DLSU Research Congress 2025
De La Salle University, Manila, Philippines
June 26 to 28, 2025

Exploring Transformer-based Approaches in Sentiment Prediction of Philippine Tweets

Marcus Calalang, Miguel Estañol, Jon Jacinto, Mauries Lopez, Eric Denver Co, John Matthew Gan, Jose Noel Noblefranca, Jason Jan Jabanes, and Edward Tighe*
Department of Software Technology and Center for Language Technologies, De La Salle University - Manila
**Corresponding Author: edward.tighe@dlsu.edu.ph*

Abstract: Sentiment analysis is a valuable tool for understanding public opinions and trends, particularly given the rapid growth of social media in the Philippines. However, several challenges must be addressed to further advancements in this field. Previous research in Philippine sentiment analysis has largely depended on traditional machine-learning approaches and manual annotation processes, resulting in smaller, less representative datasets. This study explores an emoji-based automatic annotation combined with transformer models to enhance sentiment analysis in Philippine social media. The Multidimensional Lexicon of Emojis (MLE) by Godard and Holtzman (2022) is used to leverage the sentiment scores of emojis. We used two private tweet collections from X (formerly Twitter), comprising over 17 million tweets. After data preprocessing, we labeled sentiment using MLE sentiment scores, with human annotators verifying a sample of 2,754 tweets. Spearman's rank correlation was used between MLE scores and human-annotated sentiment scores, obtaining a correlation of 0.4747 for positive and 0.6181 for negative sentiment, indicating a moderate to strong agreement. The study assessed various BERT-based transformer models, including mBERT, Tagalog-BERT, RoBERTa-Tagalog, and TwHIN-BERT, through single-output and multi-output regression configurations for sentiment analysis. The models' performance was evaluated using RMSE and R^2 metrics. The results reveal that TwHIN-BERT scored 0.1596 RMSE and 0.3018 R^2 for both single and multi-output configurations, outperforming other models, with RoBERTa-Tagalog consistently placing second in both configurations.

Key Words: Sentiment Analysis, Philippine Text Data, Emoji Lexicon, Transformer Models, Automatic Annotation

1. INTRODUCTION

Sentiment analysis is a Natural Language Processing task that identifies and classifies sentiments expressed in text. This makes it an important tool that enables organizations to gauge public opinion and trends on topics from consumer products to politics (Jurafsky & Martin, 2024). However, the increasing volume and complexity of communication on these platforms highlights several challenges, particularly in the Philippine context.

One significant challenge is the reliance on traditional machine learning models. Studies by Alcober and Revano (2021) and Delizo et al. (2020) show that these methods often struggle with the subtleties and nuances of social media texts. Taboy (2023) further demonstrated that traditional models yield limited accuracy in mixed-language environments. The limitations of these methods are further underscored by the rapid evolution of language and the inherent noise present within this communication medium.

Another major obstacle is the scarcity of large, publicly available annotated corpora. Manual annotation is a time-consuming, labor-intensive task and proves difficult to scale on increasingly larger datasets. For instance, Taboy (2023) manually labeled only 1.8k of 63.6k tweets, while Imperial et al. (2019) and Maceda et al. (2022) also reported significant reductions in dataset sizes due to annotation challenges.

Our work addresses these challenges by combining automatic data annotation with deep learning. We use an emoji-based annotation scheme that

leverages emojis as sentiment indicators for social media text data found on X (formerly Twitter). This method builds on the work of E. Co et al. (2023) and S. Co et al. (2022), whose studies showed that using sentiment scores from emoji lexicons allows emojis to serve as proxies for sentiment labels. Specifically, our scheme leverages sentiment scores from the Multidimensional Lexicon of Emojis (MLE) by Godard and Holtzman (2022).

In addition, our study employs transformer models for sentiment prediction. Transformers have been recognized for capturing long-range dependencies and contextual relationships in text. Specifically, BERT, an encoder-based transformer model, and its derivatives have been adapted across various domains and languages and have become a ubiquitous baseline in various NLP tasks (Rogers et al., 2021). Recent local studies indicate that fine-tuned BERT-based models on domain-specific text data achieve high accuracy and F1 scores in sentiment classification (Cosme & De Leon, 2024; Maceda et al., 2023). We aim to improve performance on a broader coverage of Philippine tweets by fine-tuning these models on our large-scale, automatically annotated dataset.

The findings of this study aim to enhance the field of sentiment analysis by advancing the use of transformer models and employing emoji-based automatic annotation for Philippine social media text data. This research provides valuable insights into the adaptability of BERT-based transformer models while addressing the current shortage of robust, locally tailored sentiment prediction tools, thereby establishing a foundation for future research and the development of scalable, efficient NLP pipelines in the Philippine context.

2. METHODOLOGY

This study employs a comprehensive methodological pipeline to explore transformer-based approaches in sentiment prediction of Philippine text data.

2.1 Data Source and Preparation

The datasets used in this study were sourced from two private collections of social media posts from X (formerly Twitter). The first dataset, collected by S. Co et al. (2022), consists of 10,214,176 tweets, while the second dataset, collected by E. Co et al. (2023), contains 6,966,696 tweets, for a total of 17,180,872 tweets. Both datasets consist of tweets made in the Philippines obtained via the Twitter API. All tweets were queried using a geographical bounding box[1] of the Philippines to ensure the tweets were made within the country – maximizing the chance to capture how Filipinos communicate within the social media platform.

To provide additional insight on how Filipinos tweet, we report the distribution of language tags provided by X in Table 1. The distribution shows that the data is predominantly in Tagalog[2] and English, although manual inspection of the data shows a large volume of tweets with a mix of Tagalog and English words (i.e., code-switching). We also note that the Indonesian and Spanish tags come in as the 3rd and 4th most frequent language tags; however, we attribute this to misclassification on the end of X due to overlapping vocabulary between Tagalog, Indonesian, and Spanish. Additionally, tweets labeled as Undefined can be characterized as being short in length and/or primarily containing user mentions, links, emojis, or non-language word entities, like laughter (e.g., hahaha).

Table 1. The distribution of language tags provided by X of the raw dataset (n=17,180,872).

Language	Tweet Count	Percentage(%)
Tagalog	9,277,800	54.00%
English	5,049,291	29.39%
Indonesian	479,601	2.79%
Spanish	145,027	0.84%
Undefined	1,174,029	6.83%
Others	1,055,124	6.15%

We then preprocess the raw tweets by first removing all blank entries and duplicates, resulting in 15,222,923 tweets. Removal of duplicates was done to ensure that the dataset did not overrepresent certain combinations of words. The dataset was then filtered to retain only tweets containing at least one MLE emoji. This resulted in a final set of 5,362,125 tweets. Tweets that were discarded contained only links, unrecognized emojis, text without MLE emojis, mentions, or combinations of these characteristics (e.g., text without MLE emojis and unrecognized emojis).

[1] PH bound box defined as [7.17427453, 5.58100332277, 126.537423944, 18.505227362]

[2] Language tags provided by Twitter do not distinguish between Tagalog, Filipino, or any other Philippine languages.

Lastly, the entire preprocessed dataset was split into train and test sets using a 90-10 split. The train set was then split into train and validation, also following a 90-10 split. The entire splitting process is illustrated in Figure 1.

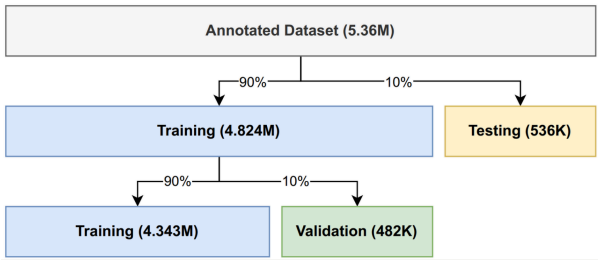

Figure 1. Visualization of how the original 5.36 million synthetically annotated tweets were divided into train, test, and validation sets.

2.2 Automated/Synthetic Sentiment Annotation

As an alternative to manually annotating, we adopted the synthetic sentiment annotation scheme proposed by E. Co et al. (2023) to capitalize on the volume of our collected data. The simple scheme utilizes the Multidimensional Lexicon of Emojis (MLE) (Godard and Holtzman, 2022), which represents positive and negative sentiments as separate dimensions. The full algorithm of E. Co et al. (2023) is shown in Figure 2.

Input: Dataset of tweets, MLE (Multidimensional Lexicon of Emojis)
Output: Positive and negative sentiment scores assigned to each tweet

for each tweet in dataset **do**
  initialize *positive_sum*, *negative_sum*, *emoji_count* ← 0
  extract *emojis_in_tweet* from tweet text
  if *emojis_in_tweet* has MLE emoji(s) **then**
    for each emoji in *emojis_in_tweet* **do**
      positive_sum+= MLE[emoji].positive
      negative_sum+= MLE[emoji].negative
      emoji_count+= 1
    end for
    positive_score ← *positive_sum*/*emoji_count*
    negative_score ← *negative_sum*/*emoji_count*
    assign *positive_score* and *negative_score* to tweet
  else
    discard tweet
  end if
end for

Figure 2. Automatic annotation scheme with MLE Sentiment Scores by E. Co et al. (2023).

The algorithm for synthetic sentiment annotation first takes an input document and extracts all MLE emojis present. For each emoji, the corresponding positive and negative sentiment scores from the MLE Lexicon are extracted and then averaged across each of the sentiment dimensions. We show an example of this process in Figure 3.

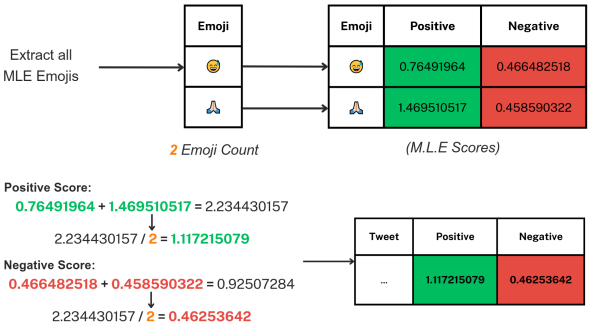

Figure 3. An example of automated sentiment annotation using emoji sentiment scores from the Multidimensional Lexicon of Emojis (MLE) (Godard and Holtzman, 2022).

2.3 Manual Annotation

To provide insight on how humans view sentiment in relation to our synthetic annotation scheme, a random sample of 3,000 tweets was taken from the final test set. This smaller set was manually annotated and the human annotations were then correlated with the synthetic labels. The 3,000 tweets were split into six sets of 500 tweets, and each set was split further into five parts to lessen the burden on manual annotators. The six sets were randomly distributed to 55 annotators participating in the manual annotation, 19 of whom completed their respective sets.

The manual annotation process required annotators to label 5 Google Forms with 100 tweets each. There were four questions for each tweet. The first three questions allowed annotators to judge the sentiment of a given tweet by each sentiment polarity (positive, neutral, and negative) using a 4-point Likert scale indicating the presence of the corresponding polarity. The last question is provided to allow the annotator to provide insight in case they cannot provide a sentiment.

2.4 Text Preprocessing

The synthetically annotated sets underwent further text preprocessing to ensure consistency and standardization across different transformer models. This preprocessing included dropping irrelevant columns, reducing excess instances, collapsing tokens, removing MLE emojis, and eliminating duplicates, which reduced the number of tweets to 4,636,366.

Multiple whitespaces, newline characters, and tab spaces were present in 799,540 tweets. Based on our observations, these elements were used for stylistic purposes to separate ideas and enhance readability. These elements were reduced to a single instance in each tweet to minimize noise while preserving their function. Additionally, other elements that contributed to noise, such as mentions, links, and hashtags, which do not provide relevant information for sentiment analysis, were collapsed into <USERHANDLE>, <URL>, and <HASHTAG> tokens, respectively.

It was observed that MLE emojis are represented differently by the transformers' tokenizer algorithm, as shown in Table 2. It was decided to remove all MLE emojis to ensure consistency and standardization, as they are not consistently represented across the used tokenizers.

Table 2. A comparison of how different transformer tokenizers represent emojis.

Tokenizer	Used by	Emoji Representation	Token Limit
BERT Tokenizer (based on WordPiece)	mBERT, Tagalog-BERT	[UNK]	512
RoBERTa Tokenizer (based on Byte Pair Encoding)	RoBERTa-Tagalog	Special Character	512
XLM-RoBERTa Tokenizer (based on SentencePiece)	TwHIN-BERT	Can represent emojis	512

Due to the token collapsing and the removal of MLE emojis, additional duplicate tweets appeared. An example of this can be seen in Table 3. These tweets previously contained unique tokens and MLE emojis, and when these tokens were collapsed and MLE emojis were removed, duplicates resulted. Duplicates can significantly affect the results by overrepresenting specific ranges of positive and negative sentiment scores. To prevent this, duplicates were dropped.

Table 3. Sample duplicate tweets found in the dataset after token collapsing and removal of MLE emojis.

Text	Positive Score	Negative Score
<URL>	0.825446	0.464058
<URL>	0.631303	0.644841
<URL>	0.900842	0.333164

Furthermore, Table 2 shows the token limits of each tokenizer. Since all BERT-based transformer models have a maximum token limit of 512, this value was used as the standardized limit across all tokenizers in this study. Additionally, tokenizer parameters such as truncation and padding were enabled. It was observed that 63 tokenized tweets exceeded the token limit: 62 tweets from RoBERTa-Tagalog and 1 from TwHIN-BERT, which is significantly low compared to the final set of tweets. To allow the tokenizer to process the data effectively, all tokenized tweets must have the same length, which is achieved through truncation and padding.

2.5 Regression Sentiment Prediction Model

After tokenization, input tokens are converted into embedding vectors and passed through transformer layers to generate contextualized representations. The final hidden state of the classification (CLS) tokens was then extracted and fed into the sentiment regression model for inference. Sentiment scores generated should align with the annotated scores derived from the MLE without additional transformations or scaling. Figure 4 illustrates our sentiment regression model.

The study explored two configurations for sentiment regression: Single-output and Multi-output. The single-output configuration trains two separate sentiment regression models, one for each polarity. Conversely, multi-output uses a single regression model to predict positive and negative scores simultaneously. Since training a model with these different configurations may yield varying performances, comparing each transformer model under both configurations is essential for evaluating their effectiveness. In both cases, we use non-fine-tuned

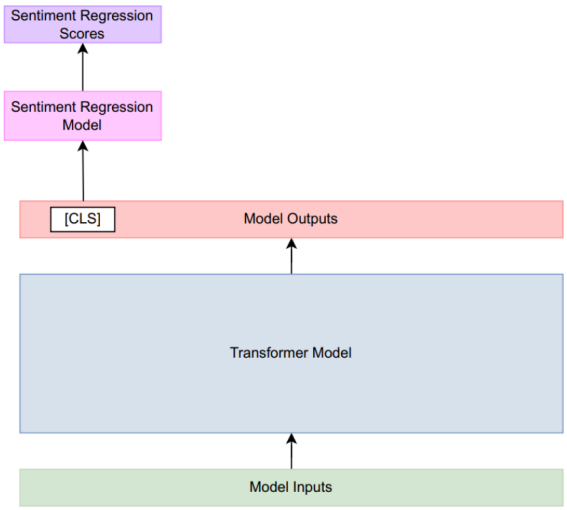

Figure 4. The neural network architecture for sentiment regression.

models where the transformer layers remain frozen, and only the regression head is updated.

Our regression architecture used a linear regression head for both single and multi-output configurations. The sentiment regression model was hyperparameter-tuned for the alpha parameter, using learning rates of 0.001, 0.01, 0.1, 1, 10, and 100, with 5-fold cross-validation. A data loader was used, configured with a batch size of 32. Finally, the models were trained for one epoch.

3. PARTIAL RESULTS AND DISCUSSION

3.1 Manual Annotation

The manual annotation process on 3,000 tweets yielded the results that are seen in the following graphs, with a separate graph for each polarity (positive, negative, and neutral) and scales ranging from 0 (sentiment not present at all) to 4 (sentiment is extremely present). Only tweets with at least one annotator giving a proper score were considered in the following graphs, leading to the sample being reduced from 3,000 tweets to 2,754 tweets. The loose inclusion criteria led to the possibility of one tweet's score being entirely affected by one annotator – a limitation of the current study and a point of improvement for future iterations.

The graphs in Figure 5, Figure 6, and Figure 7, show the distribution of the averaged manually annotated scores per tweet in each sentiment polarity, as well as the mean score for all tweets. 786 tweets from 1,425 instances were not given a sentiment score for various reasons detailed in Table 4.

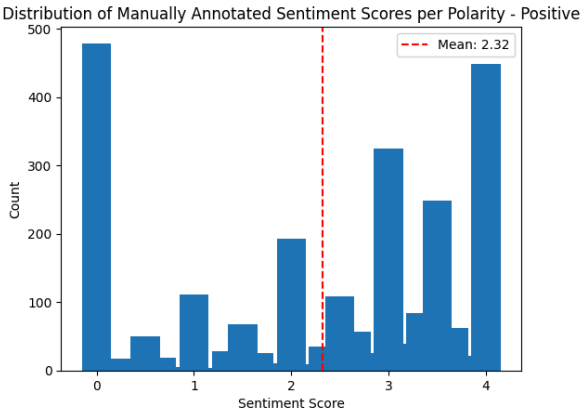

Figure 5. A histogram of positive scores of the human-annotated sample of tweets.

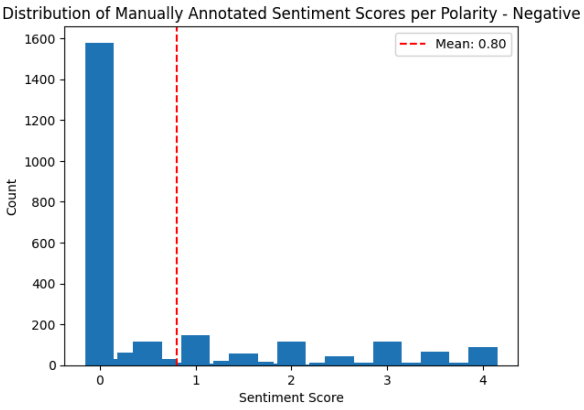

Figure 6. A histogram of negative scores of the human-annotated sample of tweets.

3.2 Correlation Between Manual Annotation Scores and MLE

The sample of 2,754 tweets that were given at least one sentiment score was automatically annotated by MLE, and a correlation test was performed between the manually annotated and automatically annotated

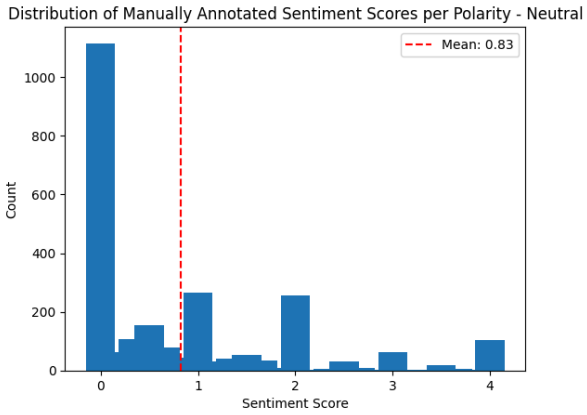

Figure 7. A histogram of neutral scores of the human-annotated sample of tweets.

Table 4. A summary of reasons when human annotators could not provide a sentiment score.

Reason	Number of Tweets	Number of Instances
Language not recognized	412	912
Unsure about sentiment	381	443
Unsure about sentiment & Language not recognized	65	69
Seems more sarcastic/a joke	1	1

data. Due to how MLE works, only the positive and negative sentiment polarities are considered for the correlation. Spearman's rank correlation will be used as the method for calculating correlation since the value ranges between the automatically annotated data and the manually annotated data are different and nonlinear (Schoeber et al., 2018).

The coefficients produced by performing Spearman's rank correlation are laid out in Table 5, with the score correlation for negative scores (0.6181) being more correlated than that of the positive scores (0.4747). We speculate that the negative label had a higher correlation with human annotation than the positive label because there were fewer tweets with a negative disposition.

3.3 Model Performance

All models were trained for one epoch. We summarize the performance of our models in Table 6 for the individual single-output models (i.e., per sentiment label), Table 7 for a macro-average perspective of the single-output models, and Table 8 for multi-output models.

Table 5. Spearman's rank coefficients between manually annotated and automatically annotated data.

Polarity	Spearman Coefficient (ρ)
Positive Score Correlation	0.4747
Negative Score Correlation	0.6181

Table 6. The performance of each transformer model using the single-output design across both positive and negative labels. Top-performing models are highlighted in bold.

Transformer	Label	Alpha	RMSE	R2
TwHIN-BERT	Positive	1	**0.21619**	**0.35915**
	Negative	1	**0.10296**	**0.24448**
RoBERTa-Tagalog	Positive	0.1	0.21955	0.33910
	Negative	0.1	0.10426	0.22522
mBERT	Positive	10	0.22810	0.28659
	Negative	10	0.10882	0.15609
Tagalog-BERT	Positive	100	0.23412	0.24847
	Negative	100	0.11029	0.13306

Table 7. A macro-average perspective of the performance of each transformer model using the single-output design. Top-performing models are highlighted in bold.

#	Transformer	Alpha	RMSE	R^2
1	TwHIN-BERT	1	**0.15958**	**0.30182**
2	RoBERTa-Tagalog	0.1	0.16191	0.28216
3	mBERT	10	0.16846	0.22134
4	Tagalog-BERT	100	0.17220	0.19076

Table 8. The performance of each transformer model using the multi-output design. Top-performing models are highlighted in bold.

#	Transformer	Alpha	RMSE	R^2
1	TwHIN-BERT	0.001	**0.15958**	**0.30182**
2	RoBERTa-Tagalog	1	0.16191	0.28216
3	mBERT	0.001	0.16846	0.22134
4	Tagalog-BERT	0.001	0.17766	0.14538

Based on our results, TwHIN-BERT achieves the best performance overall with the lowest RMSE and

highest R^2 scores across both configurations. We suspect that the likely advantage lies in its pretraining on Twitter data, which aligns closely with the domain of our current dataset. Additionally, the pretraining on multilingual likely led to a better understanding of Filipino-English code-switching.

On the other hand, RoBERTa-Tagalog consistently ranks second and outperforms mBERT despite mBERT's broader multilingual scope. While RoBERTa-Tagalog lacks domain alignment, its dedicated Tagalog pretraining appears to offer language-specific advantages not captured by mBERT. This result suggests that domain mismatch can be partially mitigated by language-focused pretraining.

Notably, the differences between the single- and multi-output designs, as seen in Table 7 and Table 8, are minimal for the top three models. One might mistake the performances to be identical at least up to the fifth decimal place; however, there are slight differences that emerge around the 7th decimal place. Regardless, Tagalog-BERT exhibits a decline in performance in the multi-output setting, which likely indicates a sensitivity to the increased modeling complexity.

4. CONCLUSION AND FUTURE WORK

Our paper illustrates the effectiveness of transformer-based models in conducting sentiment analysis on Philippine tweets using automatically generated labels. TwHIN-BERT achieved the best performance in both RMSE and R^2, outperforming the other evaluated models across both regression configurations.

Given the use of synthetic sentiment labels derived from emoji sentiment scores, we find the general performance of our models to be non-trivial in light of the inherent noise expected from transferring emoji-based sentiment scores to surrounding words. This suggests that emoji-based annotations, when used at scale, can serve as a viable source for learning sentiment in low-resource settings. We would also like to note that our proposed labeling scheme is rough in the sense that no noise has been filtered out. Additional tests to improve the transfer of sentiment knowledge from emojis to surrounding text would likely lead to interesting results.

Another potential improvement involves increasing the number of training epochs, which would entail prolonging the training duration but enabling the model to learn more complex linguistic patterns. However, this adjustment would also lead to increased training time.

Future work may also explore sampling strategies that ensure more uniform coverage across the sentiment score distribution. Our initial sample of tweets for manual annotation reveals that a good number of tweets were likely to be seen as positive rather than negative. This bias toward positive sentiment has also been observed in benchmark Twitter sentiment datasets such as SemEval-2017 (Rosenthal et al., 2017), where the number of positive tweets substantially outweighs the number of negative ones across multiple subtasks. While this bias is easier to spot when viewing the prediction task as classification, exploring sampling strategies may help mitigate bias and improve generalization for our approach to synthetic annotation.

We also recommend recording the running time for each model. Monitoring execution times throughout the training and evaluation pipeline can help pinpoint bottlenecks and optimize efficiency. In addition, tracking these metrics will support an analysis of the cost-efficiency versus accuracy trade-offs among models.

Finally, the differences in model performance raise the question of how well each model performs across different language segments. RoBERTa-Tagalog, for example, may outperform its peers when applied specifically to tweets written primarily in Filipino. Future work can look to evaluate each model by language or code-switching characteristics to better understand the strengths and limitations of each transformer in multilingual and mixed-language settings.

5. ACKNOWLEDGMENTS

We thank the insightful contributions of the panelists, whose expertise and suggestions greatly enhanced the quality of our work. We also thank all the individuals who participated in the manual annotation process. Their responses were essential in advancing our research and better understanding how well a synthetic annotation scheme to sentiment aligns with the human perspective.

6. REFERENCES

Alcober, G. M. I., & Revano, T. F. (2021). *Twitter Sentiment Analysis towards Online Learning during COVID-19 in the Philippines*. In 2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM) (pp. 1-6). IEEE.

Berkowitz, R. T., Wadden, T. A., Tershakovec, A. M., & Cronquist, J. L. (2003). *Behavior Therapy and Sibutramine for the Treatment of Adolescent Obesity* [Electronic version]. Journal of American Medical Association, 289, 1805-1812.

Co, E., Noblefranca, J., & Gan, J. M. (2023). *Affect Regression of Tweets from the Philippines using Emoji-based Labels*. [Unpublished undergraduate thesis]. De La Salle University.

Co, S. T., Custodio, N. R., Dela Cruz, A. L., Sanchez, M. C. (2022). *Sentiment Classification of Tweets from the Philippines with Emoji-based Data Annotation* [Unpublished undergraduate thesis]. De La Salle University.

Cosme, C. J., & De Leon, M. M. (2023). *Sentiment Analysis of Code-switched Filipino-English Product and Service Reviews using Transformers-based Large Language Models*. In World Conference on Information Systems for Business Management (pp. 123-135). Singapore: Springer Nature Singapore.

Delizo, J. P. D., Abisado, M. B., & De Los Trinos, M. I. P. (2020). *Philippine Twitter Sentiments during the COVID-19 Pandemic using Multinomial Naïve-bayes*. International Journal, 9 (1.3).

Godard, R., & Holtzman, S. (2022). *The Multidimensional Lexicon of Emojis: A New Tool to Assess the Emotional Content of Emojis*. Frontiers in Psychology, 13, 921388.

Imperial, J. M., Orosco, J., Mazo, S. M., & Maceda, L. (2019). *Sentiment Analysis of Typhoon Related Tweets using Standard and Bidirectional Recurrent Neural Networks*. arXiv preprint arXiv:1908.01765.

Jurafsky, D., & Martin, J. (2024). *Speech and Language Processing* [3rd ed. draft]. Retrieved from https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf

Maceda, L. L., Satuito, A. A., & Abisado, M. B. (2023). *Sentiment Analysis of Code-mixed Social Media Data on Philippine UAQTE using Fine-tuned mBERT Model*. International Journal of Advanced Computer Science and Applications, 14(7).

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A *Primer in BERTology: What We Know About How BERT Works*. Transactions of the Association for Computational Linguistics, 8, 842-866.

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 Task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 502–518). Vancouver, Canada: Association for Computational Linguistics.

Schober, P., Boer, C., & Schwarte, L. A. (2018). *Correlation Coefficients: Appropriate Use and Interpretation*. Anesthesia & Analgesia, 126(5), 1763-1768.

Taboy, C. (2023). *A Sentiment Analysis of "Filipinx" on Twitter Using a Multinomial Naïve Bayes Classification Model*. CUNY Academic Works. https://academicworks.cuny.edu/gc_etds/5234

Zhou, D.-X. (2020). *Theory of Deep Convolutional Neural Networks: Downsampling*. Neural Networks, 124, 319-327. https://doi.org/10.1016/j.neunet.2020.01.018

Zogaj, F., Cambronero, J., Rinard, M., & Cito, J. (2021). *Doing More with Less: Characterizing Dataset Downsampling for AutoML*. Proc. VLDB Endow., 14, 2059-2072. https://doi.org/10.14778/3476249.3476262