Code-Mixed Sentiment Analysis using Transformer for Twitter Social Media Data

Laksmita Widya Astuti, Yunita Sari, Suprapto
Department of Computer Science and Electronics, FMIPA UGM Yogyakarta, Indonesia

***Abstract*—The underrepresentation of the Indonesian language in the field of Natural Language Processing (NLP) can be attributed to several key factors, including the absence of annotated datasets, limited language resources, and a lack of standardization in these resources. One notable linguistic phenomenon in Indonesia is code-mixing between Bahasa Indonesia and English, which is influenced by various sociolinguistic factors, including individual speaker characteristics, the linguistic environment, the societal status of languages, and everyday language usage. In an effort to address the challenges posed by code-mixed data, this research project has successfully created a code-mixed dataset for sentiment analysis. This dataset was constructed based on keywords derived from the sociolinguistic phenomenon observed among teenagers in South Jakarta. Utilizing this newly developed dataset, we conducted a series of experiments employing different pre-processing techniques and pre-trained models. The results of these experiments have demonstrated that the IndoBERTweet pre-trained model is highly effective in solving sentiment analysis tasks when applied to Indonesian-English code-mixed data. These experiments yielded an average precision of 76.07%, a recall of 75.52%, an F-1 score of 75.51%, and an accuracy of 76.56%.**

Keywords*—Sentiment analysis; code-mixed; BERT; bahasa Indonesia

I. Introduction

Indonesia, the fourth most populous country globally, boasts a population exceeding 279 million people, according to estimates provided by the Central Intelligence Agency (CIA) World Factbook. Despite its substantial population, the Indonesian language remains inadequately represented within the realm of Natural Language Processing (NLP). According to Koto et al. [1], this paucity of representation can be attributed to three principal factors: the dearth of annotated datasets, restricted language resources, and a lack of standardization in these resources.

The prevalence of code-mixing, the linguistic phenomenon involving the amalgamation of multiple languages in communication, is influenced by sociolinguistic determinants. As highlighted by Anastassiou and Andreou [12], these determinants encompass individual speaker attributes, the linguistic milieu, the societal status of languages, and everyday language utilization. The coexistence of diverse local languages, regional dialects, and Bahasa Indonesia (the national language) engenders an environment conducive to code-mixing as individuals navigate various linguistic systems. The escalating influence of globalization has engendered a heightened reliance on the English language across the world. This trend, coupled with the multicultural nature of the community in Jakarta Selatan, further contributes to the phenomenon of code-mixing as an inherent byproduct of linguistic diversity and interactions. The South Jakarta Code-mixed phenomenon presents an avenue for NLP research to confront the intricacies of language and facilitate multilingual communication. Ongoing research endeavors in this field strive to cultivate innovative techniques adept at effectively managing the distinct challenges posed by code-mixed data. Such advancements seek to enhance comprehension and processing capabilities when dealing with multilingual text.

Code-mixing or language mixing in sentiment analysis tasks poses a unique challenge in the field of Natural Language Processing (NLP). Most research on code-mixing is conducted to address sentiment-related issues within monolingual contexts. Given that English is the second most widely used language globally, the likelihood of code-mixing between English and other languages is significant. As a result, the majority of code-mixing research focuses on combinations of English with other languages. However, publicly available code-mixed data between English and Indonesian is scarce, making it challenging to investigate and develop solutions in this area.

Several sentiment analysis studies have been conducted using code-mixed data represented in embeddings. Embeddings are used to represent words in such a way that words that are closer in vector space are expected to have similar meanings. Previous studies include Mishra et al. [2], who used Indian-English and Spanish-English with Glove and TF-IDF as embeddings; Javdan et al. [3], who used Indian-English and Bengali-English with Glove and FastText as embeddings; and Tho et al. [4], who used Indonesian-Javanese with a BERT pretrained model as embedding. Comparative studies conducted by Wang et al. [5] demonstrated that the BERT pretrained model outperforms classic embeddings that are context-independent, such as Glove and FastText. However, these three embeddings are primarily focused on a single language (English), which adversely affects sentiment classification performance.

While many research efforts have aimed to improve sentiment analysis by incorporating code-mixing data from various languages, it's clear that the models used in existing studies are predominantly centered on a single language, specifically, English. This preference for one language in model architecture, along with the neglect of others, can lead to less-than-optimal sentiment analysis results. Moreover, there is a significant scarcity of publicly available code-mixed datasets, especially those annotated with sentiment categories, particularly for the Indonesian-English language combination. This research gap constrains our understanding of the processes involved in code-mixing between the Indonesian and English languages within the field of Natural Language Processing (NLP), underscoring the necessity to consider both languages, rather than singularly focusing on one. In light of this void, this study makes the following major contributions:

- Constructing a code-mixed Indonesian-English dataset specifically designed for semantic tasks, such as sentiment analysis, and
- Developing five preprocessing scenarios based on pretrained models used in sentiment analysis tasks to address the limitations of embeddings that concentrate solely on one language.

This paper is structured as follows: Section II provides the literature review; Section III encompasses the development of the dataset, including data collection, labeling, the definition of sentiment, preprocessing, and the description of labeling outcomes. Section IV outlines the research methodology, Section V presents the results and discussion, Section VI expounds upon future research directions, and the concluding chapter wraps up the paper in Section VII.

II. Literature Review

Pretrained models in Natural Language Processing (NLP) have demonstrated their effectiveness in addressing code-mixed challenges by capitalizing on their capacity to learn from extensive and diverse text data. These pretrained models have evolved to become the state-of-the-art models for language comprehension and generation. The foundational data used to train these models is sourced from extensive corpora, including resources such as Wikipedia and book corpuses. Additionally, as outlined by Gupta, Ekbal, and Bhattacharyya [13], existing benchmark datasets across various NLP tasks can be adapted to the code-mixed environment, enabling the assessment of a model's adaptability within a multilingual framework.

Sentiment analysis, a well-established and extensively researched field within social media analysis and NLP, has primarily been conducted in monolingual settings to address sentiment-related concerns. In the realm of code-mixing research, the primary focus has been on combining English with other languages. This emphasis stems from the widespread use of English as the second most spoken language globally, resulting in the prevalence of code-mixing involving English.

In certain sentiment analysis studies, code-mixed data is represented using embeddings, which transform words into vectors, with words of similar meanings residing in close proximity within vector spaces. For example, a study conducted by Mishra, et al. [2] utilized Indian-English and Spanish-English code-mixed data, employing Glove and TF-IDF as embeddings. Similarly, Javdan, et al. [3] employed Indian-English and Bengali-English code-mixed data with Glove and FastText as embeddings, while Tho, et al [4] used Indonesian-Javanese code-mixed data with a Sentence BERT pre-trained model for embedding. In our current study, we employ IndoBERTweet, BERTweet, and Multilingual pretrained models to embed code-mixed data, including Indonesian-English pairs.

Notably, Mishra, et al. [2] conducted code-mixing research involving Indian local languages such as Hindi (HI) and Bengali (BN). The research encompassed various language pairs, including Hindi-English and Bengali-English, employing machine learning and neural networks to address challenges. Feature vectors like GloVe and TF-IDF with 2-6 n-gram characters were employed. However, the study identified several limitations, including the model's difficulty in accurately capturing sentiment polarization from ambiguous data, the limitations of word n-grams in precisely representing sentiment in sentences due to Twitter's character restrictions, and the existence of diverse spelling variations for a single word. Consequently, the use of multilingual embeddings is deemed necessary to accurately classify sentiment and represent words and phrases beyond the lexicon of a single language.

Javdan, et al. [3] conducted sentiment analysis on code-mixed Hindi-English and Spanish-English data using fastText and GloVe embeddings. However, the combination of CNN, fastText, and GloVe resulted in unexpected outcomes with subpar performance. This underperformance was attributed to the exclusion of languages other than English in the embedding process, regardless of the model used. As a result, future studies will prioritize pretraining models in both languages, employing attention mechanisms to determine the significance of each linguistic representation within each phrase.

Furthermore, Tho, et al. [4] evaluated sentiment analysis for Indonesian and Javanese code-mixed data using lexicons and transformers, incorporating an attention mechanism. The transformer model, which leverages the attention mechanism, was employed. Based on the results, the transformer model with BERT encoding demonstrated superior accuracy compared to lexicons. Additionally, a comparative analysis conducted by Wang, et al. [5] revealed that the BERT pretrained model outperforms well-known context-independent embeddings like Glove and FastText. This finding prompted the selection of the BERT pretrained model as the starting point for our current research.

Despite numerous research endeavors that have attempted to enhance sentiment analysis performance using data incorporating code-mixing across different languages, it is evident that the dataset and model architectures employed in existing studies are primarily fixated on a single language, specifically, English. The limitations associated with embedding models trained or biased toward a single language while neglecting others can result in suboptimal sentiment analysis performance. Furthermore, code-mixed datasets, particularly for semantic tasks like sentiment analysis, annotated with sentiment categories, remain notably scarce in the public domain, especially for the Indonesian-English language combination.

This research addresses critical gaps in NLP by focusing on code-mixing challenges involving English and Indonesian, offering a new dataset for sentiment analysis. It also explores preprocessing strategies to enhance sentiment analysis in code-mixed data. The study contributes to the understanding and management of code-mixed text and aligns with the broader goal of improving multilingual communication and NLP.

III. Dataset Development

A. Data Collection Procedure

The data collection methodology entails acquiring data through the submission of requests to Twitter API version 2, amalgamating data in both Indonesian and English languages. A specific authorization, denoted as "Academic research", is imperative for gaining access to Twitter API version 2. The default permission level of Twitter API, referred to as "Essential," is considerably limited and insufficient for conducting crawling operations. Registration is obligatory and necessitates the completion of various inquiries pertaining to the research project. Upon successful registration, an API key is procured. Academic access is indispensable, as it confers numerous advantages that facilitate the research process, including access to real-time public Twitter data and other features that augment accurate and equitable data collection.

A roster of desired keywords is employed as the query or filter for the endpoint, which combs through the entirety of Twitter. The keyword list is amalgamated using the OR operator, signifying a combination of keywords. Additional filters encompass "-is:retweet" to exclude retweeted or shared tweets to avert duplication, "-is:reply" to consider solely original tweets from the tweet author, and "place country:ID" to specify tweets from users located in Indonesia. The selected search terms are derived from sociolinguistic phenomena in Indonesian and English, which are popular among young individuals in the South Jakarta region [9], along with a compilation of pertinent hashtags. The keywords for the endpoint query were gathered from various publications found through Google searches utilizing the term "slang jaksel." The number of keywords obtained from these articles is restricted to 50 and stored in the software as an array. To prepare the training data, the Twitter API v2 and the Python module Tweepy are utilized for the purpose of retrieving code-mixed Indonesian and English data.

The available data spans from August 2020 to September 2022. Along with keywords, the endpoint query also includes date-based queries. The dataset is standardized by dividing it into three sections: testing, validation, and training. The evaluation and dataset distribution adhere to the same F1 value calculation as applied to the IndoLEM dataset in a manner similar to the approach outlined in a study conducted by Koto et al. [1]. The data distribution in this study employs a ratio of 3638 sentences for training, 399 for validation, and 1011 for testing. Considering that the maximum amount of data retrievable per query is 500, queries must be executed in batches.

To mitigate noise within the data, a human data cleaning procedure is executed. This procedure encompasses the removal of duplicate data, the filtration of monolingual data, and the exclusion of offensive or inappropriate terms. After the selection process, the data is annotated with emotions, specifically categorized as neutral, positive, and negative. These emotions are subsequently translated into numerical values during the training phase, with the assigned values as follows: neutral = 0, negative = 1, and positive = 2.

B. Data Labeling Procedure

The data labeling procedure involved the collaboration of five annotators with expertise in the fields of data science, English literature, and Indonesian literature. Crowdsourcing was employed to streamline the annotation process, and the services of these annotators were procured through a dedicated freelancing platform tailored specifically for Indonesian professionals (https://projects.co.id). The selection of annotators was based on their knowledge and proficiency in areas directly pertinent to the characteristics of the dataset under investigation.

The tasks assigned to the five annotators were delineated as follows:

1) *Reviewing* tweets provided via the Google AppSheet platform.

2) *Categorizing* the tweets into three distinct sentiment labels: positive, negative, and neutral.

3) *Analyzing* the sentiment groupings based on their pre-existing linguistic expertise.

In order to uphold consistency and precision throughout the annotation process, a majority vote was conducted among the three sets of labeled files. The AppSheet system was seamlessly integrated with Google Sheets, serving as the foundational database from which the annotators read and annotated the tweets. To ensure a more coherent and standardized set of annotated tweets, the final determinations regarding precise sentiment annotations were reached through mutual consensus among the annotators. To enhance comprehension, the consensus on sentiment assessment took into account the context of polarization or sentiment inclination, as well as specific instances. The sentiment categories and their corresponding degrees of polarity are delineated in Table I.

TABLE I. Sentiment Definition based on Polarity Level

Polarity Level	Definition
Positive	A response or outlook that raises the worth of someone or something.
Negative	Negative language that will reduce the perceived value of whatever is being viewed.
Neutral	Taking no sides

C. Dataset Generation Results

According to the labeling distribution determined by the five annotators, the findings revealed that negative labels were more prevalent on Twitter, constituting approximately 35.33% of the data instances, as opposed to neutral labels at approximately 36.19% and positive labels at approximately 28.48%. This distribution was observed in a total of 1,903 out of 5,068 data instances. The distribution results are comprehensively presented in Table II.

TABLE II. FIVE ANNOTATORS' RESULTS OF LABELING DISTRIBUTION

	Positive	Negative	Netral	Total
Annotator 1	1732 (≈33.98%)	1213 (≈23.76%)	2123 (≈42.26%)	5068 (100%)
Annotator 2	1421 (≈27.95%)	2185 (≈42.84%)	1462 (≈29.21%)	
Annotator 3	1039 (≈20.49%)	2458 (≈48.51%)	1571 (≈31.00%)	
Annotator 4	1714 (≈33.60%)	1927 (≈37.59%)	1427 (≈28.80%)	
Annotator 5	1784 (≈35.66%)	1732 (≈33.92%)	1552 (≈30.42%)	
Average	1538 (≈28.48%)	1903 (≈35.33%)	1627 (≈36.19%)	

Description:

- Annotator 1 possesses expertise in the field of Statistics.
- Annotator 2 specializes in Indonesian Literature.
- Annotator 3 has expertise in English Literature.
- Annotator 4 is knowledgeable in the field of English Education.
- Annotator 5 has expertise in English Literature.

D. Preprocessing

This study will employ various preprocessing methods, encompassing emoji/emoticon conversion, translation, and slang word normalization. Manual data cleaning will be undertaken to eliminate offensive remarks, obscenities, and duplicate entries. Furthermore, automatic preprocessing will be executed by removing unnecessary tokens based on a predefined regular expression. This includes the elimination of acronyms, hashtags, mentions, and URLs. The preprocessing procedure may also encompass converting text to lowercase, addressing coarse language in English, converting emoji/emoticon symbols, normalizing text, performing translation tasks, and eliminating extraneous words. The specific preprocessing techniques employed will be contingent on the selected scenario and the pretrained model in use.

IV. RESEARCH METHODOLOGY

In a general overview, the research involves five distinct phases for constructing the model. These phases encompass the data collection, data labeling, sentiment polarity determination, training and validation, and the subsequent testing and evaluation processes. The process of dataset development, which includes data collection, labeling, sentiment definition, preprocessing, and the acquisition of labeling results, is elucidated in Section III, specifically under the dataset development section. Fig. 1 depicts the research's overall diagram.

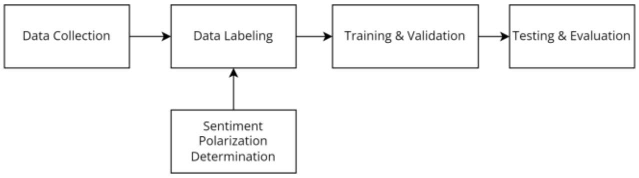

Fig. 1. The research's stages diagram.

A. Training, Testing and Validation Process

In the training process, data preparation is carried out, involving data cleaning and preprocessing. Following preprocessing, training is performed using the BERT model. The training process is conducted with three epochs, in accordance with the recommendation by Devlin et al. [7].

Validation occurs concurrently with training to find the lowest loss value among the three epochs. By analyzing the training loss and validation loss values, it can be determined whether the dataset is overfit, underfit, or properly fit to the pretrained model in use. Consequently, the output of the training process is the best model based on the validation results. Given that the primary focus of this research is to compare the use of pretrained models, the selection of epoch, batch, and learning rate values aligns with the fine-tuning procedure outlined by Devlin et al. [7]. The specific values used are adjusted for each pretrained model since they have different optimal hyperparameters. From the chosen best model, a testing process is conducted. The test data is preprocessed using the same procedure as in the training process. The testing process involves comparing the prediction results with the labels in the test data.

B. Bidirectional Encoder Representations from Transformer (BERT)

Bidirectional Encoder Representations from Transformers (BERT) is a deep learning model designed to generate comprehensive representations of unlabeled text by considering both left and right context at all levels. BERT accomplishes this by incorporating the entire set of words in the surrounding text. Pretrained models based on BERT have proven to be highly effective in representing language due to their extensive training on large corpora and specific domains. The BERT framework consists of two main stages: pretraining and fine-tuning. During pretraining, the model is trained on unlabeled data using various tasks, such as Masked Language Modeling (LM) and Next Sentence Prediction. In the fine-tuning stage, the BERT model is initialized with pretrained parameters and further fine-tuned using labeled data specific to particular tasks.

In this research, the pretrained BERT model is employed to tackle the semantic task of sentiment analysis. The utilization of pretrained models offers significant advantages in natural language processing (NLP) tasks, as it eliminates the need to train new models from scratch. The BERT-based pretrained models utilized in this study include BERTweet, trained on a collection of English tweets [8], IndoBERTweet [6], trained on a collection of Indonesian tweets, and MultilingualBERT [7], trained on a multilingual Wikipedia dataset comprising 104 languages. These pretrained models serve as a foundational point for the sentiment analysis task, leveraging their pre-established knowledge and language representations.

C. Sentiment Classification

In this research, the classification process leverages the encoding architecture of the transformer. The experiments conducted in this research are based on the implementation of the BERT model provided by Hugging Face. In its entirety, the layers are divided into three parts: the input layer, the attention layer, and the classification layer. A pretrained model, specifically BERTBase, is utilized as a reference in the training process, with adjustments made for each domain. The parameters of BERTBase used in this study have a larger configuration compared to the transformer parameters defined in previous research Vaswani et al. [19]. BERT incorporates a larger feedforward network, with 768 and 1024 hidden units/embedding sizes, as well as more attention, with 12 and 16 heads, in contrast to the default configuration in the initial Transformer paper (six encoder layers, 512 hidden layers, and eight head attention). The pooler layer in the transformer architecture is employed as input to calculate probabilities and loss values using sigmoid and BCELoss calculations. These computations are conducted in alignment with the specialized BERT implementation for semantic tasks like sentiment analysis. The sigmoid calculation produces logits, which are used to determine whether the sentiment class falls into the categories of neutral, positive, or negative. Meanwhile, the loss value is utilized for validation to select the best model and evaluate whether the trained data exhibits signs of overfitting, underfitting, or a good fit.

D. Experimental Setup

The dataset employed in this study comprises social media activity data in the form of tweets from Twitter. This dataset encompasses both textual data and emoticons/emojis, as prior research has demonstrated that the inclusion of these types of data enhances the performance of the sentiment analysis pipeline, which is based on BERT [11] and deep learning-based sentiment analysis [10].

TABLE III. FIVE ALTERNATIVE SCENARIOS

	Step Processes
Scenario 1	1. Data cleaning 2. Preprocessing 3. Emoji conversion 4. Translation to Bahasa Indonesia 5. BERT Training with IndoBERTweet pretrained model 6. Validation 7. Testing and Evaluation
Scenario 2	1. Data cleaning 2. Preprocessing 3. Emoji conversion 4. Translation to English 5. BERT Training with IndoBERTweet pretrained model 6. Validation 7. Testing and Evaluation
Scenario 3	1. Data cleaning 2. Normalization to Bahasa Indonesia 3. Training with MultilingualBERT pretrained model 4. Validation 5. Testing and Evaluation
Scenario 4	1. Training with IndoBERTweet pretrained model 2. Validation 3. Testing and Evaluation
Scenario 5	1. Training with MultilingualBERTweet pretrained model 2. Validation 3. Testing and Evaluation

A list of 50 keywords is compiled from various online articles, and some of the instances, as illustrated in Table II, are subjected to translation using Google Translate after undergoing preprocessing. The objective is to amass a diverse dataset encompassing data in multiple languages. Additionally, emoji and emoticons are transformed into plain text using the emoji 1.7.0 library incorporated in the Python package [10]. The sentiment analysis procedure is executed through five distinct scenarios, as outlined in Table III.

E. Evaluation Process

The evaluation phase is conducted based on the five experimental scenarios that have been established. All five scenarios are tested and evaluated using the same methodology. In the testing phase, labeled test data is utilized to make predictions. The predicted data is then compared to the actual label data, resulting in the creation of a confusion matrix. This confusion matrix is instrumental in determining the count of correct and incorrect predictions based on the respective classes. Metrics calculated include accuracy, precision, recall, and F-measure. Given that this research involves three classes: neutral (0), negative (1), and positive (2), the resulting confusion matrix has a 3x3 dimension, with columns representing predicted sentiment and rows representing the true sentiment. The confusion matrix provides the basis for calculating evaluation metrics, such as precision, recall, F1-score, and overall accuracy for all classes. To compute the values for each metric, one requires the count of true positives, true negatives, false positives, and false negatives.

V. RESULTS AND DISCUSSION

Leveraging pretrained BERT models has demonstrated notable effectiveness when fine-tuning the sentiment analysis task. These pretrained models provide substantial time and cost savings, as they have already undergone extensive training on large corpora, a process demanding significant computational resources. However, it is crucial to acknowledge that not all pretrained models are equally well-suited for the unique domain of code-mixed data. The comprehensive results of the experiments carried out in this study are presented in Table IV.

TABLE IV. FIVE ALTERNATIVE SCENARIOS

Scenarios	Precision	Recall	F1-Score	Accuracy
Scenario 1 IndoBERTweet (*Preprocessing)*	0.7607	0.7552	0.7551	0.7656
Scenario 2 BERTweet (*Preprocessing*)	0.7314	0.7322	0.7316	0.7389
Scenario 3 MultilingualBERT (*Preprocessing)*	0.6677	0.6622	0.6631	0.6716
Scenario 4 IndoBERTweet (Without *preprocessing)*	0.7396	0.7374	0.7358	0.7478
Scenario 5 MultilingualBERT (Without *preprocessing)*	0.6334	0.6345	0.6335	0.6370

Based on insights gained from this research, it is evident that code-mixing between English and other languages can be more optimal and effective when focused on the respective language rather than solely on English. These findings challenge previous studies conducted by Mishra et al. [2], Javdan et al. [3], and Tho et al. [4]. The utilization of the relevant language, in this case, supported by pretraining with a

Bahasa Indonesia-based BERT, adds value to enhancing the performance of a sentiment analysis task.

The experiment findings reveal that Scenario 1, which involves utilizing the pre-trained IndoBERT model along with translation into Indonesian, preprocessing steps such as emoji conversion to plain text, and data cleaning, yields the most favorable results for addressing the Indonesian-English code-mixed problem. In this scenario, an average precision of 76.07%, recall of 75.52%, an f-1 score of 75.51%, and an accuracy of 76.56% were achieved. It's worth noting that the performance, albeit slightly, decreased compared to Scenario 4, which employed the pre-trained IndoBERT model without preprocessing, resulting in an average precision of 73.96%, recall of 73.74%, an f-1 score of 73.58%, and an accuracy of 74.78%. These outcomes underscore the effectiveness of the monolingual Bahasa Indonesia translation and preparation procedure in enhancing performance.

Conversely, the pre-trained Multilingual BERT model exhibited suboptimal performance in Scenarios 3 and 5. However, the inclusion of preprocessing steps led to improved performance outcomes. This phenomenon could be attributed to the fact that the Multilingual BERT training process relies on the Wikipedia corpus, which forms its vocabulary using standard words. Consequently, it is less adept at capturing out-of-vocabulary terms in code-mixed data, including informal words. Additionally, after three epochs of training, the data for the BERTweet pre-trained model showed a good fit, but IndoBERT outperformed it in terms of performance. This may be due to the prevalence of code-mixing and its adherence to Indonesian sentence structure, rendering the translation accuracy of Google Translate less reliable. Based on the aforementioned justifications, here are the outcomes of each of the five scenarios:

- Scenario 1 (pre-trained IndoBERT with preprocessing):
- Average precision: 76.07%, recall: 75.52%, f-1 score: 75.51%, accuracy: 76.56%.
- Scenario 2 (pre-trained Multilingual BERT with preprocessing): Lower performance than Scenario 1.
- Scenario 3 (BERTweet pre-trained model): Exhibited a good fit after three epochs of training but was outperformed by IndoBERT in terms of performance.
- Scenario 4 (pre-trained IndoBERT without preprocessing): Slightly lower performance than Scenario 1.
- Scenario 5 (pre-trained Multilingual BERT without preprocessing): Lower performance than Scenario 1 but improved with preprocessing steps.

The deep learning approach that automatically generates code-mixed text from English into various languages without previous parallel data, as researched by Gupta [13], could be an option for creating a dataset. However, the perspective and role of annotators in determining how sentiment can be defined are crucial in understanding the context of a sentence to be analyzed in the labeling process. This is due to the more complex language composition compared to single language text. Leveraging their expertise, annotators can provide insights into determining sentiment through expressions, emojis/emoticons, context, and emotions. Thus, it is expected that the dataset can contribute to improving performance in various NLP tasks, regardless of the models and scenarios employed.

VI. Future Research Direction

Future research is expected to address several key challenges and explore new techniques to enhance the effectiveness and efficiency in handling code-mixed data, particularly the mixing of Indonesian and English in various NLP tasks. The following are some directions for future research:

- Future research efforts may involve enriching the lexicon of non-standard language, both in English and Indonesian, to minimize out-of-vocabulary words. Additionally, exploring the conversion of emojis and emoticons within multilingual BERT scenarios (given its training on standard corpora like Wikipedia) to make it applicable to Twitter sentences.
- As preprocessing has a substantial impact on this study, further investigation is needed to develop better and more optimal preprocessing methods for effectively handling code-mixed data.
- Research in the future could consider creating pretrained models on code-mixed data corpora without retaining the existing scenarios. This approach, while taking into account each language to avoid ambiguity, would allow sentiment analysis processes to run without the need for tuning existing monolingual pretrained models.

VII. Conclusion

This research project has successfully curated a code-mixed dataset by amalgamating Indonesian and English languages, utilizing 50 keywords derived from a sociolinguistic phenomenon observed among teenagers in South Jakarta. The presence of code-mixing, which results from the combination of two distinct languages, underscores the significance of involving language experts in determining sentiment polarity. This engagement is essential for enhancing the accuracy of sentiment analysis, given the heightened linguistic complexity compared to single-language text. By capitalizing on their expertise, annotators can offer valuable insights for ascertaining sentiment based on expressions, emojis/emoticons, context, and emotional cues. Consequently, the dataset is anticipated to make a meaningful contribution to enhancing performance across a range of NLP tasks, irrespective of the specific models and scenarios used. The dataset was meticulously designed to tackle the semantic task of sentiment analysis, incorporating three distinct label categories: positive, negative, and neutral. The annotation of the dataset was carried out by a panel of five annotators, each possessing expertise language and data science. The dataset has been made publicly available on the Github repository and can be accessed via the following link: https://github.com/laksmitawidya/indonglish-dataset.

Despite being initially trained on a monolingual corpus, the BERT pretrained model exhibits impressive performance when handling code-mixed data in sentiment analysis tasks. Furthermore, each pretrained model demands only a brief training period of approximately four hours on a CPU. This underscores the significant role of pretrained models in expediting the training process in specialized domains, such as sentiment analysis of Twitter data containing both Indonesian and English codes, while keeping computational resource requirements minimal.

Leveraging pretrained BERT models has proven effective in fine-tuning sentiment analysis tasks, resulting in substantial time and cost savings due to their extensive training on large corpora. It's worth emphasizing that not all pretrained models are equally suitable for code-mixed data, and the application of scenario-based preprocessing significantly impacts performance outcomes. The findings of this research highlight the effectiveness of concentrating on the respective language in code-mixing between English and other languages, as opposed to a sole emphasis on English. These insights challenge the conclusions drawn in prior studies. Notably, incorporating the relevant language, particularly when supported by pretraining with a Bahasa Indonesia-based BERT, emerges as a valuable approach to significantly improve the performance of sentiment analysis tasks. Specifically, Scenario 1, which combines the pre-trained IndoBERT model with preprocessing, emoji conversion, and data cleaning, delivers highly favorable results for Indonesian-English code-mixed sentiment analysis, achieving notable average precision, recall, f1-score, and accuracy values of 76.07%, 75.52%, 75.51%, and 76.56%, respectively. This emphasizes that data cleansing, translation to a unified language, and normalization during the preparation stage are adequate for enhancing performance. On the other hand, the Multilingual BERT model initially exhibits suboptimal performance in specific scenarios but shows improvement with the introduction of preprocessing steps, revealing its limitations in capturing out-of-vocabulary terms in code-mixed data. Looking ahead, future research should prioritize enriching non-standard language lexicons, optimizing preprocessing techniques, and exploring pretrained models for code-mixed data without relying on specific scenarios. These endeavors aim to enhance sentiment analysis performance in multilingual contexts, ultimately contributing to improved accuracy and efficiency.

ACKNOWLEDGMENT

This work was supported by FMIPA UGM [grant number: 42/UN1/FMIPA.1.3/KP/PT.01.03/2023].

REFERENCES

[1] Koto, F., Rahimi, A., Lau, J. H., and Baldwin, T., “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP”, Proceedings of the 28th International Conference on Computational Linguistics, pp. 757–770, 2020.

[2] Mishra, P., Danda, P., and Dhakras, P, “Code-Mixed Sentiment Analysis Using Machine Learning and Neural Network Approaches”, arXiv:1808.03299. 2018 [Online]. Available: https://arxiv.org/abs/1808.03299

[3] Javdan, S., ataei, T. S., and Minaei-Bidgoli, B, “IUST at SemEval-2020 Task 9: Sentiment Analysis for Code-Mixed Social Media Text using Deep Neural Networks and Linear Baselines”, Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1270–1275, Dec. 2020 [Online]. Available: https://aclanthology.org/2020.semeval-1.170

[4] Tho, C., Heryadi, Y., Kartowisastro, I. H., and Budiharto, W., “A Comparison of Lexicon-based and Transformer based Sentiment Analysis on Code-mixed of Low-Resource Languages”, International Conference on Computer Science and Artificial Intelligence (ICCSAI), pp. 81-85, 2021 [Online]. Available: https://ieeexplore.ieee.org/document/9609781.

[5] Wang, C., Nulty, P., and Lillis, D, “A Comparative Study on Word Embeddings in Deep Learning for Text Classification., Natural Language Processing and Information Retrieval”, pp. 18-20, Oct. 2020.

[6] Koto, F., Lau, J. H., and Baldwin, T, “IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization”, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10660–10668, 2021 [Online]. Available: https://aclanthology.org/2021.emnlp-main.833.

[7] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Google AI Language”, 2019 [Online]. Available: https://aclanthology.org/N19-1423.pdf

[8] Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). “BERTweet: A pre-trained language model for English Tweets”. Proceedings of the 2020 EMNLP (Systems Demonstrations), 9–14.

[9] Wijaya, A. D., and Bram, B., “A Sociolinguistic Analysis of Indoglish Phenomenon in South Jakarta”, Professional Journal of English Education, pp. 672-684, 2021.

[10] Ullah, M. A., Marium, S. M., Begum, S. A., and Dipa, N. S., “An algorithm and method for sentiment analysis using the text and emoticon”, ICT Express 6, pp. 357–360, 2020.

[11] Pota, M., Ventura, M., Catelli, R., and Esposito, M., “An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian”, Sensors, pp. 1-21, 2021.

[12] Anastassiou, F., Andreou, G., "Factors Associated with the Code Mixing and Code Switching of Multilingual Children: An Overview", International Journal of Linguistics, Literature and Culture (LLC) , 13-26. 2017.

[13] Gupta, D., Ekbal, A., & Bhattacharyya, P, "A Semi-supervised Approach to Generate the Code-Mixed Text using Pre-trained Encoder and Transfer Learning", Association for Computational Linguistics, 2267–2280, 2017.