Performance Comparison of BERT, ALBERT and RoBERTa for Sentiment Analysis in Critical Pilot Communication Prior to Aviation Accident.

Very Sugiarto[1], Alvionitha Sari Agstriningtyas [2], Yus Mochammad Cholily [3*]

[1,2] State Polytechnic of Malang, Malang- East Java, 65141, Indonesia
[3]University of Muhammadiyah Malang, Malang - East Java, 65144, Indonesia

[*] Corresponding author: ymcholily@gmail.com

Submitted: 29/01/2025 | Accepted: 01/03/2025 | Online: 05/03/2025 | doi: xxxx.xxxx.xxxx.xxxx/xxxxxx

Abstract:

Purpose: This research aims to analyze the emotional states of pilots during critical situations before aviation accidents by applying sentiment analysis to pilot communication. It addresses the challenge of identifying emotional cues, stress levels, and urgency in pilot dialogues, crucial for aviation safety. The study evaluates and compares the performance of transformer models—BERT, ALBERT, and RoBERTa—in analysing these emotional factors, contributing to enhanced situational awareness and safety.

Methods: The primary data used was the Cockpit Voice Recorder (CVR), which captures real-time pilot communication. BERT, ALBERT, and RoBERTa were employed for sentiment analysis, trained on domain-specific data to detect emotional distress and stress in critical situations. Model performance was evaluated using metrics such as precision, recall, F1 score, and support to assess accuracy in identifying emotional cues and overall effectiveness in classifying different emotional states.

Results: All models performed well with an accuracy of around 80%. While BERT excelled in detecting negative stressed sentiment, it struggled with neutral sentiment. RoBERTa outperformed both BERT and ALBERT by 5-10% in identifying negative stressed conversations, with higher precision and recall.

Conclusions: RoBERTa is the most effective model for sentiment analysis of pilot conversations, particularly in detecting stress and urgency, crucial for aviation safety. It was more stable and better at handling emotional variations. Further improvements in fine-tuning or exploring data augmentation could enhance its accuracy.

Keywords:

Sentiment Analysis, RoBERTa, Natural Languange Processing.

1. Introduction

Communication is one of the most important factors in the world of aviation. Included, technical maintenance, Air Traffic Controller (ATC) and pilot proficiency and ending with safety inspection at the airport terminal (Vaeng, 2012). In critical situations, miscommunication contributing as a common factor (Jones, 2003). Effective communication between the pilot, copilot, and air traffic controller is a crucial factor in

preventing accidents(Zinaida, 2022). However, analyzing communication in critical situations presents significant challenges due to its complexity and the involvement of emotion (Ashforth & Ashforth, 1986). In this context, artificial intelligence (AI)-based technology, particularly natural language processing (NLP), offers a valuable tool for understanding communication patterns(Tamrakar, 2022). Transformer models such as BERT (Bidirectional Encoder Representations from Transformers), Alberta and RoBERTa (Robustly Optimized BERT Approach) have demonstrated outstanding performance in various NLP tasks, including sentiment analysis(Azizah et al., 2023). Sentiment analysis has gained significant popularity in the airline industry. Various studies have examined sentiment analysis in the airline industry by employing different machine learning and deep learning methods (Tikayat Ray et al., 2023). Sentiment analysis in this context aims to identify emotional cues, stress levels, and urgency in pilot dialogues, which can provide valuable insights into human factors contributing to flight incidents

BERT, as a cutting-edge model, employs a bidirectional approach to grasp word context, making it significantly more effective than traditional NLP models, which process text in only one direction. By analysing both preceding and following words in a sentence, BERT provides a more advanced and accurate text representation. RoBERTa, an optimized version of BERT, enhances its efficiency by utilizing a more extensive training dataset and refined pretraining objectives. ALBERT, on the other hand, introduces a parameter-reduction technique that reduces memory consumption while maintaining performance, making it a more efficient alternative for large-scale NLP applications (Alamsyah & Girawan, 2023; Eang & Lee, 2024; Kim & Jeong, 2023). Previous research examined sentiment analysis in movie reviews and tweets using the Sentiment140 and Coronavirus Tweets NLP datasets, employing BERT, DistilBERT, and RoBERTa models. The findings indicated that BERT achieved the highest accuracy, with test accuracies of 92.76% for Sentiment140 and 90.43% for Coronavirus Tweets NLP, suggesting that pre-training on domain-specific data can further enhance model performance (Narayanaswamy, 2021). Meanwhile, RoBERTa exhibited strong performance, benefiting from smaller batch sizes, an increased number of epochs, and optimal learning rates (1e-5 to 3e-5), all of which contributed to improved accuracy. However, optimizing sequence length remains challenging, and overfitting tends to occur beyond five epochs, highlighting the need for early stopping or regularization techniques to maintain model effectiveness(Sy et al., 2024). In a related study, BERT, RoBERTa, ALBERT, and DistilBERT were compared in Question Answering (QA) tasks using the SQuAD v2 dataset(Özkurt, 2024) The results demonstrated that ALBERT outperformed the other models, primarily due to its efficient learning process requiring fewer epochs. However, it was observed that other models could achieve similar or even superior accuracy if trained for a longer duration, emphasizing the impact of training strategies on model effectiveness. These findings collectively underscore the importance of model selection and optimization techniques in improving performance across different NLP tasks.

Given the complexity of communication in aviation, comparing the performance of BERT, ALBERT, and RoBERTa in sentiment analysis of critical pilot conversations before aviation accidents is highly relevant. Sentiment analysis in this context goes beyond classifying text as positive, negative, or neutral (Aftab et al., 2023). In high-stress

situations, pilots may exhibit emotional cues such as stress, confusion, urgency, or panic, which can significantly impact how information is conveyed (NASA, 2015). These emotional factors not only influence communication but also play a crucial role in pilot decision-making, particularly in high-pressure scenarios such as landing. When stress levels escalate, pilots may struggle to process information effectively. AI-based sentiment analysis can be a valuable tool in detecting pilot stress levels and identifying linguistic patterns associated with emotional distress(Causse et al., 2013). The primary data source for sentiment analysis in aviation safety studies is the Cockpit Voice Recorder (CVR), which records real time pilot conversations during flight operations, including moments leading up to an incident or accident (Noort et al., 2021). CVR data provides critical insight into communication breakdowns, stress levels and emotional responses, making it an essential resources for understanding human factor in aviation (Kayten, 2017). By leveraging models such as BERT, ALBERT, and RoBERTa, sentiment analysis can provide insights into the emotional states of pilots in critical situations, allowing for proactive measures to enhance situational awareness and decision-making. Therefore, analysing emotions in pilot communication is a crucial step in enhancing aviation safety.

2. Methods

2.1 Datasets

For training and evaluation in sentiment analysis, the dataset used was sourced from two main platforms, namely NTSB (National Transportation Safety Board) and Kaggle.com. The data from NTSB was obtained through web scraping, allowing the collection of transcript communications from aviation accidents dating from 1962 to the most recent available data. These transcripts include communications between pilots and air traffic controllers, as well as other relevant details related to aviation accidents.

The dataset from Kaggle.com contains various flight communication data involving critical interactions between pilots and air traffic controllers. This dataset is used to provide variation and a broader scope for analysing pilot communication in critical situations.

index	Airline	last_words	label
0	Alitalia	is leaving now five thousand three six zero on the outer marker	positive
1	Alitalia	say again your last message	neutral
2	Alitalia	Say again please	negative stressed
3	Alitalia	unable to make out your last message will you please repeat	negative stressed
4	Alitalia	please say again	negative stressed
5	Alitalia	Are you coming straight in from the outer marker for landing runway two seven or making a three sixty over the outer marker then reporting leaving outer marker inbound over?	negative stressed
6	Alitalia	OK clear to the outer marker runway two seven make a three sixty on the outer marker then report the outer marker inbound for runway two seven	positive
7	Alitalia	Roger understand you will be making a three sixty over the outer marker Report leaving outer marker while proceeding making a three sixtyneutral	negative stressed
8	Alitalia	Roger will do Alitalia seven seven one	neutral

Figure 1. Sample Datasets

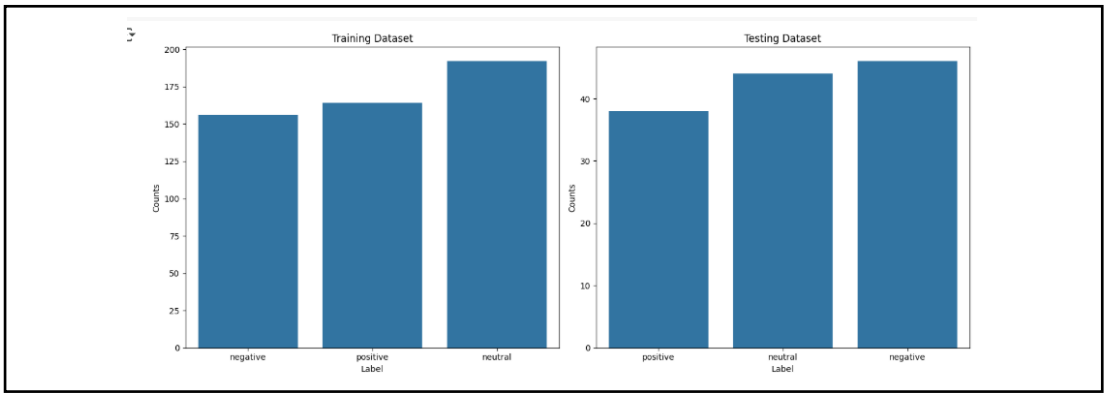

Figure 2. Datasets after Undersampling methods

Both datasets consist of pairs of questions, answers, and command sentences related to communication within the aviation context. The datasets from NTSB and Kaggle.com each contain hundreds of thousands of entries, consisting of in-depth transcript texts and questions associated with the communication. Each data item in the dataset contains communication text, along with related questions aimed at testing the system's understanding of sentiment and context in these communications.

To address class imbalance and ensure a more balanced representation of the sentiment labels, the datasets used in this study have been pre-processed using undersampling techniques. The purpose of undersampling is to reduce the dominance of the majority sentiment class (e.g., negative sentiment) by decreasing its frequency, thus balancing the number of samples for each sentiment category. This technique helps improve the model's ability to generalize across all sentiment categories, ensuring a more accurate and fair sentiment analysis, especially for the minority classes (e.g., positive and neutral sentiments).

2.2 Labelling Process

Traditionally, sentiment analysis has been conducted by manual annotators using a predefined codebook. An alternative approach is to utilize crowd-sourcing platforms instead of relying on traditional expert or undergraduate coders. Another possibility is to employ automated sentiment analysis techniques, where a computer program classifies each document as positive, neutral, or negative(van Atteveldt et al., 2021). The research demonstrates that ChatGPT outperforms MTurk, developed by Amazon, in various tasks. Through a detailed comparison of performance across multiple datasets, it was found that ChatGPT consistently provided more accurate annotations and exhibited higher levels of consistency in comparison to MTurk. This indicates that ChatGPT may serve as a more effective and efficient alternative for text annotation tasks, offering significant improvements in both time and cost reduction(Gilardi et al., 2023).

Sentiment labeling for the dataset was conducted bed on three predefined categories : positive, neutral and negatives. Prior to the labeling process, the dataset underwent a series of pre-processing steps including the conversion of all text to lowercase and removal symbols or special characters.

Table 1. Keyword Datasets

Label	Keywords
Positive	"Happy", "Good", "Well", "Excellent", "Glad", "Satisfied", "Nice"
Negative	"Panic","Afraid","Emergency","Pressure","God","Want","Jesus", "Error","Frustrated","Dissapointed", "Lost", "Allahuakbar", "Down", "Engine", "Failure", "Flap".
Neutral	"Please", "Okay", "Say Again", "Confirm", "Copy", "Clear", "Messages", "Check"

The process begins with a raw dataset containing text data that requires annotation. The first step in the process is text preprocessing, which involves standardizing the text by converting it to lowercase, cleaning unnecessary symbols, correcting slang, and removing stopwords. This step ensures that the dataset is consistent and free from elements that could hinder analysis. Following preprocessing, the text undergoes **lemmatization**, which reduces words to their base forms, thereby improving the dataset's quality for further analysis. The processed dataset is then saved and exported in CSV format. Subsequently, this CSV file is uploaded to ChatGPT for annotation. Criteria instructions are provided to ChatGPT, specifying how the text should be labelled or categorized, such as through sentiment analysis or other relevant classifications.

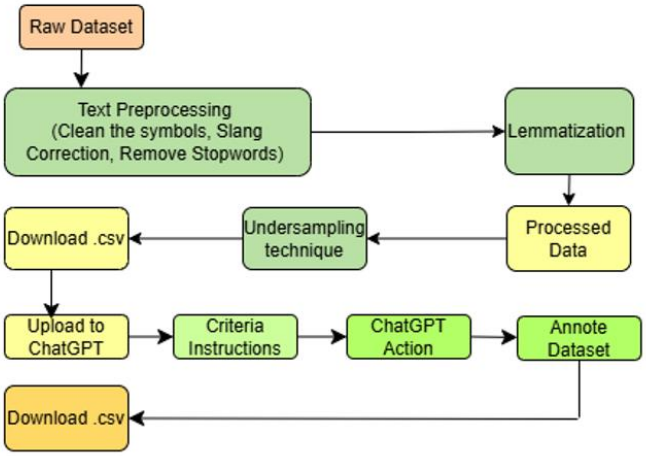

Figure 3. Datasets Processing Flow

In parallel, an undersampling technique is applied to address any class imbalance, ensuring that no particular class is overrepresented and mitigating potential bias in subsequent model training. Once the annotation process is completed, the fully annotated dataset is available for download as a CSV file. The final annotated dataset is now prepared for further analysis and can be used in model training or other research applications

2.3 Method

2.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that is frequently used in various natural language processing (NLP) applications (Oliaee et al., 2023) involving several processes such as preprocessing, embedding, pooling, and classification layers. The next stages include fine-tuning to unfreeze layers for better adaptation, training and validation, followed by prediction(Wu et al., 2024).

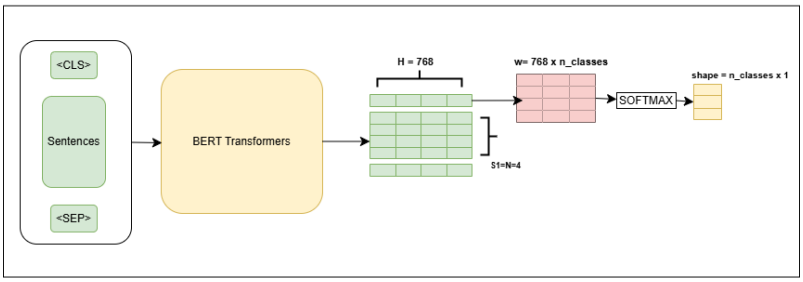

Figure 4. BERT Methods

The process begins by inputting text, which is tokenized with special tokens at the beginning, and at the end. The token serves as a marker for classification purposes, while is used to separate sentences in more complex cases, such as when the text consists of multiple sentences. The tokenized text is then processed by the BERT Transformer, which generates a representation for each token in the text. Each word or token has a representation influenced by the surrounding sentence to better understand the meaning of the word.

The output of BERT is a vector of dimension 768, which represents the input text in a high-dimensional feature space. This representation is then multiplied by a Weight Matrix (W) with a size of 768 x n_classes, where 768 represents the dimension length from BERT, and n_classes refers to the number of classes in the classification task. In this case, there are 3 tasks and 3 categories. The result of multiplying the BERT representation and the Weight Matrix is then passed to the Softmax function, which transforms it into probabilities for each class, allowing the model to predict the most relevant class for the given text. The final output is the classification probability for each class with dimensions of n_classes x 1, indicating the model's prediction for the class most suited to the input text(Sayeed et al., 2023).

2.3.2 AlBERTa

ALBERT is a lighter version of BERT, designed to reduce the number of parameters and improve training efficiency. To achieve this, ALBERT implements two primary parameter reduction techniques. The first, Factorized Embedding Parameterization, decomposes the large embedding matrix into two smaller matrices, allowing for an increase in the hidden size without significantly enlarging the embedding matrix. The second, Cross-layer Parameter Sharing, shares parameters across layers, preventing the parameters from increasing with the depth of the network. Additionally, ALBERT

adopts a self-supervised learning approach through Sentence-Order Prediction (SOP). SOP focuses on modeling inter-sentence coherence by predicting the correct order of two consecutive text segments. This approach proves more effective than Next-Sentence Prediction (NSP) used in BERT, as SOP avoids topic prediction and instead focuses on discourse coherence between sentences. By employing these techniques, ALBERT significantly reduces the number of parameters while maintaining or even improving performance on various natural language understanding tasks, as evidenced by its results on the GLUE and SQuAD benchmarks (Lan et al., 2020).

2.3.3 RoBERTa

RoBERTa is employed for various Natural Language Processing (NLP) tasks, such as text classification and sentiment analysis. The effectiveness of RoBERTa depends on its ability to extract textual information and establish semantic relationships within text. To enhance its capability to capture significant patterns from large datasets, this model incorporates a Convolutional Neural Network (CNN) layer. In evaluating input sequences and constructing contextualized representations of words in a phrase, RoBERTa—similar to BERT—is a transformer-based language model that utilizes self-attention mechanisms.

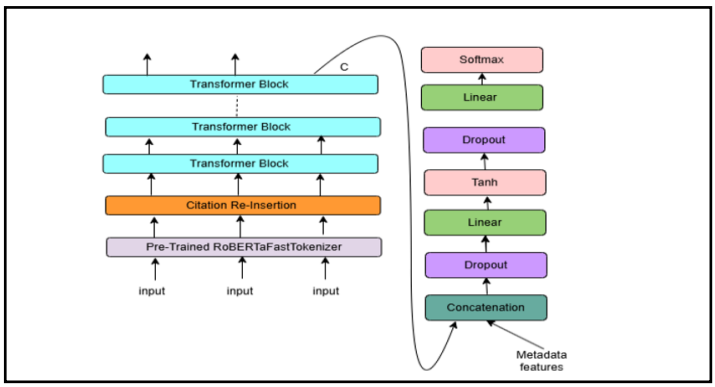

Figure 5. RoBERTa Methods

RoBERTa is trained on a substantially larger dataset and utilizes a more efficient training approach compared to BERT. While both models share similar architectural designs, RoBERTa employs a byte-level BPE tokenizer, akin to GPT-2, and a distinct pretraining strategy. Unlike BERT, which is trained on a 16GB dataset, RoBERTa is trained on nearly 160GB of uncompressed text. Additionally, RoBERTa benefits from training with: (i) full sentences without Next Sentence Prediction (NSP) loss, (ii) dynamic masking, (iii) large mini-batches, and (iv) a larger byte-level BPE tokenizer (Singla, 2024).

2.4 Training Process

The first step involves uploading the data from a .csv file and reading it using the pandas library. After the data is successfully loaded, the next step is to separate the features and labels. The data is then split into two sets: the training set and the test set,

utilizing the train_test_split function from the scikit-learn library, with an 80% allocation for training and 20% for testing.

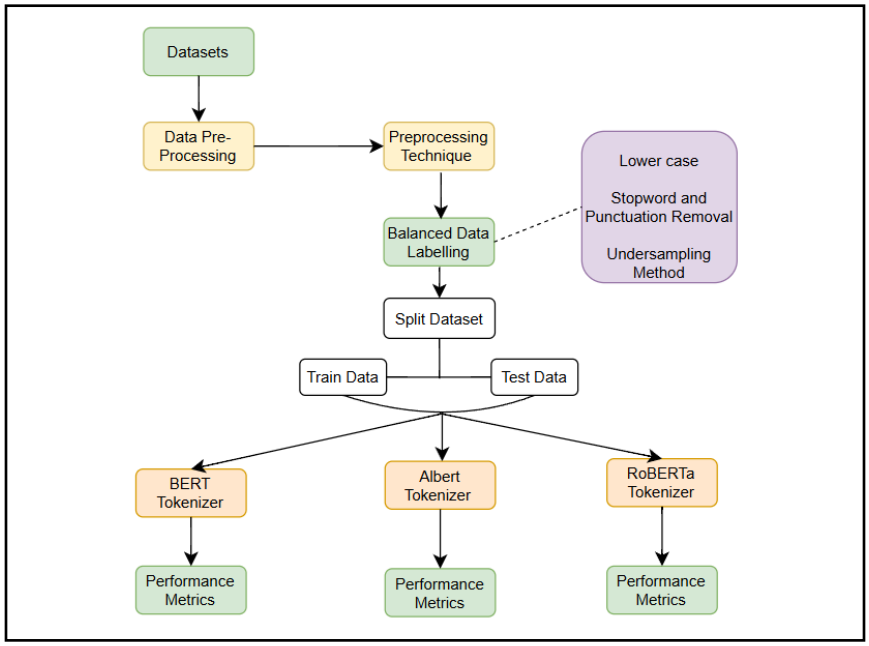

Figure 6. Training Process

The modeling process for BERT, RoBERTa, and ALBERT utilizes the respective tokenizers from the Hugging Face Transformers library. Before the tokenization process, the labels in the dataset are transformed into numerical form using LabelEncoder. Additionally, during the pre-processing stage, several techniques are applied, such as converting text to lowercase, removing stopwords, and eliminating punctuation, which are then processed into Balanced Data Labelling. The data is also balanced using the undersampling technique to address the issue of class imbalance in the dataset. The undersampling process reduces the number of samples from the majority class, making the data more balanced and improving the model's ability to handle the less-represented classes.Formodeling,BERTForSequenceClassification,RoBERTaForSequenceClassification, and ALBERTForSequenceClassification are employed, with a total of three epochs, batch size for training and evaluation, and model checkpoint saving every 500 training steps. Classification reports are then generated using the classification_report from scikit-learn to provide evaluation metrics such as precision, recall, and F1-score.

3. Results and Discussion

The BERT, RoBERTa, and ALBERT models were tested to analyze a dataset of pilot conversations before accidents, aiming to identify the sentiment contained in these conversations. Based on the classification report and confusion matrix results, all three models demonstrated an overall accuracy of around 80%. However, there were variations in precision and recall for each class.

Laporan Klasifikasi:

	precision	recall	f1-score	support
negative stressed	0.83	0.92	0.87	176
neutral	0.65	0.55	0.60	47
positive	0.75	0.47	0.58	32
accuracy			0.80	255
macro avg	0.74	0.65	0.68	255
weighted avg	0.79	0.80	0.79	255

[16/16 00:33]

Hasil evaluasi:

{'eval_loss': 0.5278079509735107, 'eval_runtime': 36.0758, 'eval_samples_per_second': 7.068, 'eval_steps_per_second': 0.444, 'epoch': 3.0}

Figure 7. BERT Classification Report

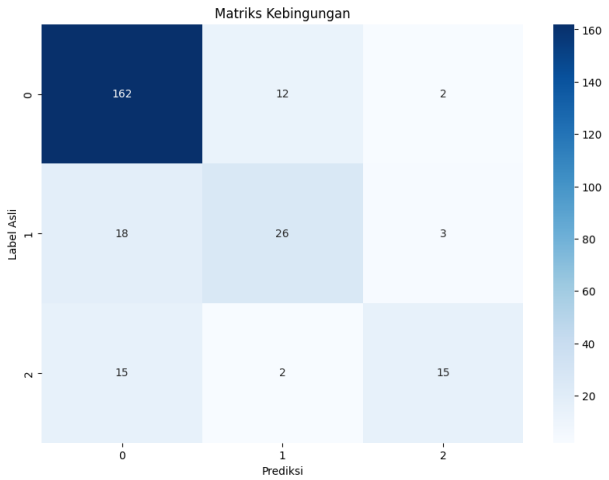

Figure 8. BERT Confusion Matrix

The BERT model achieved an accuracy of 80%, with its best performance in classifying negative stressed, where precision reached 0.80 and recall 0.96, indicating that BERT is highly effective in detecting conversations that suggest stress. However, for the neutral class, BERT showed lower performance, with a precision of 0.76, recall of 0.34, and an F1-Score of 0.47, highlighting the model's difficulty in identifying neutral conversations. Meanwhile, for the positive class, BERT achieved a precision of 0.79, recall of 0.59, and an F1-Score of 0.68, which is fairly good, though there is still room for improvement.

```
Laporan Klasifikasi:
                  precision    recall  f1-score   support

negative stressed       0.80      0.96      0.88       176
          neutral       0.76      0.34      0.47        47
         positive       0.79      0.59      0.68        32

         accuracy                           0.80       255
        macro avg       0.79      0.63      0.67       255
     weighted avg       0.80      0.80      0.78       255

[16/16 00:48]
Hasil evaluasi:
{'eval_loss': 0.727790117263794, 'eval_runtime': 52.3931, 'eval_samples_per_second': 4.867, 'eval_steps_per_second': 0.305, 'epoch': 3.0}
```

Figure 9. AlBERT Classification Report

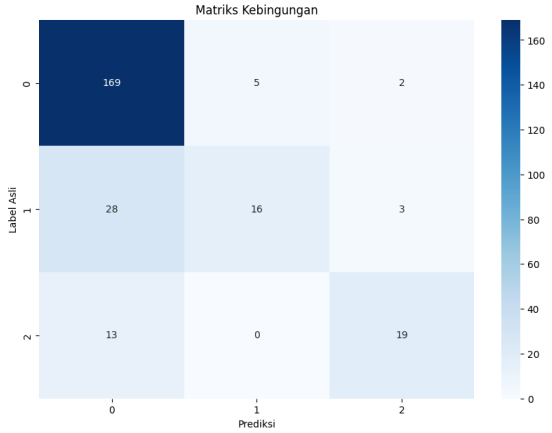

Figure 10. AlBERT Confusion Matrix

ALBERT also demonstrated an equivalent accuracy of 80%. For the negative stressed class, ALBERT achieved a precision of 0.80, recall of 0.96, and an F1-Score of 0.88, which is almost similar to BERT and slightly superior. However, for the neutral class, ALBERT showed weak performance with a precision of 0.76, recall of 0.34, and an F1-Score of 0.47, similar to BERT's results. In the positive class, ALBERT obtained a precision of 0.79, recall of 0.59, and an F1-Score of 0.68, which is nearly identical to BERT.

```
Laporan Klasifikasi:
                  precision    recall  f1-score   support

negative stressed       0.89      0.83      0.86       176
          neutral       0.60      0.64      0.62        47
         positive       0.61      0.78      0.68        32

         accuracy                           0.79       255
        macro avg       0.70      0.75      0.72       255
     weighted avg       0.80      0.79      0.79       255

[16/16 00:45]
Hasil evaluasi:
{'eval_loss': 0.6324674487113953, 'eval_runtime': 47.7889, 'eval_samples_per_second': 5.336, 'eval_steps_per_second': 0.335, 'epoch': 3.0}
```

Figure 11. RoBERTa Classification Report

The RoBERTa model performed well in sentiment analysis of pilot conversations, particularly in detecting negative stressed emotions with a precision of 0.89, recall of 0.83, and an F1-score of 0.86. It struggled more with neutral sentiment, achieving precision of 0.60, recall of 0.64, and F1-score of 0.62, indicating challenges in identifying neutral emotions.

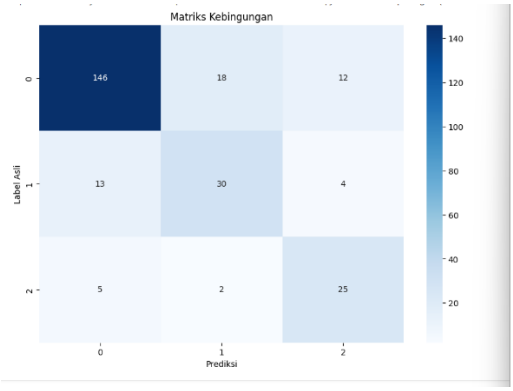

Figure 12. RoBERTa Confusion Matrix

For positive sentiment, RoBERTa performed reasonably with precision of 0.61, recall of 0.78, and F1-score of 0.68. Overall, the model achieved an accuracy of 0.79, with a macro average precision of 0.70, recall of 0.75, and F1-score of 0.72, while the weighted average values were 0.80 for precision, 0.79 for recall, and 0.79 for F1-score. These results show RoBERTa's strength in detecting stress but also point to areas for improvement in classifying neutral and positive sentiments.

All three models demonstrated good overall accuracy, but the main challenge lies in classifying the neutral class, which showed lower precision and recall across all three models. This indicates difficulty in distinguishing neutral conversations from more distinct sentiments such as positive or negative stressed. The small differences between these models suggest that transformer-based models like BERT, RoBERTa, and ALBERT excel in handling more emotional sentiments but are less effective in dealing with neutral sentiments.

4. Conclusion

Based on the classification report and confusion matrix results for the BERT, RoBERTa, and ALBERT models, all three demonstrated fairly good performance in sentiment analysis of pilot conversations before aircraft accidents, with an overall accuracy of approximately 80%. The BERT model showed solid performance, particularly in detecting negative stressed sentiment, with high precision and recall. However, it struggled to identify neutral sentiment, as reflected in its lower precision and recall for this category. Meanwhile, the RoBERTa model performed slightly better, especially in detecting negative stressed, with higher precision and recall compared to BERT. Nevertheless, RoBERTa's performance in classifying neutral sentiment remained similar to BERT. The ALBERT model produced results consistent with BERT, particularly in identifying negative stressed conversations, while its precision and recall for the positive class were nearly equivalent to the other two models.

The main challenge lies in classifying neutral sentiment. This class exhibited lower precision and recall across all three models, indicating that neutral sentiment is difficult to distinguish clearly from positive or negative sentiment. This may be due to the characteristics of neutral sentiment, which lack clear emotional cues, making it harder to predict accurately.

Among the three models, RoBERTa demonstrated the best performance compared to BERT and ALBERT. RoBERTa outperformed BERT and ALBERT by approximately 5-10% in identifying negative stressed conversations and demonstrated more stable performance across other classes, despite facing the same challenge in classifying neutral sentiment. RoBERTa successfully identified negative stressed more accurately, and although there were no significant differences in classifying neutral or positive sentiment, overall, RoBERTa was superior in handling variations in sentiment, particularly emotional ones. RoBERTa excelled in identifying critical sentiment with higher precision and recall in the negative stressed class, which is highly relevant in the context of aviation safety. Overall, RoBERTa is the most suitable model for sentiment analysis of pilot conversations, given its advantage in detecting conversations that indicate stress or critical conditions. With further improvements in handling neutral sentiment, this model could provide even more optimal results in analyzing critical conversations in aviation.

Refererences

Aftab, F., Bazai, S. U., Marjan, S., Baloch, L., Aslam, S., Amphawan, A., & Neo, T. K. (2023). A Comprehensive Survey on Sentiment Analysis Techniques. *International Journal of Technology*, *14*(6), 1288–1298. https://doi.org/10.14716/ijtech.v14i6.6632

Alamsyah, A., & Girawan, N. D. (2023). Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model. *Big Data and Cognitive Computing*, *7*(4). https://doi.org/10.3390/bdcc7040168

Ashforth, .B, & Ashforth, .B. (1986). from the SAGE Social Science Collections . Rights Reserved . *The ANNALS of the American Academy of Political and Social Science*, *503*(1), 122–136.

Azizah, S. F. N., Cahyono, H. D., Sihwi, S. W., & Widiarto, W. (2023). Performance Analysis of Transformer Based Models (BERT, ALBERT, and RoBERTa) in Fake News Detection. *2023 6th International Conference on Information and Communications Technology, ICOIACT 2023*, *November*, 425–430. https://doi.org/10.1109/ICOIACT59844.2023.10455849

Causse, M., Dehais, F., Péran, P., Sabatini, U., & Pastor, J. (2013). The effects of emotion on pilot decision-making: A neuroergonomic approach to aviation safety. *Transportation Research Part C: Emerging Technologies*, *33*(August), 272–281. https://doi.org/10.1016/j.trc.2012.04.005

Eang, C., & Lee, S. (2024). Improving the Accuracy and Effectiveness of Text Classification Based on the Integration of the Bert Model and a Recurrent Neural Network (RNN_Bert_Based). *Applied Sciences (Switzerland)*, *14*(18). https://doi.org/10.3390/app14188388

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30), 1–3. https://doi.org/10.1073/pnas.2305016120

Jones, R. K. (2003). Miscommunication between pilots and air traffic control. *Language Problems and Language Planning*, *27*(3), 233–248. https://doi.org/10.1075/lplp.27.3.03jon

Kayten, P. (2017). *The Application of CVR and FDR Data In Human Performance Investigations*. *September 1985*.

Kim, K. H., & Jeong, C. S. (2023). F-ALBERT: A Distilled Model from a Two-Time Distillation System for Reduced Computational Complexity in ALBERT Model. *Applied Sciences (Switzerland)*, *13*(17). https://doi.org/10.3390/app13179530

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: a Lite Bert for Self-Supervised Learning of Language Representations. *8th International Conference on Learning Representations, ICLR 2020*, *October*.

Narayanaswamy, G. R. (2021). Exploiting BERT and RoBERTa to Improve Performance for Aspect Based Sentiment Analysis. *Dissertations*. https://doi.org/10.21427/3w9n-we77

NASA. (2015). Effects of Acute Stress on Aircrew Performance: Literature Review and Analysis of Operational Aspects. *NASA Technical Memorandum 2015-218930*, *August*, 1–30.

Noort, M. C., Reader, T. W., & Gillespie, A. (2021). Cockpit voice recorder transcript data: Capturing safety voice and safety listening during historic aviation accidents. *Data in Brief*, *39*, 107602. https://doi.org/10.1016/j.dib.2021.107602

Oliaee, A. H., Das, S., Liu, J., & Rahman, M. A. (2023). Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types. *Natural Language Processing Journal*, *3*(April), 100007. https://doi.org/10.1016/j.nlp.2023.100007

Özkurt, C. (2024). Comparative Analysis of State-of-the-Art Q A Models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 Dataset. *Chaos and Fractals*, 0–22. https://doi.org/10.69882/adba.chf.2024073

Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). BERT: A Review of Applications in Sentiment Analysis. *HighTech and Innovation Journal*, *4*(2), 453–462. https://doi.org/10.28991/HIJ-2023-04-02-015

Singla, A. (2024). Roberta and BERT: Revolutionizing Mental Healthcare through Natural Language. *Shodh Sagar Journal of Artificial Intelligence and Machine Learning*, *1*(1), 10–27. https://doi.org/10.36676/ssjaiml.v1.i1.02

Sy, C. Y., Maceda, L. L., Canon, M. J. P., & Flores, N. M. (2024). Beyond BERT: Exploring the Efficacy of RoBERTa and ALBERT in Supervised Multiclass Text Classification. *International Journal of Advanced Computer Science and Applications*, *15*(3), 223–233. https://doi.org/10.14569/IJACSA.2024.0150323

Tamrakar, A. K. (2022). Natural Language Processing in Artificial Intelligence, NLPinAI 2021. *Studies in Computational Intelligence*, *999 SCI*(April).

Tikayat Ray, A., Bhat, A. P., White, R. T., Nguyen, V. M., Pinon Fischer, O. J., & Mavris, D. N. (2023). Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS). *Aerospace*, *10*(9). https://doi.org/10.3390/aerospace10090770

Vaeng, K. A. (2012). *School of Hotel Management Master ' S Thesis*. 1–127.

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Wu, Y., Jin, Z., Shi, C., Liang, P., & Zhan, T. (2024). Research on the application of deep learning-based BERT model in sentiment analysis. *Applied and Computational Engineering*, *71*(1), 14–20. https://doi.org/10.54254/2755-2721/71/2024ma

Zinaida, R. S. (2022). *Optimizing Air Traffic Controller Communication for Enhanced Flight Safety : A Case Study of AirNav Indonesia Palembang*. *03*(01), 1307–1314. https://doi.org/10.12928/sylection.v3i1.14517