ASPECT-BASED SENTIMENT ANALYSIS ON CHATGPT IN TWITTER USING BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

[1]HANDRIZAL, [2]ANANDHINI MEDIANTY NABABAN, [3]RICKY ALAN

[1,2,3]Department of Computer Science, Faculty of Computer Science and Information Technology,
Universitas Sumatera Utara, Jl. University No. 9-A, Medan 20155, Indonesia

E-mail: handrizal@usu.ac.id

ABSTRACT

ChatGPT, a chatbot developed by OpenAI, is currently being widely discussed, especially on social media platforms like Twitter. Many users have provided positive feedback regarding this chatbot. However, some have provided negative feedback. To understand the sentiment of users regarding ChatGPT, this research conducted aspect-based sentiment analysis using the Bidirectional Encoder Representations from Transformers (BERT) model. The aspects analyzed include general aspects, functions, performance, and potential of ChatGPT. The data used in this research was obtained from the social media platform Twitter. The BERT model used is IndoLEM/IndoBERT-base-uncased, which has been pre-trained on 220 million Indonesian language words from various sources. From the conducted tests, the model was able to achieve an f1-score of 84% for the general aspect, 89% for the functional aspect, 98% for the performance aspect, and 98% for the potential aspect.

Keywords: *Sentiment Analysis, Aspect Based Sentiment Analysis, ChatGPT, BERT, Twitter.*

1. INTRODUCTION

In today's digital era, social media is one of the most popular and widely used sources of information [1]. We can discuss and exchange opinions with others through social media [2]. One of the most popular social media today is Twitter. Twitter allows users to share posts or tweets with a maximum length of 140 characters. Tweets usually contain people's opinions about certain topics [3].

ChatGPT, a chatbot developed by OpenAI, has become a hot topic of conversation on Twitter. Many users gave a positive view of ChatGPT, due to its ability to provide answers and solutions to the questions. However, some expressed concerns about the possibility of ChatGPT replacing humans [4]. Research [5] showed that tweets about ChatGPT can generally be grouped into 3 categories, general, functional, and potential of ChatGPT. The study also showed that some users also discussed the ability of ChatGPT to provide answers.

Understanding user sentiment about ChatGPT can provide information about this technology's potential advantages and disadvantages [4].

Sentiment analysis is a field of Natural Language Processing (NLP) that is widely used to extract views or opinions from textual data [6]. Sentiment analysis automatically extracts and identifies information from the text to determine whether the text contains positive, negative, or neutral sentiments. Sentiment analysis has been widely applied in various fields, such as business, politics to health [2]. However, sentiment analysis generally only focuses on identifying sentiment in the text without knowing the aspects of sentiment [7]. Texts can contain sentiments about different aspects. Therefore, analyzing sentiment based on aspects is necessary to gain a more detailed understanding of opinions or views on a product or service [8].

BERT (Bidirectional Encoder Representations from Transformers) is a Transformer-based model that only uses the encoder layer. As the name implies, BERT is a bidirectional model, meaning that BERT can process text from both directions, from left to right and right to left [9]. BERT uses attention mechanisms to better understand the relationships between each word [10]. BERT can also handle ambiguity in understanding language and can achieve near-human performance [11].

Based on the description above, this research was conducted "Aspect Based Sentiment Analysis on ChatGPT in Twitter Using Bidirectional Encoder Representations from Transformers (BERT)". In this research, new knowledge is created in the form of a model that can classify the sentiments of Twitter users towards ChatGPT.

2. FORMULATION OF THE PROBLEM

In this research, the formulation of the problem discussed is how to perform aspect-based sentiment analysis on ChatGPT-related tweets on Twitter by implementing Bidirectional Encoder Representations from Transformers (BERT)

3. RESEARCH OBJECTIVE

The main objective of this research is to identify the aspects and sentiments contained in tweets that discuss ChatGPT on Twitter by implementing BERT.

4. SCOPE AND LIMITATION

The scope and limitations of this research are as follows:

- Data used from social media Twitter.
- The data obtained amounted to 7844 tweets.
- The analysis was performed using the bidirectional encoder representations from the transformers (BERT) model.
- The analysis is only done on the tweet data about ChatGPT.
- Aspects analyzed are general aspects, functions, performance, and potential.

The limitations of this research are needed to focus on one problem, help identify problems to be discussed, limit the range of processes discussed, provide an overview of the things to be researched, and tested so that this research is more effective, efficient, and directed in finding problem-solving.

5. LITERATURE REVIEW

Aspect-based sentiment analysis research conducted by Hoang, Bihorac, and Rouces titled "Aspect-Based Sentiment Analysis Using BERT" uses a combination model to classify aspects and sentiments at once. The model obtained an accuracy value of 87.5% for the restaurant dataset, 78.7% for the laptop dataset, and 87.3% for the hotel dataset [7]. Research conducted by Sun, Huang, and Qiu titled "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence" obtained an F1-Score value of 92.18% [8]. Another research conducted by Abdelgwad titled "Arabic Aspect Based Sentiment Classification Using BERT" was conducted on 3 datasets, namely Arabic Hotel Reviews, Arabic News, and HAAD (Human Annotated Arabic Dataset). The research obtained an accuracy value of 89.51% on the Hotel Reviews dataset, 85.73% on the Arabic News dataset, and 73.23% on the HAAD dataset [11]. The above research shows that BERT can be used for aspect-based sentiment analysis and has excellent results.

5.1. Deep Learning

Deep learning is a part of machine learning that is inspired by the human brain system, namely neurons. As in the brain, each neuron will receive information and pass the information to other neurons. Generally, neurons in deep learning are organized into different layers. The input layer will receive input (can be an image or text) and the output layer will produce output (can be a classification result). Between the input and output layers, there is a hidden layer that is tasked to map the input to the corresponding output [12].

Deep learning is generally used to learn high-dimensional data such as images, sound, and text because of its excellent performance, outperforming machine learning algorithms [12].

5.1.1. Transformers

Transformer is a deep learning model architecture that uses the attention mechanism to determine the relationship between each element in the data. Attention allows modeling relationships without regard to the distance of each data element. Each element will be assigned a weight or value based on its relationship with other elements in the data sequence [13].

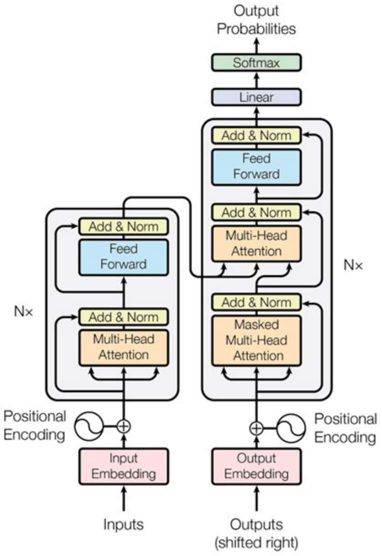


Figure 1: Transformers Architecture

Figure 1 above shows the Transformers architecture which consists of two main components, namely the encoder and decoder. The encoder is responsible for converting the input into a contextual representation through a self-attention block and a feed-forward neural network. The self-attention block allows the encoder to calculate the weight of the relationship between tokens in the input. The feed-forward neural network block will then process the representation generated from self-attention. The encoder part receives the representation generated by the encoder and uses it to generate the corresponding output. The self-attention block in the decoder allows the decoder to learn the relationship between the input and output. The resulting representation will be processed by the feed-forward neural network to produce the output [13].

Transformers have an advantage over architectures such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) that require longer training time [13].

5.1.2. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a developmental model of Transformers that only uses an encoder layer. BERT is a bidirectional model, which means that BERT can process text from both directions, either from left to right or from right to left [9].

Figure 2 below shows the same BERT architecture as its original implementation, Transformers. However, BERT only adopts the encoder part.

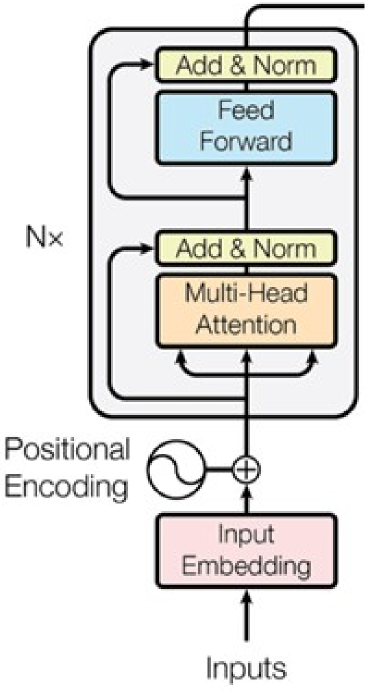


Figure 2: BERT Architecture

BERT has been pre-trained on two different tasks, namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) which can then be fine-tuned on various other tasks [9].

1. Masked Language Modeling (MLM)

   In this task, 15% of tokens will be randomly selected from the dataset. 80% of the selected tokens will be replaced with [MASK] tokens, 10% will be replaced with random tokens, and if the remaining 10% of tokens are not replaced, then BERT will try to predict the replaced tokens.
2. Next Sentence Prediction (NSP)

   In this task, BERT is trained to understand the relationship between two sentences A and B. In 50% of the dataset, B is the next sentence of A, which is labeled next. While in the other 50% of the dataset, B is not the next sentence of A, which is labeled NotNext.

Pre-trained BERT can be fine-tuned on various tasks by adding only one output layer. Unlike pre-training, fine-tuning BERT is relatively faster and does not require large computer resources [9].

5.2. Sentiment Analysis

Sentiment analysis automatically extracts and identifies information from the text to determine whether the text contains positive, negative, or neutral sentiment [2]. Sentiment analysis allows us to know the public's views on a particular topic. The results of sentiment analysis can be taken into consideration for decision-making [14]. Sentiment analysis has been widely applied in various fields, such as business, politics, and health [2].

3 levels in sentiment analysis indicate the level of analysis performed [15].

- Document Level: At this level, sentiment analysis is performed on the entire document or text, The purpose of this analysis is to determine the general sentiment of the document, whether it is positive, negative, or neutral.
- Sentence Level: At this level, sentiment analysis is performed on each sentence in the document. The focus is to identify the sentiment contained in each sentence, whether positive, negative, or neutral.
- Aspect Level: At this level, sentiment analysis is performed on specific aspects or features of an entity.
-

5.2.1 Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis focuses on identifying the aspects present in the text, along with the sentiment expressed towards those aspects. Aspect-based sentiment analysis was first introduced at SemEval-2014 to analyze restaurant and laptop review datasets [7]. Aspect-based sentiment analysis can help better understand the public's opinion or view of a product or service [8].

5.3. Social Media and Twitter

Social media has become one of the most important platforms for interacting and sharing information among internet users [2]. One of the popular social media platforms today is Twitter. Twitter allows users to send and read messages called "tweets" with a certain character limit. Tweets can contain opinions or sentiments on certain topics [3].

5.4. ChatGPT

ChatGPT is a chatbot developed by OpenAI, trained using the GPT (Generative Pre-trained Transformer) architecture to respond to questions, generate text, and interact in human-like conversations. ChatGPT has been trained on a wide variety of text data, allowing it to provide answers in a very broad context. ChatGPT is widely used for various applications and can provide answers in multiple languages [16].

5.5. Regularization

One of the problems that often occur in training machine learning models, especially neural networks is overfitting. Overfitting occurs when the model gives excellent results on training data but fails to predict new data that has never been seen before [17].

Regularization can be used to deal with model overfitting. Commonly used regularization techniques are L1 and L2 regularization. These techniques prevent the model weight from becoming too large by adding a penalty to the loss function. Besides L1 and L2 regularization, another commonly used regularization technique is dropout. This technique will randomly remove neurons during training, thus reducing the complexity of the model [17].

5.6. F1-Score

F1-Score is a commonly used evaluation metric to assess classification performance. F1-Score is the harmonic mean of precision and recall. F1-Score takes into account false positive and false negative values. A good F1-Score has low false positive and false negative values. A score of 1 is the best result of the F1-Score [18]. Figure 3 below shows the formula for calculating F1-Score.

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

Figure 3: F1-Score Formula

6. METHODOLOGY

The system designed in this research is a system to perform aspect-based sentiment analysis on ChatGPT on Twitter. The sentiment analysis system is designed using the Bidirectional Encoder Representations from the Transformers (BERT) model. The BERT model will be fine-tuned with data about ChatGPT obtained from Twitter.

The fine-tuning of the model begins with collecting data about ChatGPT from Twitter. The data is collected using the scrape library from the Python programming language. The data collected is Indonesian text data containing the word 'chatgpt'. The data will then go through a preprocessing stage, including text cleaning, labeling, and tokenization. Next, the clean data will be used to train the Bidirectional Encoder Representations from the Transformers (BERT) model. After the model is trained, it will be evaluated using the f1-score metric to determine how well it performs aspect-based sentiment analysis.

The research methodology used in this study is as follows:

- Data collection
- Data preprocessing (text cleaning, labeling data, data tokenization)
- Training models
- Model evaluation

6.1. Data Collecting

Data will be collected from Twitter using a scrape library from Python. The data collected is data containing the word 'chatgpt'. Only Indonesian data is collected to ensure consistency in analysis. Data will be collected from January 01, 2023, to February 28, 2023.

The data retrieved from Twitter is only date data and user tweets in text form. A total of 13,961 tweet data were obtained at this stage. The data obtained will then be saved into a CSV file. This data collection was carried out on May 21, 2023.

Data collection was carried out using language filters directly from the scrape library, to ensure that only Indonesian language data was obtained. However, some data in other languages are retrieved in the results.

6.2. Data Preprocessing

The data that has been collected needs to be preprocessed first before being used to train the model. This process consists of 3 steps, text cleaning, data labeling, and data tokenization.

6.2.1.Text cleaning

In this step, the tweets will undergo cleaning to remove links, mentions, hashtags, and emoticons that are irrelevant or do not contribute to the analysis. This cleaning is done so that the data becomes clean and only contains the relevant text.

After that, a language filter will be performed on the data to ensure that all data used to train the model will be Indonesian data. The language filter is done using a machine-learning model that is trained separately. The amount of data after text cleaning is 8,278 data.

6.2.2. Data labeling

This step involves labeling each tweet data based on the sentiment contained in the text for each aspect to be analyzed. Tweets will be labeled with positive, negative, neutral, and sentiment for each aspect analyzed. The label none means that the sentiment is not intended for that aspect. 4 aspects will be analyzed in this research:

- General Aspects, discussing general things about ChatGPT, technology, interactions, and user reactions to ChatGPT.
- Functional Aspects, discussing the use of ChatGPT to perform certain tasks.
- Performance Aspects, discussing the ability of ChatGPT to provide answers.
- Potential Aspects, discussing the potential impact that could occur due to ChatGPT.
-

6.2.3. Data tokenization

Tokenization is the process of breaking down text into smaller words or sub-words called tokens. Each token is given an index that corresponds to the one in the dictionary used by the BERT model. After that, the tokenizer will convert each token into an embedding. This process allows BERT to understand the context of the word in the sentence. Data tokenization is done with the help of the Transformers library, BertTokenizerFast.

BertTokenizerFast will convert each data into tokens with a specified maximum length of 64. Data will be padded if the data length does not reach the maximum size. Data that exceeds the maximum length will be truncated. At this stage, special tokens [CLS] and [SEP] are also added. The [CLS] token is added at the beginning of a sentence and the [SEP] token is added at the end of a single sentence.

6.3. Model Training

At this stage, the pre-trained BERT model will be trained using data that has gone through the preprocessing process. The model to be used is the IndoLEM/IndoBERT-base-uncased model which has been pre-trained on 220 million Indonesian words from various sources. This research uses Transformers and Tensorflow libraries to train the model.

The Transformers library is used to download the pre-trained IndoLEM/IndoBERT-base-uncased model through the TFBertModel function. The TensorFlow library adds classification layers for each aspect and runs the training process.

The model trained is a multilabel and multiclass classification model. The model receives input in the form of input_ids or token_ids and attention masks generated from the tokenizer. The model will produce 4 outputs for each aspect (label) analyzed. For each aspect, the model will classify the sentiment (class) of the text, whether it is positive, negative, neutral, or none.

The model will be trained for 5 epochs using categorical_crossentropy loss function and Adam optimizer with learning_rate = 0.00005.

6.4. Evaluation

At this stage, the trained model will be evaluated using the f1-score metric. From this metric, we can measure how well the model performs aspect-based sentiment analysis.

Figure 4 below shows the f1-score values obtained during model training.

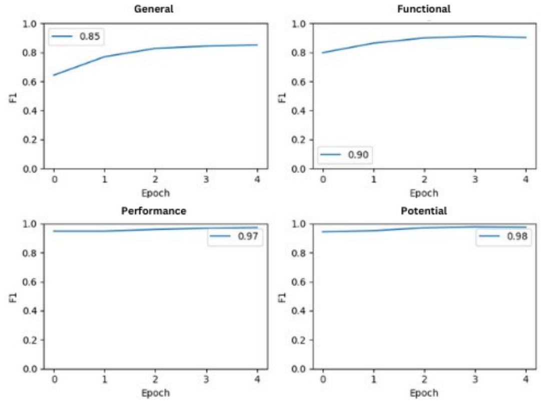


Figure 4: F1-Score for Each Aspect

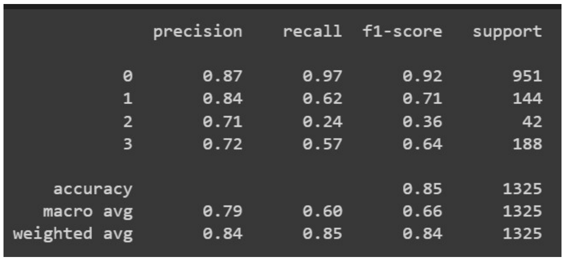
The model was able to achieve an f1-score of 85% for the general aspect, 90% for the functional aspect, 97% for the performance aspect, and 98% for the potential aspect.

These values indicate that the model has excellent performance, especially for the performance and potential aspects.

7. RESULTS

Model testing is performed on new data that the model has never seen before.

Figures 5 - 8 below show the classification report for each aspect analyzed.

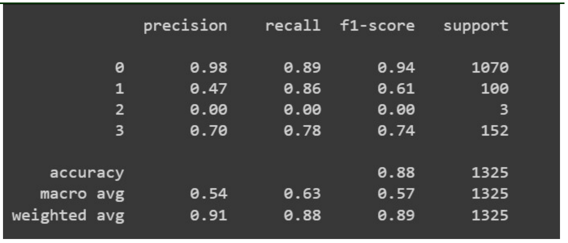


```
              precision    recall  f1-score   support

           0       0.87      0.97      0.92       951
           1       0.84      0.62      0.71       144
           2       0.71      0.24      0.36        42
           3       0.72      0.57      0.64       188

    accuracy                           0.85      1325
   macro avg       0.79      0.60      0.66      1325
weighted avg       0.84      0.85      0.84      1325
```

Figure 5: Classification Report for General Aspect

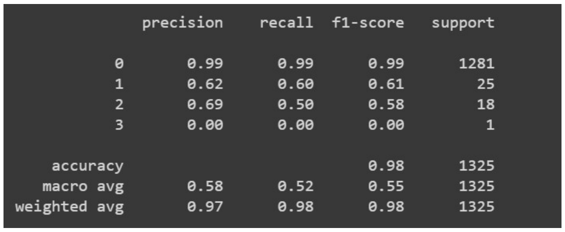


```
              precision    recall  f1-score   support

           0       0.98      0.89      0.94      1070
           1       0.47      0.86      0.61       100
           2       0.00      0.00      0.00         3
           3       0.70      0.78      0.74       152

    accuracy                           0.88      1325
   macro avg       0.54      0.63      0.57      1325
weighted avg       0.91      0.88      0.89      1325
```

Figure 6: Classification Report for Functional Aspect

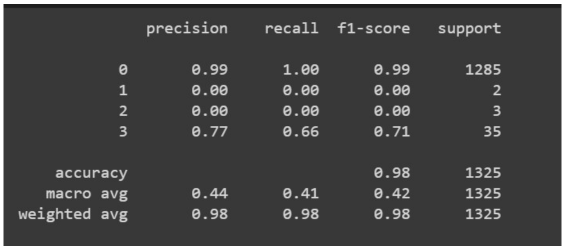


```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      1281
           1       0.62      0.60      0.61        25
           2       0.69      0.50      0.58        18
           3       0.00      0.00      0.00         1

    accuracy                           0.98      1325
   macro avg       0.58      0.52      0.55      1325
weighted avg       0.97      0.98      0.98      1325
```

Figure 7: Classification Report for Performance Aspect



```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1285
           1       0.00      0.00      0.00         2
           2       0.00      0.00      0.00         3
           3       0.77      0.66      0.71        35

    accuracy                           0.98      1325
   macro avg       0.44      0.41      0.42      1325
weighted avg       0.98      0.98      0.98      1325
```

Figure 8: Classification Report for General Aspect

From each classification report created, the value that is considered is the weighted average f1-score value because this value is more suitable for use if there is an imbalance in the amount of data. From the tests conducted, the model was able to obtain f1-score results of 84% for general aspects, 89% for function aspects, 98% for performance aspects, and 98% for potential aspects.

8. DISCUSSIONS

The model was also tested on data about ChatGPT collected from March 01, 2023, to March 31, 2023. Figures 9 - 12 below show the model classification results for each aspect analyzed.

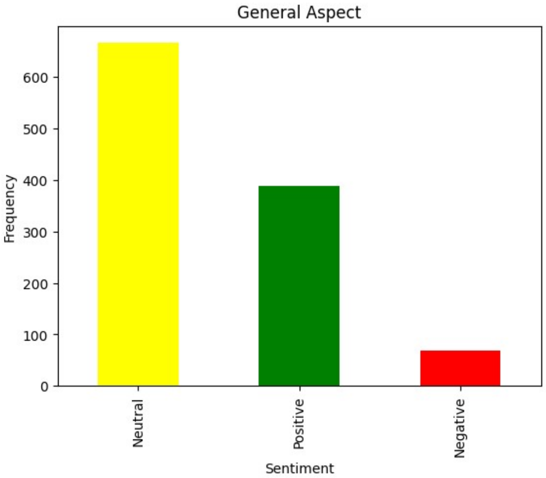


Figure 94: Classification Result for General Aspect

From Figure 9 above, it can be seen that for general aspects, user sentiment towards ChatGPT tends to be neutral with very few negative sentiments.

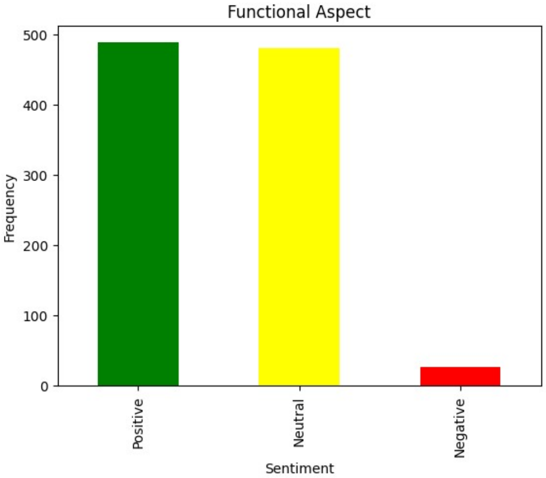


Figure 5: Classification Result for Functional Aspect

From Figure 10 above, it can be seen that in terms of ChatGPT usage, user sentiment tends to be positive or neutral. There are rarely users who give negative sentiments.

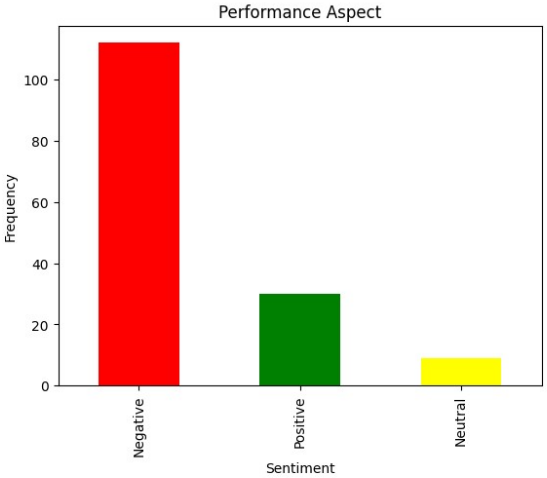


Figure 11: Classification Result for Performance Aspect

From Figure 11 above, it can be concluded that in terms of ChatGPT's performance in providing answers, users tend to give negative sentiments. There are rarely users who give positive or neutral sentiments.

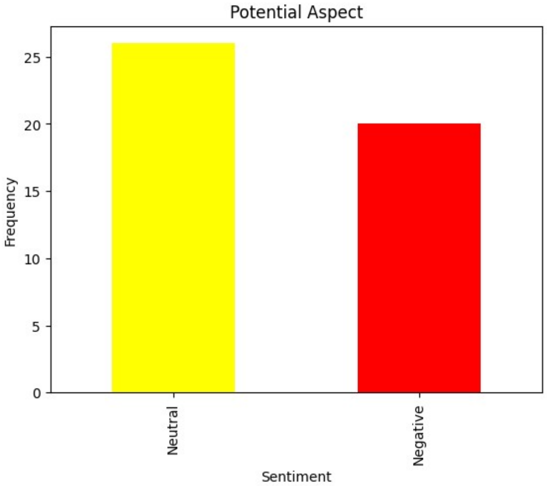


Figure 12: Classification Result for Potential Aspect

From Figure 12 above, it can be seen that user sentiment toward the potential of ChatGPT tends to be neutral. Some users gave negative sentiments. However, there are no positive sentiments obtained from the classification results.

From Figure 9 - 12 above it can be seen that the results of the classification model for each aspect analyzed can provide information about the classification of sentiments on ChatCPT on Twitter For future research, the researcher can identify more in-depth and relevant aspects related to ChatGPT and can also be studied by combining BERT with other methods to get maximum results.

REFERENCES:

[1] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges, and trends," *Knowledge-Based Systems,* vol. 226, p. 107134, 2021.

[2] N. C. Dang, M. N. Moreno-García and F. De la Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics,* vol. 9, no. 3, p. 483, 2020.

[3] M. Al-Tawil, A. Huneiti, R. Shahin, A. A. Zayed, and R. Al-Dibsi, "Twitter Sentiment Analysis Approaches: A Survey," *International Journal of Emerging Technologies in Learning (iJET),* vol. 15, no. 15, pp. 79-93, 2020.

[4] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse and H. Ahmad, "I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data," 2022.

[5] V. Taecharungroj, "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. Big Data and Cognitive Computing," vol. 7, no. 1, p. 35, 2023.

[6] Y. Wang, J. Guo, C. Yuan, and B. Li, "Sentiment Analysis of Twitter Data," *Applied Sciences,* vol. 12, no. 22, p. 11775, 2022.

[7] M. Hoang, O. A. Bihorac and J. Rouces, "Aspect-Based Sentiment Analysis Using BERT," *Proceedings of the 22nd Nordic Conference on Computational Linguistics,* pp. 187-196, 2019.

[8] C. Sun, L. Huang and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* p. 380–385, 2019.

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805,* 2018.

[10] M. Khadhraoui, H. Bellaaj, M. B. Ammar and H. Hamam, "Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study," *Applied Sciences,* vol. 12, no. 6, p. 2891, 2022.

[11] M. M. Abdelgwad, "Arabic aspect-based sentiment classification using BERT," *arXiv preprint arXiv:2107.13290,* 2021.

[12] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets,* vol. 31, no. 3, pp. 685-695, 2021.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems,* vol. 30, 2017.

[14] A. Ligthart, C. Catal and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artif Intell Rev,* vol. 54, p. 4997–5053, 2021.

[15] P. Ray and A. Chakrabarti, "A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis," *Applied Computing and Informatics,* vol. 18, no. 1/2, pp. 163-178, 2022.

[16] Ö. Aydın and E. Karaarslan, "OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare," *Emerging Computer Technologies 2,* pp. pp. 22-31, 2022.

[17] H. Soumare, A. Benkahla, and N. Gmati, "Deep learning regularization techniques to genomics data," *Array,* vol. 11, 2021.

[18] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," *Chaos, Solitons & Fractals,* vol. 140, p. 110120, 2020.