to new, unseen data; and Contextual Understanding, Models like transformers allow deep contextual comprehension, which is essential for accurate emotion detection.

Barbieri et al. [11] developed deep LSTM-based architectures for emoji prediction in Twitter texts, achieving significant improvements over bag-of-words baselines. Their work inspired follow-up studies leveraging transformer-based models. For example, Felbo et al. [12] introduced the DeepMoji model, a pre-trained attention-based network trained on 1.2 billion tweets containing emojis, which was later fine-tuned for emotion classification and sarcasm detection tasks. Another direction explored multimodal emoji classification, integrating both text and image signals. Wang et al. [13] proposed a fusion model that utilizes both textual and visual embeddings to better understand emoji usage in social media platforms such as Instagram and TikTok. Meanwhile, contextual language models like BERT and RoBERTa have also been fine-tuned for emoji-related tasks, showing strong performance [14].

Lastly, challenges remain in disambiguating emojis with multiple meanings and handling culturally specific emoji use. Pohl et al. [15] highlighted the variations in emoji usage across demographic groups, which poses a challenge for creating universally accurate classifiers. These studies collectively underscore the growing complexity and richness of emoji classification, motivating continued exploration into more robust, context-aware models.

**Existing Techniques**

Several prior studies have explored the intersection of emojis and emotion analysis. A comparative overview between the proposed study and existing research is presented in TABLE 1.

**TABLE 1**. Comparison of the Proposed Study with Existing Related Works

|  |  |  |  |
| --- | --- | --- | --- |
| **Ref.** | **Data** | **Technique** | **Label** |
| [6] | Emoji | Manual | Sentiment: positive, negative, and neutral |
| [7] | Text | Automatic | Emotion: anger, joy, love, fear, sadness, neutral |
| [5] | Emoji | Manual | Non-label |
| proposed | Emoji | Automatic | Emotion: anger, joy, love, fear, and sadness |

Below is a summary of key related works. In [6], an emoji list was used as one of the features for sentiment classification in textual data. The emojis were manually listed and grouped into three sentiment categories: positive, negative, and neutral. However, a major limitation of this study was the small number of emojis included in the dataset. Another study that utilized emoji data was conducted by [5], which combined linguistic, semantic, and lexical features for Twitter data classification. Emojis were among the features employed; however, the emoji set was limited and lacked proper labelling, which reduced the effectiveness of their integration. In contrast to the two previous studies, [7] focused on classifying Twitter data based on emotions without including emojis. This study used six emotion classes: anger, sadness, happiness, love, fear, and neutral. The proposed study adopts the same set of emotion classes, except for the neutral category.

This proposed research differs significantly from the aforementioned works. It aims to automatically classify emoji data based on emotion categories using deep learning techniques. The expected output is an emotion-labeled emoji dataset.

# METHODS

The research methodology of this study is illustrated in FIGURE 1, which consists of three primary stages: data preparation, classifier modelling, and evaluation. Each stage is designed to systematically support the research objective of building an emotion-based emoji classification model using a deep learning approach.

A diagram of a deep learning model

Description automatically generated

**FIGURE 1**. Emotion-based emoji classification model using a deep learning approach

The primary data utilized in this study consists of emoji data. The emoji data is sourced from publicly available emoji dictionaries, which are openly accessible and provide a comprehensive list of emoji characters along with their associated meanings. The dataset structure is illustrated in FIGURE 2, which includes relevant attributes such as the emoji symbol and the description.

For the classification model, we employ a deep learning approach by implementing the pre-trained "Indonesian RoBERTa-base" model for emotion classification. The use of the Indonesian RoBERTa model in emotion analysis represents a deep learning approach based on the transformer architecture. RoBERTa (Robustly Optimized BERT Pretraining Approach) is a refined variant of BERT that has been pre-trained with a larger dataset, for a longer duration, and with optimizations such as dynamic masking and the removal of the Next Sentence Prediction objective. Key advantages of RoBERTa over BERT include More Extensive Pretraining, it leverages larger corpora and longer training cycles, resulting in improved language understanding; Dynamic Masking: Training data is masked dynamically for each batch, providing greater variability and generalization; The Indonesian RoBERTa model is a version of RoBERTa that has been pre-trained or fine-tuned specifically on Indonesian language datasets. This specialization allows the model to better comprehend local linguistic structures, idioms, and nuances.

A screenshot of a phone

Description automatically generated

**FIGURE 2**. The overview of the Emoji dataset

Benefits of using Indonesian RoBERTa for emotion analysis include: Deep Contextual Understanding, The model excels at capturing meaning and sentiment in context, essential for accurate emotion classification; Local Language Adaptation, As it is trained on Indonesian data, the model is more adept at understanding Indonesian grammar, expressions, and cultural references; and Domain Versatility, The model is applicable across various domains—such as emotion detection in social media, customer reviews, or formal documents. By employing Indonesian RoBERTa, emotion analysis in the Indonesian language can be conducted with greater precision, particularly in capturing subtle emotional expressions and contextual meanings.

Our model is designed to identify five emotion categories, namely: anger, sadness, joy, love, and fear. In the first step, the preprocessed text is input into the Indonesian RoBERTa model, which processes the text to generate vector-based representations of words and sentences. These contextual embeddings capture the semantic nuances of the input.

The IndoRoBERTa Emotion Classification model is an emotion classification system based on the IndoRoBERTa architecture. This model was trained using the EmoT dataset from the IndoNLU benchmark. By applying transfer learning, the base IndoRoBERTa model was adapted to function as an emotion classifier. The model was trained for 7 epochs with a learning rate of 2e-5, resulting in various performance metrics, as shown in the section below.

Subsequently, the emotion classification is performed using the output vectors from RoBERTa, which are passed through an additional classification layer (e.g., a fully connected layer or softmax layer) to predict the most relevant emotion label, such as joy, sadness, anger, or fear, according to the defined classification schema. The model is trained using a labelled dataset containing Indonesian-language text samples annotated with their corresponding emotion categories.

To evaluate the performance of the emoji classification, we will measure how well the model performs using a standard statistical metric, namely accuracy. This will involve cross-checking the predicted emotion classes produced by the model against human-annotated corrections provided by reviewers. As part of the initial evaluation plan, two human reviewers will be involved to assist in validating and reviewing the classification results.

# RESULTS AND DISCUSSION

To evaluate the model's performance, we employed common statistical metrics widely used in machine learning: precision, recall, F1-score, and accuracy. Accuracy is a metric to measure the ratio between the number of correct predictions and the total number of data predictions, while the formula of accuracy is defined as equation (1). Precision is a metric to calculate the proportion of true positive predictions out of all predicted positives, whether correct or not, while the formula of precision is defined as (2). Recall is a metric to measure the proportion of actual positives that were correctly predicted out of all actual positives, while the formula of recall is defined as (3). The F1-score represents the harmonic mean of precision and recall, while the formula of the F1-score is defined as (4), where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

(1)

(2)

(3)

(4)

We combined various input features to achieve optimal classification performance. The types of input features combined include emoji features, semantic meaning features, combined emoji and meaning, comparison of emoji and meaning, cleaned meaning, combined emoji and cleaned meaning, and comparison of emoji and cleaned meaning. TABLE 2 describes each input feature:

**TABLE 2.** Variations of Input Features

| **Feature** | | **Description** | **Example** |
| --- | --- | --- | --- |
| 1 | Meaning | Feature taken from the “meaning” column in the dataset, containing the textual interpretation of the emoji. | "Injured Face" |
| 2 | Emoji | Emoji icons represent various emotions. | 🤕 |
| 3 | Combined Meaning and Emoji | Combines the textual meaning and the emoji using a delimiter (e.g., ‘+’). | "Injured Face" + ' ' + 🤕 |
| 4 | Comparison of Meaning and Emoji | Compares the classification scores of both the meaning and the emoji, selecting the higher score. | Score (“Injured Face”), Score (🤕) |
| 5 | Cleaned Meaning | Preprocessed (cleaned) version of the meaning text. | "injured" |
| 6 | Combined Cleaned Meaning and Emoji | Combines cleaned meaning with emoji and selects the higher score. | "injured" + ' ' + 🤕 |
| 7 | Comparison of Cleaned Meaning and Emoji | Compares classification scores of cleaned meaning and emoji, selecting the higher one. | Score (cleaned meaning), Score (🤕) |

In FIGURE 3, the columns Label1 to Label6 display the classification results using the different input feature variations. These results show that different input combinations can lead to different emotion-label predictions. We present visualizations of the classification outcomes for each feature variation to better illustrate these differences.

A screenshot of a phone

Description automatically generated

**FIGURE 3**. Classification Results Using Different Feature Variations

On the other hand, FIGURE 4(a) presents a visual diagram showing the distribution of emotion classification results from the RoBERTa model using Feature 1 (input: Meaning). Based on the experiment results, the dominant label is "happy" which is represented in blue color, covering 671 emojis. Following with “sad" label represented in red color with 643 emojis. Other than that, the remaining labels (“anger”, “love”, and “fear”) appear less frequently, with 116, 46, and 37 emojis, respectively. Meanwhile, FIGURE 4(b) displays the classification result distribution using Feature 2 (Emoji only). In this case, the dominant emotion is again “happy” label with 1164 emojis, followed by “sad” with 335 emojis. The remaining emotions (“anger”, “love”, and “fear”) are represented by 1, 13, and 0 emojis, respectively. FIGURE 4(c) illustrates the results from using Feature 3 (input: Combined Meaning + Emoji). The emotion “happy” dominates once more, with 1211 emojis, followed by the “sad” label with 268 emojis, while the other labels “anger”, “love”, and “fear” are represented by 8, 23, and 3, respectively.

FIGURE 5(a) shows a visualization of the emotion distribution classified by the RoBERTa model using Feature 4. Based on the experiment results, it can be observed that the dominant label is the “happy” which is represented with blue color by 1.092 emojis, followed by the “sad” label represented with red color in 368 emojis. Meanwhile, the remaining labels (“anger”, “love”, and “fear”) are represented by 36, 13, and 4 emojis, respectively. FIGURE 5(b) displays the visualization of emotion distribution using Feature 5. The chart shows that the dominant label is “sad”, representing red color with 772 emojis, followed by the “happy” label which is represented by the blue color with 452 emojis, while the other labels “anger”, “love”, and “fear” are represented 182, 64, and 40, respectively.

|  |  |  |
| --- | --- | --- |
| A pie chart with different colored circles  Description automatically generated | **A blue and red pie chart with a number of different emotions  Description automatically generated** | A blue pie chart with red and blue text  Description automatically generated |
| (a) | (b) | (c) |

**FIGURE 4.** (a) Classification results using Feature 1 (input: meaning/text); (b) Classification results using Feature 2 (input: emoji); (c) Classification results using Feature 3 (input: merged features)

|  |  |
| --- | --- |
| A pie chart with numbers and a red circle  Description automatically generated | A pie chart with different colored circles  Description automatically generated |
| (a) | (b) |

**FIGURE 5**. (a) Classification results using Feature 4 (Input: feature combination); (b) Classification results using Feature 5 (Input: cleaned meaning/text)

FIGURE 6(a) illustrates the distribution of classified emotions using Feature 6. The “happy” label dominates with 1.104 emojis, followed by “sad” with 366 emojis. The remaining emotions are “anger”, “love”, and “fear” achieved 8, 28, and 7, respectively. FIGURE 6(b) presents the emotion distribution using Feature 7. The “happy” emotion again dominates with 967 emojis, followed by “sad” with 454 emojis, while the other labels “anger”, “love”, and “fear” achieved 61, 22, and 9, respectively.

|  |  |
| --- | --- |
| A pie chart with a number of different features  Description automatically generated with medium confidence | A pie chart with different colored circles  Description automatically generated |
| (a) | (b) |

**FIGURE 6**. (a) Classification results using Feature 6 (Input: cleaned merged features); (b) Classification results using Feature 7 (Input: cleaned feature combination)

FIGURE 7 presents the comparison of classification results using different features. It displays the emotion class distribution for all feature combinations using the RoBERTa model. Based on the experiment results, it demonstrated that the most features lead to the prediction of “happy” as the dominant class represented by blue bars, except for Feature 5, where the dominant emotion is “sad” represented by red bars.

A graph with different colored bars

Description automatically generated  
**FIGURE** **7**. Emotion class distribution comparison across all feature combinations

We also compared the evaluation results for all variations of input features, as shown in FIGURE 8. Several evaluation metrics were used, including accuracy, F1-score, precision, and recall. Based on the experiment results, the best performance was achieved by the RoBERTa model combined with Feature 6, illustrated by the yellow line. This combination produced the highest accuracy of 0.48, with an F1-score of 0.41, precision of 0.70, and recall of 0.48. The evaluation results for the other feature combinations can be seen in TABLE 4.

Based on the results of using the RoBERTa model in combination with various input features for emoji classification, it can be concluded that input variation significantly influences the model's predictions. Among all the tested feature combinations, the RoBERTa model with Feature 6 provided the best overall performance, achieving the highest accuracy.

A graph with colorful lines

Description automatically generated

**FIGURE** **8**. Comparison of evaluation metrics (accuracy, F1-score, precision, and recall) for emotion classification using the RoBERTa model with various input features.

**TABLE 4**. Evaluation Results of Emoji Classification using the RoBERTa Model with Different Input Features

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Evaluation** | **Feature 1** | **Feature 2** | **Feature 3** | **Feature 4** | **Feature 5** | **Feature 6** | **Feature 7** |
| Accuracy | 0.35 | 0.36 | 0.46 | 0.39 | 0.38 | 0.48 | 0.41 |
| F1-score | 0.32 | 0.25 | 0.38 | 0.29 | 0.35 | 0.41 | 0.37 |
| Precision | 0.39 | 0.19 | 0.58 | 0.45 | 0.48 | 0.7 | 0.6 |
| Recall | 0.35 | 0.36 | 0.46 | 0.39 | 0.38 | 0.48 | 0.41 |

# CONCLUSIONS

This study has successfully classified emoji data based on emotional categories. The approach utilized is a deep learning-based classification technique using the RoBERTa model, a pre-trained model tailored for classifying Indonesian text emojis. Based on the experimental results, which involved combining the RoBERTa model with various input feature sets, it can be concluded that input variation significantly influences the model's predictions. Among all combinations, the RoBERTa model with Feature 6 achieved the highest accuracy compared to the other feature sets.

As for future research, it is recommended to focus on improving the accuracy of the emotion classification model used for emoji classification. This improvement can be achieved by exploring other deep learning models that may be better suited for the task and by experimenting with emoji datasets from alternative sources.

# Acknowledgments

This research was funded by the *blockgrant* program of Engineering Faculty, University of Muhammadiyah Malang, through “PUSKAREKATEK 2024”.

# References

1. N. Hayatin, “Implementasi Multinomial Naive Bayes Untuk Klasifikasi Data Tweets Mengandung Term Depresi,” in Prosiding SENTRA (Seminar Teknologi dan Rekayasa), No. 6 (2021), pp. 344–349.
2. T. Lecompte and J. Chen, “Sentiment Analysis of Tweets Including Emoji Data,” in Proc. Int. Conf. Comput. Sci. Comput. Intell. CSCI (2017), pp. 793–798.
3. T. I. Sari, Z. N. Ardilla, N. Hayatin, and R. Maskat, “Abusive comment identification on Indonesian social media data using hybrid deep learning,” IAES Int. J. Artif. Intell. 11, 895–904 (2022).
4. M. S. Saputri, R. Mahendra, and M. Adriani, “Emotion Classification on Indonesian Twitter Dataset,” in Proc. Int. Conf. Asian Lang. Process. IALP (2018), pp. 90–95.
5. R. Wahyuni and I. Budi, “Combining Linguistic, Semantic and Lexicon Feature for Emoji Classification in Twitter Dataset,” Procedia Comput. Sci. 135, 194–201 (2018).
6. N. Hayatin, M. Mentari, and A. Izzah, “Opinion Extraction of Public Figure Based on Sentiment Analysis in Twitter,” IPTEK J. Eng. 1, 1 (2014).
7. Riccosan, K. E. Saputra, G. D. Pratama, and A. Chowanda, “Emotion dataset from Indonesian public opinion,” Data Brief 43, 108465 (2022).
8. P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, “Sentiment of emojis,” PLoS ONE 10, e0144296 (2015).
9. B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel, “emoji2vec: Learning emoji representations from their description,” arXiv:1609.08359 (2016).
10. A. A. Arifiyanti and E. D. Wahyuni, “Emoji and emoticon in tweet sentiment classification,” in Proc. 6th Information Technology International Seminar (ITIS) (IEEE, 2020), pp. 145–150.
11. F. Barbieri, F. Ronzano, and H. Saggion, “What does this emoji mean? A vector space skip-gram model for Twitter emojis,” in Proc. LREC (2016).
12. B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, “Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm,” arXiv:1708.00524 (2017).
13. F. Barbieri, M. Ballesteros, F. Ronzano, and H. Saggion, “Multimodal emoji prediction,” arXiv:1803.02392 (2018).
14. M. O. Nusrat, Z. Habib, M. Alam, and S. A. Jamal, “Emoji prediction in tweets using BERT,” arXiv:2307.02054 (2023).
15. H. Pohl, C. Domin, and M. Rohs, “Beyond just text: semantic emoji similarity modeling to support expressive communication👫📲😃,” ACM Trans. Comput.-Hum. Interact. 24, 1–42 (2017).