International Journal of Computing and Digital Systems
2025, VOL. 17, NO. 1, 1-12

http://dx.doi.org/10.12785/ijcds/1570999256

Leveraging ALBERT for Sentiment Classification of Long-Form ChatGPT Reviews on Twitter

Wanda Safira[1], Benedictus Prabaswara[1], Andrea Stevens Karnyoto[2] and Bens Pardamean[1,2]

[1]*Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480*

[2]*Bioinformatics and Data Science Research Center, Bina Nusantara University Jakarta, Indonesia, 11480*

Received 16 February 2024, Revised 29 June 2024, Accepted 3 August 2024

Abstract: Sentiment analysis of user-generated content on social media sites reveals important information about public attitudes toward emerging technologies. Researchers face challenges in understanding these impressions, ranging from cursory evaluations to in-depth analyses. Analyzing detailed, long-form reviews exacerbates the difficulty of achieving accurate sentiment analysis. This research addresses the challenge of accurately analyzing sentiments in lengthy and unstructured social media texts, specifically focusing on ChatGPT reviews on Twitter. The study introduces advanced natural language processing (NLP) methodologies, including Fine-Tuning, Easy Data Augmentation (EDA), and Back Translation, to enhance the accuracy of sentiment analysis in such texts. The primary objectives of this research are to improve the accuracy of sentiment analysis for long-form social media texts and to evaluate the effectiveness of the ALBERT transformer-based language model when augmented with data augmentation techniques. Results demonstrate that ALBERT, when augmented with EDA and Back Translation, achieves significant performance improvements, with 81% and 80.1% accuracy, respectively. This research contributes to sentiment analysis by showcasing the efficacy of the ALBERT model, particularly when combined with data augmentation techniques like EDA and Back Translation. The findings highlight the model's capability to accurately gauge public sentiments toward ChatGPT in the complex landscape of lengthy and nuanced social media content. This advancement has implications for understanding public attitudes toward emerging technologies, with potential applications in various domains.

Keywords: Sentiment Analysis, AlBERT, Natural Language Processing, ChatGPT, Long-Form Review

1. Introduction

Sentiment analysis of social media posts is crucial for understanding public perceptions and opinions on emerging technologies like ChatGPT [1][2]. The tremendous amount of content users create on these platforms provides enormous helpful information for assessing public opinion. However, lengthy and unstructured text from platforms like Twitter poses challenges for accurate classification[3]. Texts containing long sentences make it difficult to analyze sentiment, leading to less accurate results. This research addresses sentiment analysis for texts with long sentences, utilizing advanced natural language processing (NLP) methodologies for improved accuracy.

Social media platforms are invaluable digital spaces connecting users and facilitating online social interactions [4][5]. By providing a platform for users to communicate with each other, these digital spaces become vibrant hubs where individuals can not only stay current on news and information about today's world but also actively engage in 24-hour digital social exchanges [6][7]. Through these dynamic networking platforms, people can virtually participate in a wide range of social activities and access up-to-date content from anywhere at any time, fostering a rich and interconnected online community [8][9].

ChatGPT, short for Chat Generative Pre-trained Transformer, is a prominent model in natural language processing (NLP) developed by OpenAI. It is built on the Transformer architecture and has been trained on various internet materials to produce human-like responses to specified cues [10]. ChatGPT has gained attention for its capacity to comprehend and reply to natural language requests. It serves as an efficient tool for various NLP tasks, including question-answering, summarization, sentiment analysis, and more. [11]. The model's versatility and potential applications have sparked significant interest and research in AI and NLP [12].

Twitter opinions can be categorized into positive, neutral, or negative sentiments through Sentiment Analysis. Sentiment Analysis is a Natural Language Processing (NLP) subfield that automatically classifies subjective text opinions. It detects if an opinion or viewpoint carries an

E-mail address: wanda.safira@binus.ac.id, benedictus.prabaswara@binus.ac.id

underlying negative, positive, or neutral tone [13]. This is accomplished by applying NLP techniques to process the linguistic context of the text and determine its emotional inclination or polarity [14]. For example, when a user tweets a ChatGPT review, Sentiment Analysis can automatically evaluate whether the opinion is praiseful or critical based on the text composition. The system then outputs the results of its sentiment categorization as either a positive, negative, or neutral label on the opinion. This allows large volumes of tweets to be efficiently sorted and tagged by underlying sentiment [15].

The rise of transformer-based language models has marked a significant advancement in sophistication and accuracy [16]. Among these models, ALBERT: A Lite BERT for Self-Supervised Learning of Language stands out as a prominent contender, celebrated for its efficient parameter reduction techniques and robust performance across various NLP tasks [17]. Nevertheless, to understand ALBERT's effectiveness and competitive edge comprehensively, it is imperative to compare it with other transformer models such as BERT, XLNet, DistilBERT, and ELECTRA. ALBERT can outperform these models based on its architecture and reduce running time; therefore, the name Lite version of BERT. While ALBERT excels in optimizing efficiency without compromising performance, each model brings distinct strengths and innovations to the table [17]. BERT, renowned for its groundbreaking masked language modeling approach, has established a standard for contextual language comprehension [18]. XLNet, with its bidirectional context understanding, has garnered recognition for capturing intricate contextual dependencies [19]. DistilBERT offers a lightweight alternative with its compact architecture [20]. Meanwhile, ELECTRA introduces a fresh perspective on adversarial training through its discriminator model [21].

The problem of accurately analyzing sentiments in long-form social media texts is multifaceted, involving issues of text complexity, informal language, dynamic content, data imbalance, and computational constraints. Our approach, which integrates the ALBERT model with data augmentation techniques, offers a novel solution that enhances the robustness, efficiency, and accuracy of sentiment analysis in this challenging domain.

In this study, we contribute to evaluating the capability of the ALBERT model in conducting sentiment analysis effectively, especially when faced with datasets containing numerous words in each data instance. We utilize ALBERT, a leading deep learning model, based on numerous comparative studies highlighting its superior performance [22]. Additionally, we contribute to investigating various methodologies to improve the accuracy of the ALBERT model on datasets characterized by many lexical items per data entry. The methodologies we will explore include Fine-Tuning and Data Augmentation techniques such as Easy Data Augmentation and Back Translation [23]. Here are our main contributions can be outlined as follows:

- Utilization of ALBERT, a prominent deep learning model, based on comparative studies showcasing its superior performance.
- Investigation of methodologies to enhance ALBERT's accuracy on datasets with many lexical items per data entry.
- Exploration of Fine-Tuning and Data Augmentation techniques, including Easy Data Augmentation and Back Translation.
- Evaluation of the ALBERT model with performance evaluation metrics, confusion matrix, and results from test data analysis.

The novelty of our approach lies in the strategic combination of data augmentation techniques with the ALBERT model to improve sentiment analysis accuracy in long-form social media texts. This integration not only enhances the model's robustness and contextual understanding but also offers a parameter-efficient solution that can be effectively deployed in various practical applications.

The structure of this paper is outlined as follows: Section 2 offers a comprehensive review of pertinent literature, underscoring the difficulties inherent in social media posts and introducing transformer-based language models like ALBERT. In Section 3, the study's methodology is elaborated upon, covering data collection, preprocessing, and the ALBERT model's implementation. Section 4 delves into performance evaluation metrics and results, presenting confusion matrices and sentiment prediction outcomes. Lastly, Section 5 summarizes the key findings, contributions, and implications for sentiment analysis, especially concerning emerging technologies such as ChatGPT.

2. Literature Review

Sentiment analysis on social media has become crucial for evaluating public sentiment, especially regarding emerging technologies like ChatGPT [24]. However, unstructured and lengthy posts on platforms like Twitter pose accuracy challenges [25]. Lengthy texts with extended phrases make sentiment categorization more manageable, leading to decreased performance. Sentiment analysis, a vital technique in natural language processing (NLP), assesses subjective opinions as positive, neutral, or negative based on linguistic context [26][27]. Transformer models such as ALBERT, BERT, DistilBERT, XLNet, and ELECTRA have propelled improvements in sentiment analysis through contextual language modeling and efficiency [28]. ALBERT, known for its efficient parameter reduction techniques, significantly advances sentiment analysis [17].

The rise of transformer-based language models has significantly advanced sentiment analysis. These models, known for their ability to understand contextual relationships within text, have outperformed traditional machine

learning approaches. BERT (Bidirectional Encoder Representations from Transformers) and its variants have set new benchmarks in various NLP tasks, thanks to their masked language modeling approach that captures intricate contextual dependencies [16][17][18]. However, BERT's large number of parameters can be computationally expensive and memory-intensive, which limits its practicality for real-time applications or for organizations with limited computational resources. To address these limitations, Lan et al. introduced ALBERT (A Lite BERT for Self-Supervised Learning of Language Representations). ALBERT reduces the number of parameters through techniques like parameter sharing and factorized embedding parameterization, making it more efficient while maintaining or even improving performance on various NLP tasks. ALBERT's innovations, such as cross-layer parameter sharing and the Sentence Order Prediction (SOP) task, enhance its contextual understanding capabilities while reducing computational demands [17].

Several comparative studies have evaluated the performance of transformer models like BERT, ALBERT, XLNet, DistilBERT, and ELECTRA in various NLP tasks. These studies highlight the strengths and weaknesses of each model. For instance, XLNet's autoregressive pretraining method captures bidirectional context better than BERT, while DistilBERT offers a lightweight alternative by distilling knowledge from BERT, making it faster and more efficient [19][20]. ELECTRA introduces a novel approach by pretraining text encoders as discriminators rather than generators, showing competitive performance with fewer computational resources [21]. Comparative evaluations reveal that while ALBERT excels in parameter efficiency and contextual understanding, each model brings unique advantages depending on the specific task and dataset. These findings underscore the importance of selecting the appropriate model and fine-tuning strategies to optimize performance for specific applications [28][29][30].

Several studies have explored the application of transformer models and data augmentation techniques in sentiment analysis. Alamoudi and Alghamdi evaluated the efficacy of deep learning and transfer learning models for sentiment analysis on Yelp reviews. Their research introduced an unsupervised technique for aspect extraction based on semantic similarity and pre-trained language models, demonstrating that ALBERT outperformed other models, providing higher accuracy and valuable insights for businesses from customer feedback [29]. Dwivedi et al. investigated the use of transformer models for sentiment analysis in social media, highlighting the importance of model interpretability and the challenges posed by unstructured social media text. Their results emphasized the potential of models like BERT and ALBERT in accurately capturing public sentiment [24]. Isnan et al. conducted sentiment analysis for TikTok reviews using VADER sentiment analysis and SVM models, demonstrating the effectiveness of combining traditional machine learning methods with modern NLP techniques to analyze social media data [25]. Chaurasia and Sherekar analyzed Twitter data using convolutional neural networks (CNNs) for sentiment analysis, providing insights into the performance of deep learning models on short text data and showcasing the challenges and opportunities in social media sentiment analysis [13]. Shirke and Agrawai investigates the effectiveness of token-based text augmentation techniques for text classification tasks in Indic languages. The findings demonstrate that token-based augmentations can significantly enhance the performance of text classification models, particularly in the context of Indic languages[31].

Fine-tuning and data augmentation techniques are assessed to enhance the precision of transformer-based language models such as ALBERT, BERT, DistilBERT, XLNet, and ELECTRA [28][30].

Fine-tuning is a crucial technique in deep learning, where a neural network's weights are initialized using pretrained models. This method leverages previously learned features, allowing the network to adapt to new tasks more efficiently and with fewer training resources compared to training from scratch [32]. This method capitalizes on the knowledge gained by the pre-trained model on a large and diverse dataset, enabling the network to learn general features and representations [32]. By starting with these pre-existing weights, the model benefits from the wealth of information encoded in the initial parameters [33].

Data augmentation is a crucial technique for improving the performance of NLP models, especially when dealing with limited labeled data. Techniques like Easy Data Augmentation (EDA) and Back Translation have shown significant improvements in model robustness and accuracy. EDA involves four simple yet effective operations: synonym substitution, random insertion, position swapping, and word dropping. These operations increase the diversity of the training data, making the model more robust to different linguistic patterns and reducing overfitting. Wei and Zou demonstrated that EDA can improve the performance of text classification models by increasing data diversity and providing varied training examples [28]. Back Translation is a robust technique that entails converting text into another language and then translating it back to the original language. The process generates diverse sentence structures while retaining the original meaning, enhancing the model's ability to generalize to different linguistic patterns and improving performance on machine translation and text classification tasks [34].

The literature highlights the evolution and advancements in sentiment analysis through transformer-based models and data augmentation techniques. he problem of accurately analyzing sentiments in long-form social media texts is multifaceted, involving issues of text complexity, informal language, dynamic content, data imbalance, and computational constraints. Our approach, which integrates the ALBERT model with data augmentation techniques, offers a novel so-

lution that enhances the robustness, efficiency, and accuracy of sentiment analysis in this challenging domain. ALBERT stands out for its efficiency and robust performance, making it a promising model for analyzing sentiments in long-form social media texts. The incorporation of data augmentation methods like EDA and Back Translation further enhances the model's accuracy, demonstrating their importance in NLP research. By building on these findings, our study aims to contribute to the field by evaluating the effectiveness of ALBERT in sentiment analysis, particularly for lengthy and unstructured social media content.

3. Method

An overview of how the research will be conducted in a flowchart. Figure 1 shows the steps for this research involving collecting the dataset, preprocessing the data that has been collected, initializing and fine-tuning the model, following the data is augmented, and then evaluating the model by analyzing the results.

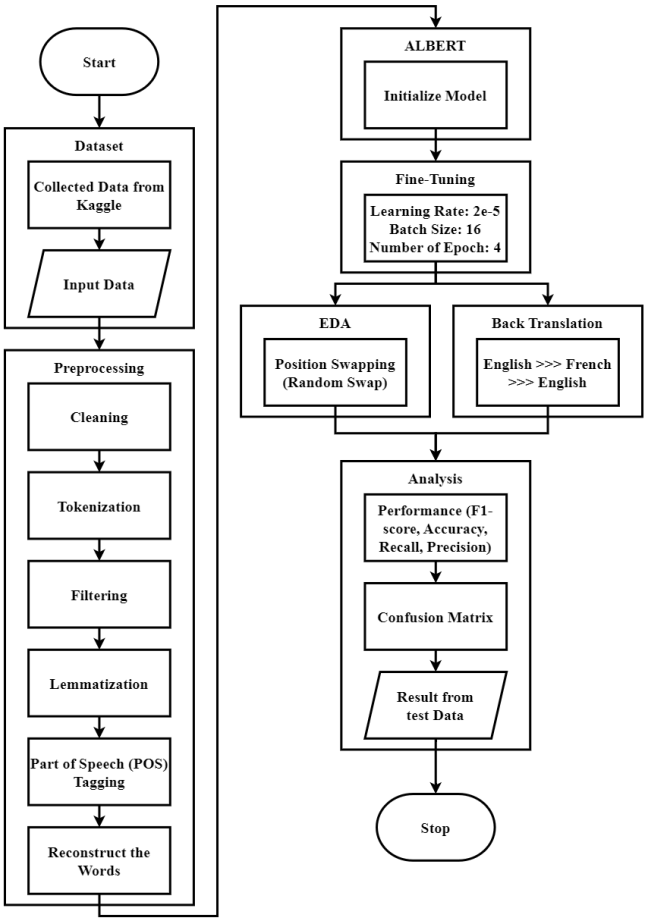

Figure 1. Flowchart Sentiment Analysis

A. Dataset

The dataset was sourced from Kaggle, specifically curated to include Twitter reviews of ChatGPT. This dataset was chosen because it provides a diverse range of opinions and sentiment expressions, making it ideal for training a sentiment analysis model. Initially, the dataset contained numerous raw tweets. We filtered these tweets to exclude irrelevant or non-English tweets, ensuring that the dataset was focused and relevant to the study's objectives. Each tweet in the dataset was pre-labeled with one of three sentiment categories: Positive, Neutral, or Negative. These labels were verified to ensure accuracy and consistency in the sentiment analysis.

Prior research conducted by Kaggle user Sujal Neupane involved testing the DeBERTa model, resulting in an evaluation accuracy of 28.7%, a performance level considered suboptimal for a transformer model. Based on Table I, the long-format dataset consists of several data entities labeled as 10539 Positive, 10539 Neutral, and 10539 Negative. Further analysis of this dataset is expected to provide in-depth insights into opinions and sentiments surrounding the use of ChatGPT.

TABLE I. Data Distribution

Polarity	Number of Data
Positive	10539
Neutral	10539
Negative	10539

The data analysis divides the dataset into three categories: long, medium, and short, using quartiles. The specifications are as follows: the long category has a word range from 25 to 64 words, the medium category has a word range from 15 to 24 words, and the short category has a word range under 15 words. In this context, only the dataset with the long category is selected because it is considered the most relevant to achieving the objectives of this study.

TABLE II. Data Split

Segment	Number of Data
Training	18492
Validation	6563
Test	6562

As shown in Table II, the balanced dataset was then split into three segments: Training, Validation, and Testing. The Training set comprised 60% of the data, while the Validation and Testing sets each comprised 20%. This split ensured that the model was trained on a substantial amount of data while retaining enough data for robust validation and testing.

B. Pre-processing

Preprocessing is necessary in sentiment analysis to prepare and transform raw text data before feeding it into machine learning models [35]. The effective preprocessing analysis makes patterns in the text more apparent for the

model to learn from and improves accuracy on unseen real-world data [36][37]. The primary procedures of the pre-processing stage are outlined as follows:

- Cleaning: The cleaning process systematically removes Twitter handles, which are identified by the '@' symbol, and hashtags that begin with '#'. Retweet marks like 'RT' are stripped away to avoid duplication of content. Links, which often clutter the text, are also removed by detecting common URL patterns. Last, non-word characters such as punctuation marks and special symbols are filtered out, ensuring that only meaningful words remain, creating a cleaner and more uniform dataset for analysis.

- Tokenization: The process of of dividing cleaned text into smaller, manageable units known as tokens. Using NLTK word tokenizer, the text is split into words and punctuation marks. This step creates a list of tokens that can be easily manipulated in subsequent stages.

- Filtering: This process involves eliminating stop words, which are frequent terms like 'the', 'is', and 'and' that do not contribute significant meaning in sentiment analysis. By filtering out these stop words, we reduce noise in the data. The remaining tokens, which include nouns, verbs, adjectives, and adverbs, are considered significant and are retained for further analysis.

- Lemmatization: The process of standardizing tokens to their root form, ensuring consistency in the text data. Words like 'writing' are converted to 'write', and 'studying' is converted to 'study'. This step reduces the complexity of the text by grouping different forms of a word into a single term. Using tools such as WordNetLemmatizer from nltk, tokens are lemmatized with consideration of their part-of-speech tags, accurately converting them to their base forms.

- Part Of Speech (POS) Tagging: Each token is assigned a part-of-speech label, identifying it as a noun, verb, adjective, or another grammatical category, which aids in understanding the text's grammatical structure and meaning. Using nltk's POS tagging tools, each token is tagged appropriately. This step allows us to select only the tokens with significant POS tags, such as nouns, verbs, and adjectives, for sentiment analysis.

- Reconstruct the words: The preprocessed tokens are reconstructed into a cleaned sentence string, ready for sentiment analysis. The filtered and lemmatized tokens are joined back together to form coherent sentences. This final step prepares the cleaned sentence string for input into sentiment analysis models, ensuring that the text is in the optimal format for accurate analysis.

TABLE III. Clean Data Comparison

Uncleaned Data	**Cleaned Data**
#ChatGPT will replace @Google as search engine as it's only limited to crawling and not understanding exact search. Not atleast as effectively as this model. Internet is about to be so much cognitively advance soon @sama	replace search engine exact search at least effectively internet much cognitively advance soon
keep seeing tiktoks about how chatgpt is gonna ruin all education and so im putting in some of my old homework problems to see how it does and so far its doing stuff but just completely WRONG Ÿ˜ like thermo stuff is gonna be done by hand for a few more years at least	keep ruin education old homework problem far stuff completely wrong thermo stuff hand year least
It's amazing how people on Twitter are excited about #ChatGPT. If you think media manipulation is bad just wait for the AI manipulation IF it becomes mainstream. Don't know why @elonmusk is happy about it... After all that Twitter Files show...	twitter medium manipulation bad wait ai manipulation mainstream know happy twitter file show

Following the execution of various pre-processing steps detailed earlier.

Table III shows that the preprocessing task shows excellent results from the uncleaned data. The preprocessing task makes the uncleaned data more assertive and more understandable for the model to learn.

C. ALBERT

The ALBERT framework is a comprehensive development of the BERT (Bidirectional Encoder Representations from Transformers) architecture aimed at increasing efficiency and scalability in language representation [12]. A key innovation lies in introducing parameter sharing across layers, reducing redundancy, and rendering the model more

parameter-efficient. This is

achieved by sharing weights and biases within the feedforward neural network (FFN) across layers [12]. The embedding layer is factorized into token embeddings and token-type embeddings, reducing the overall number of parameters. Cross-layer parameter sharing is also introduced, further optimizing parameter utilization. The training objective involves Sentence Order Prediction (SOP), wherein the model predicts the correct order of sentences in a document, facilitating improved contextual understanding.

Figure 2 shows the overall architecture of ALBERT. ALBERT uses multiple transformer layers interspersed with a more comprehensive particular transformer layer to learn contextual relationships between words in the input text. The output embeddings are fed to a task-specific classifier. The goal is to achieve BERT-level performance with significantly fewer parameters.

$$h_i = LN\left(h_{i-1} + FNN\left(LN\left(h_{i-1} \cdot W_i + b_i\right)\right)\right) \quad (1)$$

where:

- h_i is the output layer of the i-th layer,
- LN is layer normalization,
- FNN is a feedforward neural network,
- W_i and b_i are the shared weights and biases across layers.

The parameter sharing is achieved through and which are the same for all layers.

The key idea behind ALBERT is parameter sharing across layers. In BERT, every layer possesses its unique set of parameters. , leading to many parameters in the model. ALBERT reduce the number of parameters by sharing them across different layers. This enables a more parameter-efficient model while maintaining performance.

D. Fine-tuning

Fine-tuning is a transfer method of learning that involves retraining a previously learned neural network on a fresh dataset or task [32]. The model in this research was tuned using the Adam optimizer, and the fine-tuned parameters are:

Learning Rate: 2e-5

Batch Size: 16

Number of Epoch: 4

We utilize fine-tuning to adapt a pre-trained ALBERT language model to our specific text classification task. ALBERT is a transformer-based model trained on a large corpus of unlabelled text data using a novel permutation language modeling objective [28].

E. Easy Data Augmentation (EDA)

EDA, which stands for Easy Data Augmentation, is a simple yet effective approach for improving the robustness and preventing overfitting text classification models. It involves applying four straightforward data transformation operations to the training data:

- Synonym Substitution: Instead of randomly selecting non-halt terms and replacing them with arbitrary synonyms, randomly choose n non-stop words in each sentence and replace them with random synonyms.
- Random Insertion: Select an arbitrary synonym for an arbitrary non-stop word in the sentence. Insert that synonym at an arbitrary placement within the sentence. Redo n times.
- Position Swapping: Randomly select two words from the sentence and switch their locations. Repeat n times.
- Word Dropping: Randomly eliminate each word in the sentence with probability p.

In this research, we will use Position Swapping (Random Swap) as it results in better performance and accuracy than other methods. Using Random Swap from Easy Data Augmentation on the training dataset, the original training length is 18,492 and the augmented length is 36,984.

F. Back translation

Back translation augmentation is a powerful technique employed in natural language processing, text data is translated into another language and then translated back into the original language [34]. After using Back Translation, the augmented length of the data becomes 20583, initially 18492. This means the Back Translation added 2092 new text data that has been translated to French and back to English.

Figure 3 shows how Back Translation proceeds by translating the original data to the target language and translating it back to the original language, producing different sentences with the same meaning.

G. Algorithm

The comprehensive step-by-step procedure is meticulously outlined in Algorithm 1.

Algorithm 1: Sentiment Analysis Using ALBERT with Data Augmentation.

Input: Raw Twitter reviews dataset DDD, ALBERT model MMM, and Data Augmentation Techniques (EDA, Back Translation).

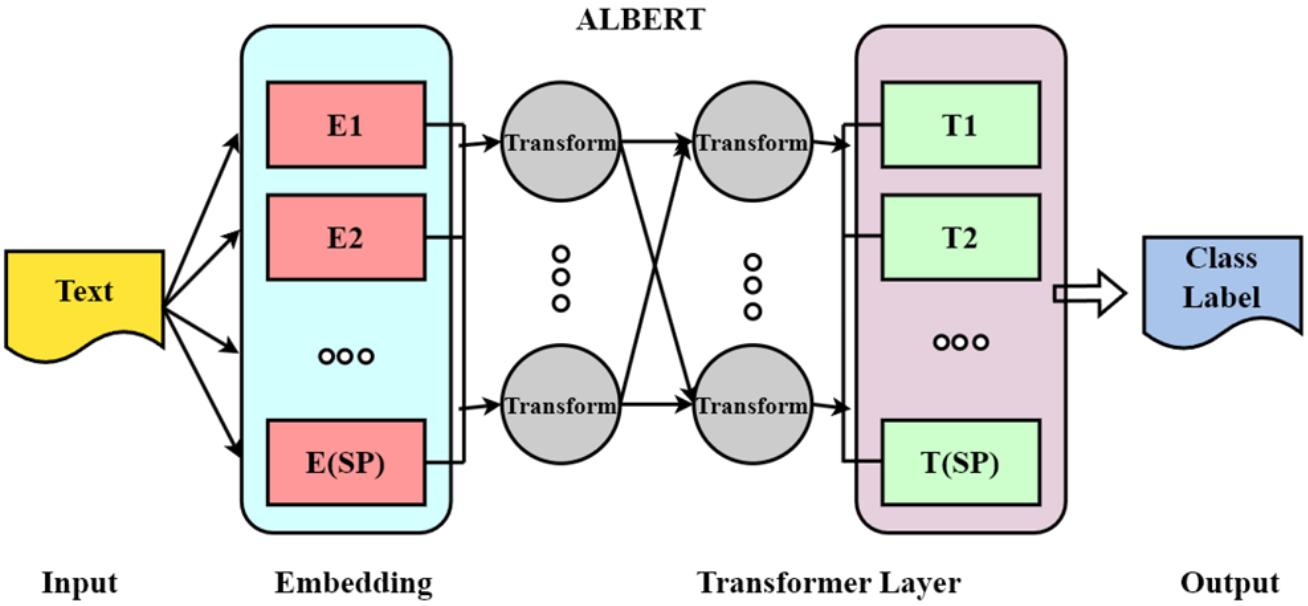

Figure 2. ALBERT Architecture

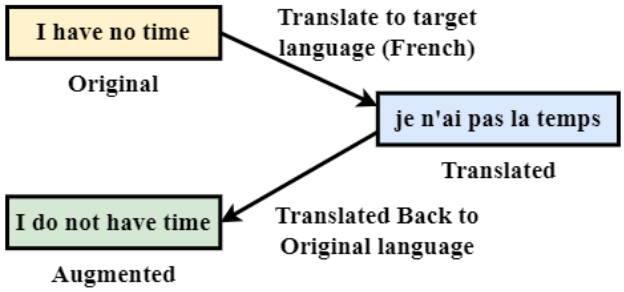

Figure 3. Back Tranlsation Process

Output: The sentiment analysis model was developed to analyze and categorize textual data into positive, negative, or neutral sentiments.

1. **Data Collection**:

- Collect the dataset DDD containing Twitter reviews of ChatGPT from Kaggle.

2. **Data Preprocessing**:

- **Step 1**: Filter out non-English tweets.
- **Step 2**: Remove Twitter handles, hashtags, retweet marks, links, and non-word characters using regular expressions.
- **Step 3**: Tokenize the text into individual tokens (words) using NLTK's word tokenizer.
- **Step 4**: Remove stop words and retain significant tokens.
- **Step 5**: Lemmatize tokens to their root form.
- **Step 6**: Perform Part-of-Speech (POS) tagging and retain significant POS tags.
- **Step 7**: Reconstruct cleaned sentence strings from preprocessed tokens.

3. **Sample Selection**:

- **Step 1**: Divide the dataset into three categories based on tweet length: long (25-64 words), medium (15-24 words), and short (less than 15 words).
- **Step 2**: Select only long-form tweets for the study.
- **Step 3**: Balance the dataset to have equal representation of Positive, Neutral, and Negative sentiment categories.
- **Step 4**: Divide the balanced dataset into three parts: 60% for training, 20% for validation, and 20% for testing.

4. **Data Augmentation**:

Easy Data Augmentation (EDA):

- **Step 1**: Apply synonym substitution on randomly selected non-stop words.

- **Step 2**: Perform random insertion of synonyms at arbitrary positions.
- Step 3: Conduct position swapping by swapping positions of randomly selected words.
- **Step 4**: Execute word dropping by randomly removing words with a certain probability.

Back Translation:

- **Step 1**: Translate text to another language (e.g., French) and then back to English to create diverse sentence structures.
- **Combine**: Apply EDA and Back Translation to augment the training dataset, increasing data diversity and robustness.

5. **Model Initialization**:

- Initialize the ALBERT model MMM with pre-trained weights.

6. **Fine-Tuning the Model**:

- **Step 1**: Configure the model's hyperparameters: The learning rate , batch size , and the number of epochs .
- **Step 2**: Fine-tune the ALBERT model MMM on the training dataset using the Adam optimizer.

7. **Model Evaluation**:

- **Step 1**: Evaluate the trained model on the validation dataset to tune hyperparameters and avoid overfitting.
- **Step 2**: Assess the final model's performance on the test dataset using accuracy, F1-score, precision, and recall metrics.
- **Step 3**: Generate confusion matrices to visualize classification performance and identify areas of improvement.

8. **Output**:

- Output the fine-tuned ALBERT model capable of performing sentiment analysis on long-form Twitter reviews.

4. Result and analysis

A. *Performance evaluation*

The model performance analysis for sentiment analysis of long-form texts reveals notable trends and insights. Among the models evaluated, ALBERT demonstrates strong performance, particularly when augmented with Easy Data Augmentation (EDA) or Back Translation techniques, achieving 81% and 80.1% accuracy, respectively. Table IV compares model performance metrics for sentiment analysis of long-form texts.

The performance of Naive Bayes and LSTM + Random Forest from previous research demonstrates noteworthy characteristics. Naive Bayes, characterized by its simplicity and efficiency, achieved an accuracy of 72%, indicating a reasonable level of predictive capability. However, its F1-Score, Precision, and Recall metrics are comparatively lower at 65%, 71%, and 66% respectively. Notably, BERT and ELECTRA also exhibit competitive performance, with accuracies ranging from 79.3% to 79.5%. Furthermore, integrating data augmentation techniques consistently enhances model performance across various metrics, including F1-Score, Precision, and Recall.

However, models like XLNet and DistilBERT show slightly lower accuracy and F1-Score, with 78.7% and 77.1%, respectively, indicating the potential for improvement. Fine-tuning strategies or alternative data augmentation methods may enhance the performance of these models. These findings underscore the practical utility of transformer-based models in sentiment analysis of long-form text.

B. *Training accuracies*

Training accuracies measure how well a sentiment analysis model fits the data it was trained on. While training accuracy is essential for tracking model progress throughout development, validation accuracy should drive model selection to avoid overfitting. Figures 4 and 5 show the training accuracy from the ALBERT model using EDA and Back Translation.

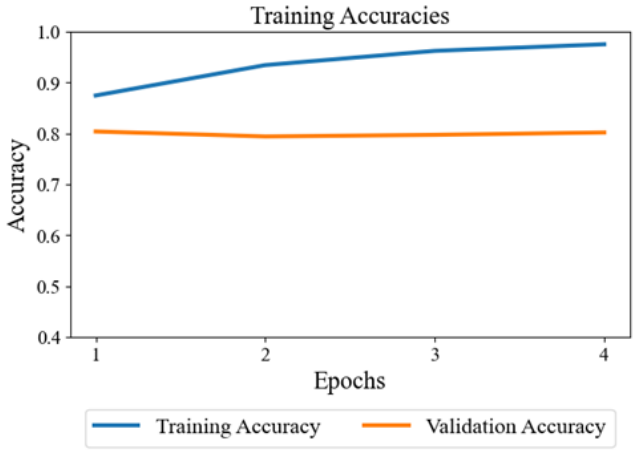

Figure 4. The Training Accuracy of ALBERT with EDA.

C. *Confusion matrix*

A confusion matrix is a tool used to evaluate the accuracy of a classification model by comparing its predictions to the actual outcomes on a test dataset. It provides a visual

TABLE IV. Comparison of Model Performance Metrics

Model	Accuracy	F1-Score	Precision	Recall
LSTM + Random Forest	55%	53%	55%	54%
Naive Bayes	72%	65%	71%	66%
Electra + EDA	78.9%	79.2%	79.8%	78.9%
XLNet + EDA	78.7%	78.4%	78.7%	78.4%
BERT + EDA	79.3%	79.5%	80.1%	79.3%
DistilBERT + EDA	77.1%	77.5%	79.3%	77.1%
ALBERT + EDA (Proposed Method)	**81%**	**81.2%**	**81.5%**	**81%**
ALBERT + Back Translation (Proposed Method)	80.1%	80.8%	81.4%	80.1%

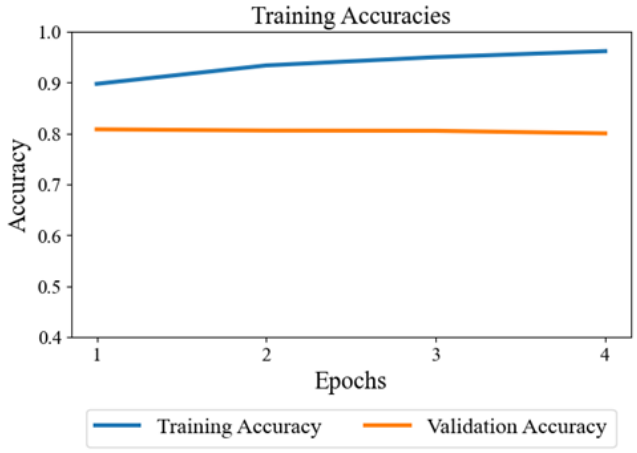

Figure 5. The Training Accuracy of ALBERT with Back Translation

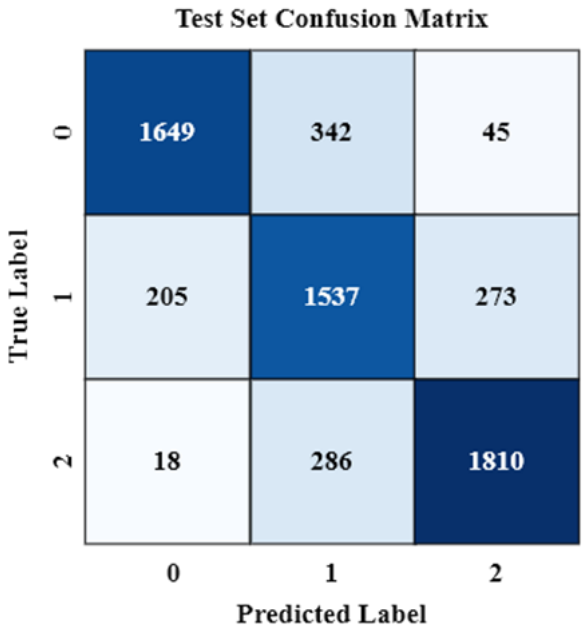

Figure 6. The Confusion Matrix of ALBERT with EDA.

representation of the model's performance by displaying the correlation between the true labels and the predicted labels. In the matrix, each row corresponds to the actual class instances, and each column corresponds to the predicted class instances.

Based on the accuracy results of the model, Figure 6 and Figure 7 show the confusion matrix of the ALBERT model using EDA and Back Translation. Mapping words into negative, neutral, and positive, the confusion matrix displays a list of words that the model accurately and inaccurately classifies into each sentiment category.

The confusion matrix analysis provides valuable insights into the strengths and weaknesses of the ALBERT model when augmented with EDA and Back Translation techniques. These results guide further improvements and highlight the model's effectiveness in handling complex, long-form social media texts.

The confusion matrices reveal that both EDA and Back Translation techniques significantly improve the model's performance in sentiment analysis. However, certain challenges remain, such as:

1. **False Positives and False Negatives**:

- For EDA, the relatively high number of neutral instances classified as positive (342) and neutral instances classified as negative (273) indicate that the model occasionally struggles with distinguishing between neutral sentiments and other sentiments.

- For Back Translation, the number of neutral instances classified as positive (348) and neutral instances classified as negative (154) are similar to those of EDA, showing that while Back Translation improves overall accuracy, fine-tuning may still be necessary to reduce these misclassifications.

2. **Impact on Real-World Applications**:

- High precision and recall rates for positive sentiments indicate that the model is reliable for identifying positive feedback, which is crucial for applications like customer satisfaction analysis and brand monitoring.

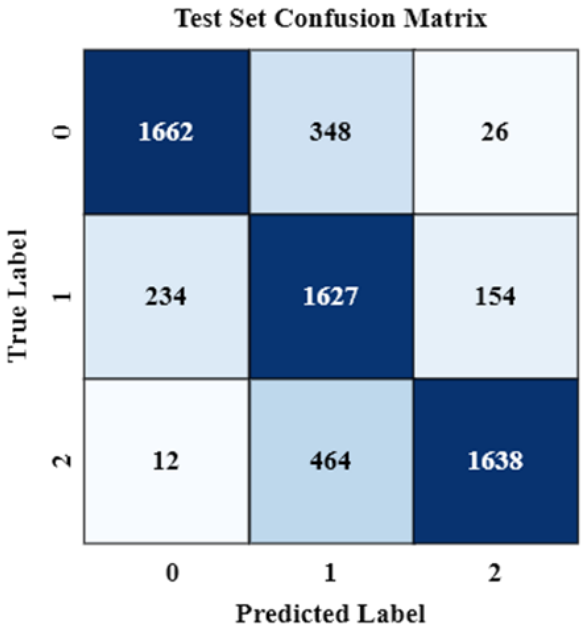

Figure 7. The Confusion Matrix of ALBERT with Back Translation.

TABLE V. Sentiment Prediction Result on ALBERT with EDA Test Data

Polarity	Text
Positive	I have been fascinated my #OpenAi release of the conversational GPT-3 model, #ChatGPT. This incredible language model, finely tuned for conversations, creates human-like responses and is capable of remembering the conversation and building upon the stored knowledge.
Neutral	My daughter just asked me how to convert .tpl file to .abr file. Not knowing what they were, I was able to give her step-in-step instructions under 10 seconds. Between me and #ChatGPT, I am better, with ... https://t.co/daprlCr1XI
Negative	OpenAi GPT3 and beyond, this is the single most threat to computer programming jobs. \n\nTrust me, itâ€™s a long road ahead for devs. \n\nThe ChatGPT should be a source of worry for Alphabet.

- The balanced performance across all sentiment categories ensures that the model can be effectively used in diverse applications where understanding public opinion across a spectrum of sentiments is necessary.

D. Result from test data

After conducting the analysis, Table V and Table VI display the sentiment prediction results on the ALBERT with EDA and Back Translation test data. Testing with this new data is performed to validate further how accurately the model predicts sentiments in data it has not seen before.

TABLE VI. Sentiment Prediction Result on ALBERT with Back Translation Test Data

Polarity	Text
Positive	ChatGPT is neat, but using open source code in closed source coding is not; know your attribution:\n"Output generated by code generation features of our Services, including OpenAI Codex, may be subject to third party licenses, including, without limitation, open source licenses."
Neutral	This article was written 90% by @OpenAI's ChatGPT. I was working on a SwiftUI app today and prompted it to help me change the background color. Initially the answer was incorrect, but after suggesting that it use a ZStack, the answer was
Negative	ChatGPT is really bad at math. It can't even pick the larger of two numbers.\n\nBut it's not just for humanitiesâ€"it's actually really good at writing code.\n\nSo, running with @sjwhitmore's latest thread on building your own ChatGPT-like tool, I made ChatGPT's math-nerd alter ego.

5. Conclusion

This research focuses on using the ALBERT model to analyze sentiments in long Twitter reviews about ChatGPT. It highlights the need for advanced natural language processing (NLP) methods to improve sentiment analysis accuracy due to the complexity of social media content. Results show that ALBERT, combined with data augmentation techniques, achieves high accuracy rates of 81% and 80.1%. This study demonstrates the effectiveness of ALBERT in handling lengthy social media texts and emphasizes the importance of data augmentation for better model performance. It suggests that transformer-based models like ALBERT are valuable for understanding public opinions on emerging technologies like ChatGPT.

The high accuracy achieved by the ALBERT model in this study signifies a substantial advancement in sentiment analysis for long-form social media content. Future research could explore additional data augmentation techniques and fine-tuning strategies to further enhance model performance. Limitations of this study include the focus on English tweets only, which may not generalize to other languages. By addressing these challenges and exploring future directions, the potential applications of this model can be broadened, further contributing to the field of natural language processing.

6. Acknowledgement

The authors would like to express their gratitude towards the ChatGPT Sentiment Analysis dataset collectors and

Sujal Neupane's contributions, reviewers, and peers for insightful feedback and research participants who unintentionally contributed through Kaggle comments to support the execution of this research.

References

[1] Y. Su and Z. J. Kabala, "Public perception of chatgpt and transfer learning for tweets sentiment analysis using wolfram mathematica," *Data*, vol. 8, no. 12, Dec. 2023.

[2] A. A. Hidayat and B. Pardamean, "Count time series modelling of twitter data and topic modelling: A case of indonesia flood events," in *Procedia Comput. Sci.*, vol. 227, Aug. 2023, pp. 805–812.

[3] F. Kateb and J. Kalita, "Classifying short text in social media: Twitter as case study," *Int. J. Comput. Appl.*, vol. 111, pp. 1–12, Feb. 2015.

[4] J. Davis, *Social Media*, 2016.

[5] K. Purwandari, J. W. C. Sigalingging, T. W. Cenggoro, and B. Pardamean, "Multi-class weather forecasting from twitter using machine learning approaches," in *Procedia Comput. Sci.*, vol. 179, Jan. 2021, pp. 47–54.

[6] K. Purwandari, R. B. Perdana, J. W. C. Sigalingging, R. Rahutomo, and B. Pardamean, "Automatic smart crawling on twitter for weather information in indonesia," in *Procedia Comput. Sci.*, vol. 227, Aug. 2023, pp. 795–804.

[7] I. Nurlaila, R. Rahutomo, K. Purwandari, and B. Pardamean, "Provoking tweets by indonesia media twitter in the initial month of coronavirus disease hit," in *2020 International Conference on Information Management and Technology (ICIMTech)*. Bandung, Indonesia: IEEE, Aug. 2020, pp. 409–414.

[8] Y. K. D. et al., "Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *Int. J. Inf. Manag.*, vol. 66, p. 102542, Oct. 2022.

[9] K. Purwandari, A. S. Perbangsa, J. W. C. Sigalingging, A. A. Krisna, S. Anggrayani, and B. Pardamean, "Database management system design for automatic weather information with twitter data collection," in *2021 International Conference on Information Management and Technology (ICIMTech)*, Aug. 2021, pp. 326–330.

[10] P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Jan. 2023.

[11] B. Lund and T. Wang, "Chatting about chatgpt: How may ai and gpt impact academia and libraries?" *Libr. Hi Tech News*, vol. 40, Feb. 2023.

[12] D. Cortiz, *Exploring Transformers models for Emotion Recognition: a comparision of BERT, DistilBERT, RoBERTa, XLNET and ELECTRA*, 2022.

[13] S. Chaurasia and S. Sherekar, "Sentiment analysis of twitter data by natural language processing and machine learning," 2023, pp. 59–70.

[14] K. Svensson and J. Hagelbäck, "Sentiment analysis with convolutional neural networks."

[15] S. Pipin, F. Sinaga, S. Winardi, and M. Hakim, "Sentiment analysis classification of chatgpt on twitter big data in indonesia using fast r-cnn," *J. MEDIA Inform. BUDIDARMA*, vol. 4, pp. 2137–2148, Oct. 2023.

[16] B. G. Bokolo and Q. Liu, "Deep learning-based depression detection from social media: Comparative evaluation of ml and transformer techniques," *Electronics*, vol. 12, no. 21, p. 4396, Oct. 2023.

[17] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv*, Feb. 2020.

[18] S. R. et al., "Analyzing sentiments regarding chatgpt using novel bert: A machine learning approach," *Information*, vol. 14, no. 9, Sep. 2023.

[19] T. Han, Z. Zhang, M. Ren, C. Dong, X. Jiang, and Q. Zhuang, "Text emotion recognition based on xlnet-bigru-att," *Electronics*, vol. 12, no. 12, Jan. 2023.

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv*, Feb. 2020.

[21] K. Clark, M.-T. Luong, and Q. V. Le, "Electra: Pre-training text encoders as discriminators rather than generators," 2020.

[22] S. F. N. Azizah and W. Widiarto, "Performance analysis of transformer based models (bert, albert and roberta) in fake news detection."

[23] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *J. Big Data*, vol. 8, no. 1, p. 101, Jul. 2021.

[24] K. R. Manda, "Sentiment analysis of twitter data using machine learning and deep learning methods."

[25] M. Isnan, G. N. Elwirehardja, and B. Pardamean, "Sentiment analysis for tiktok review using vader sentiment and svm model," in *Procedia Comput. Sci.*, vol. 227, Aug. 2023, pp. 168–175.

[26] K. Purwandari, R. Rahutomo, J. W. C. Sigalingging, M. A. Kusuma, A. Prasetyo, and B. Pardamean, "Twitter-based text classification using svm for weather information system," in *2021 International Conference on Information Management and Technology (ICIMTech)*, Aug. 2021, pp. 27–32.

[27] K. Purwandari, T. W. Cenggoro, J. W. C. Sigalingging, and B. Pardamean, "Twitter-based classification for integrated source data of weather observations," *IAES Int. J. Artif. Intell. IJ-AI*, vol. 12, no. 1, Mar. 2023.

[28] S. Casola, I. Lauriola, and A. Lavelli, "Pre-trained transformers: an empirical comparison," *Mach. Learn. Appl.*, vol. 9, p. 100334, Sep. 2022.

[29] E. Alamoudi and N. Alghamdi, "Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings," *J. Decis. Syst.*, vol. 30, pp. 1–23, Jan. 2021.

[30] E. Zhang, A. Cheok, Z. Pan, J. Cai, and Y. Yan, "From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models," *Sci*, vol. 5, p. 46, Dec. 2023.

[31] R. Shirke and A. Agrawal, "Performance analysis of token-based text augmentation techniques on text classification tasks in indic languages," in *2023 9th International Conference on Smart Computing and Communications (ICSCC)*, 2023, pp. 168–172.

[32] P. D. Alfano, V. P. Pastore, L. Rosasco, and F. Odone, "Top-tuning: A study on transfer learning for an efficient alternative to fine tuning for image classification with fast kernel methods," *Image Vis. Comput.*, vol. 142, p. 104894, Feb. 2024.

[33] M. U. et al., "Etcnn: Extra tree and convolutional neural network-based ensemble model for covid-19 tweets sentiment classification," *Pattern Recognit. Lett.*, vol. 164, pp. 224–231, Dec. 2022.

[34] M. Fadaee and C. Monz, "Back-translation sampling by targeting difficult words in neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 436–446.

[35] D. Sudigyo, A. A. Hidayat, R. Nirwantono, R. Rahutomo, J. P. Trinugroho, and B. Pardamean, "Literature study of stunting supplementation in indonesian utilizing text mining approach," in *Procedia Comput. Sci.*, vol. 216, Jan. 2023, pp. 722–729.

[36] T. N. Ph.D and A. Amalanathan, "Data preprocessing in sentiment analysis using twitter data," vol. 3, pp. 89–92, Jul. 2019.

[37] R. Rahutomo, F. Lubis, H. H. Muljo, and B. Pardamean, "Preprocessing methods and tools in modelling japanese for text classification," in *2019 International Conference on Information Management and Technology (ICIMTech)*, Aug. 2019, pp. 472–476.