

EXPLORING TRANSFORMER-BASED APPROACHES IN SENTIMENT PREDICTION OF PHILIPPINE TWEETS

A Thesis Proposal

presented to

the Department of Software Technology
College of Computer Studies
De La Salle University

In partial fulfillment
of the requirements for the degree of

Bachelor of Science in Computer Science
Major in Software Technology

by

CALALANG, Hosea
ESTANOL, Miguel
JACINTO, Jon
LOPEZ, Mauries

Mr. Edward TIGHE
Adviser

December 2, 2024

Abstract

Transformers are widely recognized as state-of-the-art models in natural language processing (NLP) tasks like sentiment analysis due to their ability to capture contextual relationships in texts. Although some studies have employed transformers for sentiment analysis of Philippine text, they made use of a limited amount of data due to manual labeling. Such dataset sizes are not optimal for properly fine-tuning transformers. To address this gap, we aim to utilize an emoji-based sentiment scheme to synthetically label a large number of tweets. Using the abundant data provided by this automatic annotation scheme, this paper aims to explore the use of pre-trained transformer models from various domains and the impact of fine-tuning on their performance in sentiment prediction of Philippine text data.

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

Contents

1	Research Description	1
1.1	Overview of the Current State of Technology	1
1.2	Research Objectives	3
1.2.1	General Objective	3
1.2.2	Specific Objective	3
1.3	Scope and Limitations of the Research	3
1.4	Significance of the Research	4
2	Related Works	6
2.1	Sentiment Analysis in the Philippine Context	6
2.2	Sentiment Analysis in the Context of Low Resource Languages . .	8
2.3	Emojis and Automatic Annotation	9
2.4	Transformers and Sentiment Analysis	10
3	Theoretical Framework	12
3.1	Sentiment Analysis	12
3.2	Multidimensional Lexicon of Emojis	13
3.3	Text Processing	14

3.3.1	Noise Replacement	14
3.3.2	Tokenization	15
3.4	Machine Learning	15
3.4.1	Regression as a Task	16
3.4.2	Hyperparameter Tuning	16
3.5	Deep Learning Algorithms	18
3.5.1	Neural Networks	18
3.5.2	Transformers	18
3.5.3	BERT	20
3.6	Model Evaluation	21
3.6.1	Mean Squared Error	22
3.6.2	Root Mean Squared Error	22
3.6.3	Coefficient of Determination	23
3.6.4	Spearman’s Rank Correlation	24
4	Methodology	25
4.1	Data Source	26
4.2	Automated Annotation	28
4.3	Text Preprocessing	29
4.4	Neural Network Architechture	30
4.5	Experimental Setup	31
4.5.1	Data Split	32
4.5.2	Transformer Models	32
4.5.3	Sentiment Prediction Module	33

4.5.4	Hyperparameter Tuning	34
4.5.5	Model Evaluation	34
4.5.6	Experiment Summary	34
A	Research Ethics Documents	36
A.1	Research Ethics Clearance Form	36
A.2	General Research Ethics Checklist	39
A.3	Research Checklist A - Human Participants	43
A.4	Research Checklist G - Internet	51
	References	57

Chapter 1

Research Description

1.1 Overview of the Current State of Technology

Sentiment analysis is an approach to natural language processing (NLP) that focuses on extracting and classifying sentiments expressed in textual data. This technique is widely employed to gauge the positive or negative orientation expressed by authors toward various objects, such as films, books, products, or political entities (Jurafsky & Martin, 2024). Given its ability to capture public opinion and sentiments, sentiment analysis has found applications across industries like business (Aliyu, bt Sarlan, et al., 2024), finance (Mishev et al., 2020; Gondaliya et al., 2021), and law (Abimbola et al., 2024; Alqaryouti et al., 2020). The advent and rapid expansion of the internet, and the consumers' active engagement on online platforms have made sentiment analysis a global and dynamic field.

Recent studies indicate that sentiment analysis in the Philippines still relies on traditional Machine Learning (ML) methods for developing sentiment analysis models. Studies such as Alcober and Revano (2021) employed Naïve Bayes, logistic regression, and random forest in analyzing Twitter sentiments toward online learning during the COVID-19 pandemic. Similarly, Delizo et al. (2020) used Multinomial Naive Bayes and TF-IDF to study Philippine Twitter sentiments during the pandemic. Taboy (2023) also utilized the Multinomial Naive Bayes model with TF-IDF in their analysis of "Filipinx" on Twitter. However, traditional ML models are increasingly viewed as inadequate due to the complexities and nuances of modern sentiment analysis tasks, which require more advanced approaches to accurately capture context, especially considering the diverse language and inherent noise present in communication on online platforms.

Another significant challenge within the Philippine context is the lack of a large, annotated corpora, which leads researchers to create and manually annotate their datasets. This process is time-consuming and becomes increasingly difficult with larger datasets. These issues result in smaller datasets for training and evaluation. For instance, Taboy (2023) manually labeled 2,000 tweets out of a total of 52,059 tweets, resulting in a final dataset of 1,853 tweets. Similarly, Imperial, Orosco, Mazo, and Maceda (2019) manually labeled 3,900 out of 39,867 typhoon-related tweets for initial training data, while Cureg et al. (2019) hired three psychology majors to annotate 11,897 tweets.

The growing complexities of sentiment analysis tasks, coupled with the significant challenges in the Philippine landscape, underscore the need for more advanced and adaptable approaches. Transformers present significant opportunities for sentiment analysis, with two local studies already demonstrating early success. Cosme and De Leon (2024) sentiment-annotated corpus of Filipino-English reviews and fine-tuned sentiment analysis models, achieving high accuracy and F1 scores. Similarly, Maceda et al. (2023) utilized a fine-tuned multilingual model to analyze sentiments regarding the Universal Access to Quality Tertiary Education (UAQTE) program in the Philippines, which achieved an accuracy of 80.21 percent and an F1 score of 81.14 percent.

Another promising direction involves leveraging an emoji-based automatic sentiment annotation, which offers the potential to handle large datasets efficiently, increasing the amount of available data and reducing reliance on manual annotation. For instance, S. Co et al. (2022) explored the use of emoji-based sentiment annotation to create a large dataset of annotated tweets using the Emoji Sentiment Ranking lexicon by Kralj Novak et al. (2015). E. Co et al. (2023) expanded on this by employing a multi-dimensional, emoji-based sentiment annotation through the use of the Multidimensional Lexicon of Emojis by Godard and Holtzman (2022) to assign affect labels.

Building on these insights, this study proposes to explore transformer-based approaches for sentiment prediction of **Philippine** texts with emoji-based automatic sentiment annotation as initial training data. Specifically, it seeks to determine the effectiveness of these approaches to navigate and manage the inherent noise in **Philippine** text data without relying on expert-labeled data.

1.2 Research Objectives

1.2.1 General Objective

This research aims to explore transformer-based approaches in sentiment prediction of **Philippine** text data.

1.2.2 Specific Objective

1. **To investigate the issues in automated mixed language Filipino-Text sentiment analysis.**
2. To collect **Philippine** text data and implement an automatic annotation scheme using an emoji-based sentiment lexicon.
3. **To develop a data processing and training pipeline for sentiment prediction models using various transformer models.**
4. **To evaluate the performance of our models.**

1.3 Scope and Limitations of the Research

This study investigates the challenges and opportunities in automated sentiment analysis of Philippine text data, focusing on integrating different types of transformer models and an emoji-based automatic annotation scheme.

The datasets used in this study were created as part of the research by S. Co et al. (2022) **and** E. Co et al. (2023). These datasets were generated using the official Twitter API, which allows tweets to be filtered by the location where the tweet was made. A bounding box that covers the land territorial extent of the Philippines was used to determine whether the tweets were generated within the Philippines (S. Co et al., 2022; E. Co et al., 2023). Data from Twitter can generate a significant number of emojis, which our study will leverage for sentiment annotation.

For this purpose, we will utilize the automated sentiment annotation scheme of E. Co et al. (2023), which assigns sentiment scores to tweets based on emoji content. Our annotation approach will make use of sentiment scores from the Multidimensional Lexicon of Emojis (MLE) developed by Godard and Holtzman (2022) on tweets that contain emojis listed in said emoji lexicon. MLE consists of

359 common emojis’ sentiment directions, two of which are sentiment polarities, positivity and negativity, which the study plans to focus on.

While this scheme offers a scalable solution for generating annotated datasets, it is important to note that it is synthetic in nature, and that some of the data produced may not be completely usable. We also recognize the inherent noisiness of social media data and its sensitivity to time and events. Moreover, reliance on the Twitter API during the collection period represents another limitation of this study. Due to the dataset’s size, tweets will not be manually validated for satiric tones, such as sarcasm or irony, which may convey sentiments that differ from their literal textual meaning. As a result, the annotation scheme may not effectively capture such sentiments.

The study is focused on predicting sentiment using pre-trained transformer models, which will be fine-tuned for the task rather than trained from scratch. As such, domains beyond sentiment prediction, such as sarcasm or irony detection, will not be explored. We aim to assess the adaptability and effectiveness of Transformer models for sentiment prediction using Philippine text data. Thus, we aim to explore various transformer models, including general-purpose models as well as models specifically trained for sentiment analysis, social media data, Filipino text data, and multilingual data. The study primarily focuses on encoder-based architectures, which have been widely used in sentiment prediction literature. While decoder-based models could be explored, they may not be as effective as their encoder counterparts. We will use a regression approach in sentiment prediction to align with the sentiment scores from the MLE. We are considering a validation approach that involves splitting the annotated data into training, validation, and testing sets and performing hyperparameter tuning.

As we treat sentiment prediction as a regression task, our models will be evaluated using regression metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Coefficient of Determination (R^2). We will also use a correlation metric such as Spearman’s Correlation to correlate our models’ performance with existing tools such as the Valence Aware Dictionary and Sentiment Reasoner (VADER) to validate their effectiveness in sentiment prediction.

1.4 Significance of the Research

This study contributes to the growing field of sentiment analysis by examining the performance of transformer models on Philippine social media text data. Fur-

thermore, analyzing the performance of different transformer models on Philippine text data provides a broader understanding of its use through insights into how these models interpret Philippine text data alongside other social media elements and emoji annotations. The findings contribute to a better understanding of transformers adaptability to local social media data and offer perspectives on fine-tuning techniques to optimize model performance for sentiment prediction tasks.

Developing sentiment prediction models addresses the lack of readily available tools for processing Philippine text data. While this study does not fully resolve issues associated with the Philippine scene, this study serves as a foundational effort toward creating more robust tools for sentiment analysis in the Philippine context. The resulting methods and insights can guide future research in building efficient and scalable sentiment prediction pipelines tailored to Philippine data. This work is expected to benefit researchers by promoting advancements in the local Natural Language Processing (NLP) scene, specifically sentiment analysis, where transformer models play a growing role.

Chapter 2

Related Works

The following literature review explores sentiment analysis, primarily focusing on the Philippine context. It examines the use of traditional machine-learning methods in the Philippines and addresses the associated challenges. The review underscores emojis’ potential as sentiment indicators, especially in multilingual and code-mixed environments. Additionally, it discusses the difficulties faced in low-resource languages (LRLs) and explores various strategies to mitigate these challenges. Lastly, the review delves into the impact of transformer models in recent years and the different approaches to enhancing sentiment analysis performance in low-resource and social media settings.

2.1 Sentiment Analysis in the Philippine Context

In the Philippines, sentiment analysis has relied on traditional machine learning approaches, often employed in domain-specific applications. For instance, Delizo et al. (2020) employed Multinomial Naïve Bayes to analyze sentiments during the COVID-19 pandemic, which resulted in an accuracy of 72%. Taboy (2023) use of Multinomial Naïve Bayes for Sentiment Analysis of “Filipinx” yielded 61.74% accuracy score. According to S. Co et al. (2022), these studies no longer represent state-of-the-art approaches and perform poorly compared to more recent methodologies.

The research scene has begun to adopt deep learning approaches for sentiment analysis, with studies utilizing models such as Convolutional Neural Networks

(S. Co et al., 2022), Recurrent Neural Networks Imperial et al. (2019) and Long Short Term Memory networks (Frias et al., 2023). Transformers have also been used in Sentiment Analysis of Philippine data. However, its use in the local context remains in its infancy. Existing studies are often limited to specific topics, such as Maceda et al. (2023) use of Universal Access to Quality Tertiary Education (UAQTE) social media data for sentiment classification tasks and Cosme and De Leon (2024) study on code-switched Filipino-English product and service reviews. Nevertheless, these pioneering studies represent significant advancements in the application of transformers for Philippine sentiment analysis, laying the groundwork for future research and development.

Annotating datasets for use in Philippine Sentiment Analysis has primarily been done using two approaches:

- Lexicon-based annotation strategies rely on pre-defined word lists to assess sentiment but are limited by domain specificity and lack flexibility across contexts. This is particularly challenging in the Philippine context, as readily available Lexicons are predominantly English-based. Alcober and Revano (2021) use of the “bing” and AFINN lexicons, which are general-purpose English sentiment datasets used in R programming, would omit Filipino text in sentiment computations. Contreras et al. (2018) worked around this limitation by combining an English Opinion lexicon with Tagalog words from a Kaggle dataset to analyze airline reviews. The limited scope and rigid sentiment rules of these lexicons, no matter the context, make them less effective in broader sentiment analysis tasks.
- Given the pitfalls of lexicon-based annotation, manual annotation by experts would be the more prominent approach in the field, where experts would annotate private datasets on sentiment polarities. Manual annotation would also overcome the difficulties in defining sentiments in Philippine texts. Such a strategy was employed in studies conducted by Boquiren et al. (2022), who hired three domain experts to annotate 32,036 tweets, or Cureg et al. (2019), hiring three psychology professionals to categorize tweets into five labels. However, manual annotation is time-consuming and labor-intensive, which limits scalability. As a result, experts have annotated only a fraction of the collected datasets. This, in turn, questions the extent to which the manually annotated dataset can capture the sentiments of their respective domains.

Table 2.1: Dataset Reduction in Various Philippine Sentiment Analysis Studies

Study	Initial Size	Final Size	Reduction	Retained
Taboy (2023)	63,622 tweets	1,823 tweets	97.09%	2.91%
Imperial et al. (2019)	92,040 tweets	3,900 tweets	95.76%	4.23%
Cureg et al. (2019)	11,897 tweets	5,900 tweets	57.97%	42.03%
Maceda et al. (2023)	13,332 data points	3,900 data points	70.75%	29.25%

2.2 Sentiment Analysis in the Context of Low Resource Languages

These limitations are not unique to the Philippine context, however. They are common challenges across general Low-Resource Languages (LRLs). With the diversity of languages included in the Low Resource classification, studies have worked on new tools and have worked with models and methods to cope with the inherent properties of LRLs.

A significant challenge for LRLs is the lack of high-quality, publicly-available annotated datasets, which limits advancements in sentiment analysis and other NLP tasks. Researchers in certain low-resource language groups have undertaken a more comprehensive approach to address this challenge by creating large annotated corpora for various NLP tasks. Efforts to address this challenge have been exemplified by initiatives such as IndoLEM (Koto et al., 2020) in Indonesia and AfriSenti (Muhammad et al., 2023) in the African continent. IndoLEM introduced a comprehensive dataset comprising seven NLP tasks, including sentiment analysis for the Indonesian language. Similarly, AfriSenti, the largest sentiment analysis benchmark for African languages, provides over 110,000 annotated tweets across 14 languages from diverse language families.

Low-resource languages (LRLs) have made notable strides in adopting deep learning methods for sentiment analysis of morphologically rich languages. For example, Ullah et al. (2022) explored a CNN-LSTM hybrid model for sentiment analysis in Roman Urdu, achieving 93.3% accuracy by combining word embeddings and deep neural networks. Similarly, Gupta et al. (2021) developed an integrated CNN-RNN model for Hindi tweets, demonstrating an 85% accuracy rate. Shehu et al. (2024) found that CNNs slightly outperformed RNN and hierarchical attention networks (HAN) for Hausa sentiment analysis. Still, overall model accuracy remained modest (68.48%, 63.22%, 67.90%, for CNN, RNN, and HAN, respectively). These advancements signify a shift towards sophisticated sentiment analysis techniques within the Low-Resource NLP landscape.

Preprocessing is another area of consideration for LRLs. Preprocessing approaches such as machine translation, where text is translated to or from a resource-rich language, have been utilized in studies such as that of Ghafoor et al. (2021) and Sabri et al. (2021). However, as Aliyu, bt Sarlan, et al. (2024) noted, translation is highly dependent on the source language’s complexity and the tool’s accuracy, among other factors. Beyond translation, many studies perform standard preprocessing tasks like lowercasing and removal of certain textual elements, which, while justified because it saves on memory and computational resources, do remove information within a text.

2.3 Emojis and Automatic Annotation

One such element frequently removed in sentiment analysis of text data is emojis. Studies across a spectrum of methods, from traditional machine learning to deep learning, commonly omit emojis as they are considered noisy elements that hinder accurate analysis. Several studies, however, argue that emojis further emphasize the sentiment of a text by providing a language-independent method of conveying emotion and intention (Hakami et al., 2022; Li et al., 2018; Shiha & Ayvaz, 2017). (Hakami et al., 2022) specifically highlighted that emojis “can behave as an emphaziser, an indicator, a mitigator, a reverser, or a trigger of either negative or positive sentiment within a text.”

Given the language-independent nature of emojis, researchers have adopted them as novel features for sentiment annotation through the use of emoji sentiment lexicons. One such lexicon is the Emoji Sentiment Ranking (ESR) by Kralj Novak et al. (2015)., which introduced a lexicon of 751 frequently used emojis mapped by sentiment. ESR is the emoji sentiment lexicon of choice in S. Co et al. (2022) research, which enabled them to annotate over 2.5 million tweets. A more recent lexicon, the Multidimensional Lexicon of Emojis (MLE), was developed by Godard and Holtzman (2022) using data from over 3 million tweets and ratings from 2,230 human annotators. E. Co et al. (2023) utilized the MLE to automatically annotate on ten affect dimensions, resulting in a dataset of 1.81 million tweets.

Despite the advantages of automatic sentiment annotation using emojis, there are challenges to this approach. One limitation of emoji-based annotation is the inherent ambiguity of emojis, which can complicate sentiment classification. This is because some emojis are used with the intention of irony or sarcasm (Chen et al., 2018), examples of which being the crying face emoji (😭) and the face with tears of joy emoji (😂), both of which were explicitly specified in a study concerning S. Co et al. (2022). These lexicons, given time, could also show begin

their age. The ESR was developed in 2015, and thus, it excludes newer emojis introduced since then, which limits the ability include these emojis in automatic annotation schemes, a drawback also noted by S. Co et al. (2022).

2.4 Transformers and Sentiment Analysis

In addition to emojis and emoji lexicons, the advent of transformers for various natural language processing (NLP) tasks presents opportunities for further advancing sentiment analysis. Vaswani et al. (2017) introduced the transformer model that fundamentally changed how we approach NLP tasks by leveraging self-attention mechanisms. This allows the model to weigh the importance of different words in a sentence, regardless of their position.

Unlike traditional ML models and lexicon-only approaches, transformers excel in sentiment analysis because they can understand the context and syntax of words in a sentence (Mishev et al., 2020). Traditional ML models are limited by their inability to capture the full context of a sentence, and their poor performance on social media data has been observed (Drus & Khalid, 2019). Lexicon approaches that depend on predefined lists may not account for the complexity and variability of natural language. Transformers address these limitations through transfer learning, where models leverage knowledge from large pre-trained corpora. This, in turn, enables effective performance on target domains (Zhuang et al., 2020). The strong contextual understanding of transformers allows sentence inference from using words in context.

BERT (Bidirectional Encoder Representations from Transformers) and its derivatives have been especially influential in establishing new benchmarks across various NLP tasks. In a survey of 150 studies, Rogers et al. (2020) concluded that BERT has become a ubiquitous baseline for NLP Deep Learning Architecture of its kind, two years after the model’s release. This inevitably caught the attention of Low-Resource studies in various NLP applications, which presents new challenges and opportunities. Two primary strategies have emerged to address these challenges: continued pre-training and fine-tuning

Continued pretraining primarily involves further training a pre-existing model on additional, specific data. Training begins with a robust model like BERT or RoBERTa and involves feeding said models with large amounts of target data. This approach has shown promise in code-mixed sentiment analysis, where models are trained on a mix of available language resources. One notable example of this approach is the AfriBERTa model (Ogueji, 2022), which was initially based on a

pre-trained mBERT model. AfriBERTa was further trained using a large corpus of text from 11 African languages. The study by Aliyu, Sarlan, et al. (2024) using the said model shows the model’s remarkable performance on established benchmarks, with an F1 score of 81% and an accuracy rate of 89%. Another is the WangchanBERTa model by Lowphansirikul et al. (2021), which takes the pre-trained RoBERTa model and fine-tunes it on a large corpus of Thai text.

In addition to low-resource languages, this approach has also been applied to social media data, resulting in models like BERTweet (Nguyen et al., 2020). BERTweet was developed by training a RoBERTa model on a large corpus of English tweets designed to handle the unique linguistic features and informal language prevalent on Twitter. Transformers under this approach demonstrate enhanced performance in understanding the target data while leveraging an existing model’s strengths. However, this approach requires substantial computational resources and the availability of sufficient training data.

Fine-tuning transformers is another process for low-resource tasks wherein a pre-trained model is further trained on a smaller, task-specific dataset. This approach has been widely used in low-resource sentiment analysis. This task typically requires calibration of the model’s parameters, a process better known as hyperparameter tuning. Comparative finetuning in low-resource languages has been explored by Azhar and Khodra (2020) and Raychawdhary et al. (2023), wherein pre-trained transformers with different hyperparameter configurations yielded varying accuracies in sentiment classification tasks.

In the Philippines, the application and adaptation of transformer models are still in their early stages. Notable pioneering efforts in pretraining include Cruz and Cheng (2020, 2022) development of Tagalog-BERT and RoBERTa-Tagalog. Maceda et al. (2023), on the other hand, advanced the local sentiment analysis field with one of the first publicly available studies on fine-tuning for downstream tasks by applying the mBERT model to classify sentiments in social media data related to Universal Access to Quality Tertiary Education (UAQTE).

The current underutilization of these models in the local context presents an untapped opportunity for extending social insights in a country where online discourse and media consumption are significant (Mateo, 2024). By leveraging both transformer-based approaches emoji sentiment annotation, there is potential to foster a more nuanced understanding of Filipino online engagement and sentiments.

Chapter 3

Theoretical Framework

3.1 Sentiment Analysis

Sentiment, as defined outside Computer Science, Natural Language Processing, and Computational Linguistics, is deeply rooted in human social interaction and reflective processes. Cooley, as cited in Stets (2003), describes sentiment as emotions shaped by social interaction and reflection. Unlike raw instincts, sentiments are uniquely human feelings that develop through our capacity to imagine and empathize with others' perspectives.

Expanding on this, Gordon, as cited in Stets (2003) further identifies sentiment as a socially constructed combination of autonomic responses, expressive behaviors, and shared meanings often centered around another person or a collective of people.

Sentiment analysis refers to a field of research dedicated to defining automatic tools for extracting this subjective information—or "sentiment"—from natural language text (Pozzi et al., 2017). Applications of sentiment analysis span social evaluations, analysis in social media networks, and the extraction of insights about products or brands. Within sentiment analysis, various subtasks are explored, including polarity classification, subjectivity classification, sarcasm detection, and opinion summarization.

Models performing sentiment analysis tasks fall into two categories: *regression* and *classification*.

- *Regression models* contain a number or a series of numbers as an output.

- *Classification models* contain categories or text labels (positive & negative, spam & not spam, harmful & not harmful messages as an output.

3.2 Multidimensional Lexicon of Emojis

A lexicon of emojis based on the NRC Emotion Lexicon is the first emoji lexicon to include eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, as an addition to the basic sentiment categories of negative and positive sentiments. The emojis were collected from three million tweets produced at three-time points. Moreover, emojis that are highly context-dependent and potentially affect their usage by current events were excluded, such as national and regional flags. The NRC Emotion Lexicon (Mohammad & Turney, 2013) is a crowd-sourced word-emotion association lexicon, which contains the basis for the MLE’s categorical scheme of eight emotions and two polarities (Godard & Holtzman, 2022).

To ensure a fair representation of emojis collected from the tweets, only those present in at least 50 tweets from each of the three-time points were retained. Furthermore, tweets were scored for each of the eight emotions, as well as the negative and positive sentiments, using the NRC Emotion Lexicon. The NRC Emotion Lexicon scores represent the number of words associated with each emotion or sentiment. For instance, if a tweet contains five words related to joy, a score of 5 is given for joy. Each emoji is then assigned a score for the eight emotions and two sentiment categories by averaging the NRC Emotion Lexicon scores of all tweets containing that specific emoji. Finally, average the NRC Emotion Lexicon scores of tweets containing each emoji and for each of the ten dimensions of the emoji across all three-time points. These scores comprise the final MLE (Godard & Holtzman, 2022).

Furthermore, they recruited participants to complete an online survey in which they were asked, "To what extent does this emoji communicate the following emotions?" using a 4-point Likert scale, where 4 indicated "a lot." Additionally, they were asked to rate the emoji’s positivity and negativity using the same scale. Then, the scores of each emoji or sentiment were generated by averaging the ratings provided by each participant (Godard & Holtzman, 2022).

A total of 1,458 emojis were identified in the 678,789 emoji-containing tweets, with 359 emojis (25%) meeting the criteria of appearing in at least 50 tweets at each time point.

Table 3.1 presents the overall statistics for positive and negative sentiments

Table 3.1: Descriptive Statistics of MLE’s Positive and Negative Dimensions

	Positive	Negative
Min.	0.57	0.20
Max.	2.80	1.03
Mean	1.07	0.47
Median	1.06	0.44
σ	0.33	0.15

per emoji. It highlights that more emojis convey positive sentiments than negative ones. As shown, the positive sentiment exhibits a broader range of scores, indicating a stronger and more varied expression of positive sentiment across emojis. In contrast, the negative sentiment scores are more concentrated, reflecting a narrower range of variation for negative sentiment.

Moreover, Spearman’s correlations were used to determine the strength of the association between the scores based on the NRC Emotion Lexicon and human-generated dimensions. The correlation coefficients of each dimension are shown in Table 3.2.

Table 3.2: Spearman’s Correlation between Positivity and Negativity using NRC Emotion Lexicon and Human-Generated Data in MLE

Dimension	Spearman’s correlations
Negativity	0.83
Positivity	0.61

Godard and Holtzman (2022) indicated that the value for the negativity dimension is good while adequate for the positive dimension.

3.3 Text Processing

3.3.1 Noise Replacement

Noise replacement refers to the process of identifying and handling elements in text data that may interfere with model generalization. In the context of social media content, elements like URLs and user mentions may be replaced with placeholder tokens to preserve sentence structure.

3.3.2 Tokenization

Tokenization is the process of breaking down text into smaller units, known as tokens. Traditionally, text could be tokenized into either words or individual characters, but a third option, subword tokenization, has become the standard in modern language models. Subword tokenization is a method in natural language processing that breaks down words into smaller units or "subwords." This approach allows models to handle out-of-vocabulary words by decomposing them into familiar subword units. Transformers make use of the following tokenization algorithms based on subword tokenization:

Byte Pair Encoding (BPE), adapted by Sennrich et al. (2016) for NLP, is designed to reduce out-of-vocabulary issues by breaking down words into subword units. The BPE algorithm operates by iteratively replacing the most frequent pair of characters or character sequences in a corpus with a new token. Initially, every unique character in a text sequence is treated as a single token. BPE identifies and merges the most frequent adjacent pair of tokens repeatedly until reaching a predefined vocabulary size.

WordPiece is a tokenization algorithm developed by Google (2018) to address the tokenization needs of the BERT language model. Like BPE, WordPiece decomposes words into subwords and learns to merge the most frequent token pairs. However, unlike BPE, which prioritizes the most frequent pair at each step, WordPiece selects the pair that maximizes the likelihood of the resulting subwords across the corpus.

SentencePiece is a language-independent tokenizer designed to simplify tokenization without requiring language-specific pre-tokenization (Kudo & Richardson, 2018). Sentencepiece implements both BPE and Unigram Language model. Distinctively, SentencePiece encodes each input as a sequence of Unicode characters, which is especially advantageous for multilingual corpora. SentencePiece also introduces a unique whitespace-handling feature: it replaces whitespace with the Unicode symbol U+2581, also known as the lower one-quarter block character, effectively preserving spacing information throughout tokenization.

3.4 Machine Learning

Machine learning is a branch of artificial intelligence that focuses on developing algorithms that enable systems to learn from data and make predictions without explicit programming. These models identify patterns within data and improve

their performance over time through iterative training and tuning. The subsequent subsections introduce some supervised learning tasks commonly employed by deep neural networks.

3.4.1 Regression as a Task

Regression is a supervised learning algorithm aimed at predicting the value of a target variable for new, unseen data based on one or more input features. It particularly works for problems where the predictions are in real or continuous values.

Regression models come in various forms, one of the most common being the Linear Regression Model. Linear Regression is a method for modeling the relationship between a dependent and one or more independent variables. It aims to fit a straight line to the data by minimizing the differences between observed and predicted values. Linear Regression is expressed by the equation:

$$\hat{y} = \theta_1 x + \theta_0$$

where \hat{y} is the dependent variable, x is the independent variable θ_1 is the slope and θ_0 is the intercept or model parameter.

3.4.2 Hyperparameter Tuning

Hyperparameter tuning refers to the process of selecting the most optimal values for a model's hyperparameters. Unlike parameters, which are learned during the training process, hyperparameters are preconfigured and define the architecture and behavior of the model even before the learning process begins. Bischl et al. (2023) noted that selecting hyperparameter values with care is important to achieve optimal performance due to their substantial influence on the model's complexity, behavior, and speed.

In the context of Deep Learning Algorithms, examples of hyperparameters that require calibration include:

- **Batch Size** refers to the number of training examples one GPU uses in one training step.
- **Learning Rate** is the rate an algorithm updates during each iteration.

- **Training Epoch** corresponds to one complete pass over the training data during training.
- **Number of Neurons per Layer** refers to the number of nodes in each hidden layer. It determines the width of the model
- **Number of Hidden Layers** determines the depth of the model

Tuning is accomplished through the following techniques:

Train-Test Split: The train-test split is a method used to evaluate the performance of machine learning models by dividing a dataset into two parts: a train set and a test set. The train set is used to fit the model; the test set is used to assess the model's ability to generalize to new and unseen data. A variation of this is the train-validate-test split, where an additional validation set is used to fine-tune hyperparameters. The validation set provides a benchmark for model performance and prevents overfitting.

K-Fold Cross-Validation: K-fold cross-validation is a method in machine learning for model evaluation wherein a dataset is split into k equally sized subsets or "folds." For each iteration, one fold is used for testing, while the remaining k-1 folds are used for training. This process is repeated k number of times, after which performance scores from each k run are averaged.

Grid Search: Grid search is a hyperparameter tuning method that systematically evaluates all possible combinations of a predefined set of hyperparameter values. The goal is to find the best-performing model given the combination of hyperparameters. Given its simplistic and exhaustive nature, grid search can become computationally expensive, especially with large datasets or a wide range of hyperparameters.

Random Search: Random search is another hyperparameter tuning method that randomly selects for combinations of a predefined set of hyperparameter values. The model is then trained and evaluated for each random combination, and the best combination is selected. Random search is less systematic and effective at finding the optimal hyperparameters than grid search but is computationally more efficient.

3.5 Deep Learning Algorithms

3.5.1 Neural Networks

Neural networks are computational models inspired by the structure of the human brain, consisting of layers of interconnected nodes called neurons. These neurons are organized into three primary types: input neurons, which receive external data; hidden neurons, which perform intermediate computations; and output neurons, which provide the final result. Information travels through the network via weighted connections, where each neuron applies a threshold to determine if it will activate and pass information forward.

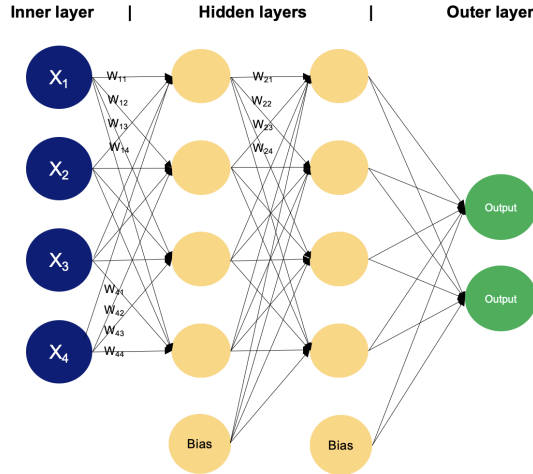


Figure 3.1: Fully Connected Layer (Jaiswal, 2024)

Fully Connected Layers, also known as dense layers, are a specific type of neural network layer in which each neuron is connected to every neuron in the adjacent layers. This architecture enables the network to perform linear transformations on the data, followed by an activation function. Multilayer Perceptrons (MLPs) are an example of a feedforward neural network composed entirely of Fully Connected Layers. MLPs are trained through a feedforward process, where input data moves layer by layer without feedback, and weights are updated using backpropagation.

3.5.2 Transformers

The Transformer is a neural network architecture designed to learn from sequential data, analyze patterns across the entire input, and generate new data based on

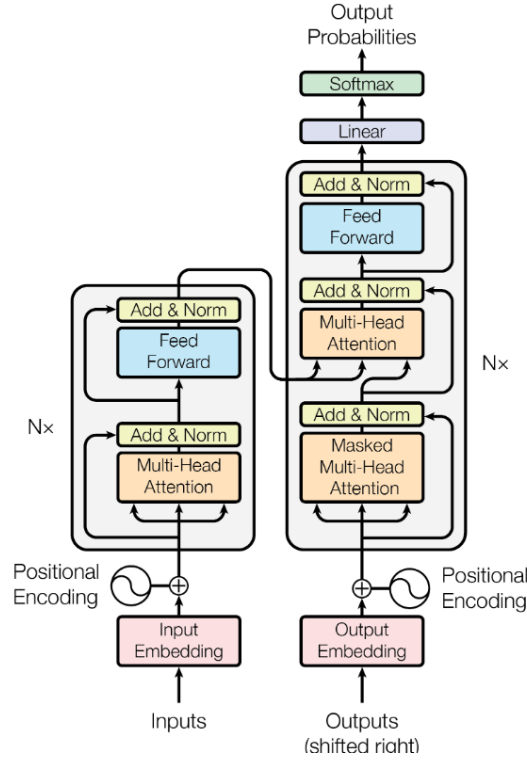


Figure 3.2: Transformer Architecture from Vaswani et al. (2017)

learned patterns. It achieves this by using self-attention mechanisms, which allow the model to focus on different parts of the input simultaneously rather than processing elements in sequence like traditional models. Transformers follow the encoder-decoder structure, where the encoder processes the input sequence into a set of context-rich representations, and the decoder uses these representations to generate outputs, one token at a time.

Transformers mainly consist of three parts:

The Input Encoding Component converts discrete tokens into vector representations through embedding layers. The input of a transformer model is often a tensor of shape $\mathbb{R}^B \times \mathbb{R}^N$, where B is the batch size and N is the sequence length. The first input passes through an embedding layer, which converts each one-hot token into a d_{model} dimensional embedding, resulting in a tensor of shape $\mathbb{R}^B \times \mathbb{R}^N \times \mathbb{R}^{d_{\text{model}}}$. Additionally, transformers incorporate positional encoding, which allows the model to differentiate between tokens based on their positions within the data. The token embeddings and positional encodings are combined before passing through the transformer blocks.

Once tokens are embedded, the transformer processes these vectors using stacked layers of Transformer Blocks. Each block consists of two sublayers: multi-head attention and feed-forward layers. Both layers are connected through residual connections and layer normalization.

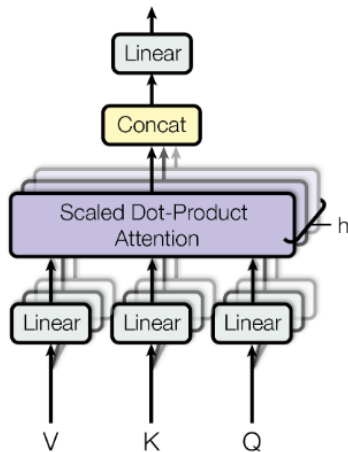


Figure 3.3: Multi-Head Attention from Vaswani et al. (2017)]

The key feature of transformers is self-attention mechanisms. This allows the model to weigh the importance of each input element relative to other elements in a sequence. It enables the transformer to capture dependencies across long distances without sequential processing. Transformers use multiple attention heads, each computing attention scores for input tokens in parallel. After multi-head self-attention operations, the outputs are passed into the Feed-Forward layer. The Feed-Forward Network (FFN) is a fully connected 2-layer network responsible for linear and non-linear transformations.

The Language Modelling Head serves as the final component of the transformer architecture. It takes the output of the final transformer block and converts it into predictions through a softmax layer. This head is responsible for the next token in a sequence during training or generating a token during inference.

3.5.3 BERT

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2019), is a state-of-the-art transformer model built upon the Transformer architecture developed by Vaswani et al. (2017). BERT improves upon the Transformer architecture by introducing bidirectional encoding. Bidirectionality allows it to capture context from both the left and right sides of a token, unlike

unidirectional models, which read text sequentially. BERT is pre-trained using two unsupervised tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). This pre-training is carried out on massive corpora, including BooksCorpus (800 million words) and English Wikipedia (2.5 billion words).

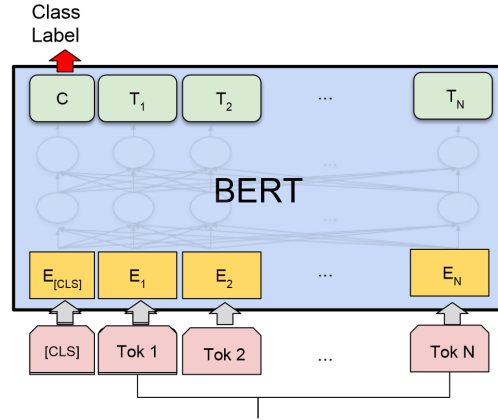


Figure 3.4: BERT used for Single Sentence Classification (Devlin et al., 2019)

BERT’s architecture is designed to be versatile and readily adaptable for fine-tuning across various downstream tasks. One such task, classification, is achieved by simply adding a task-specific layer, which in this case is classification, on top of its architecture. In this setup, BERT’s architecture leverages special tokens such as [CLS], which captures a pooled representation of the input sequence for classification purposes and the [SEP] token to denote segment boundaries. The final hidden state of the [CLS] token is used to represent the entire sequence for classification tasks.

3.6 Model Evaluation

Model evaluation is the process of determining whether a model needs improvement by analyzing its performance or assessing if it is making accurate predictions. Different evaluation metrics are used for classification and regression tasks. Several evaluation metrics are used for regression tasks, such as Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2). Moreover, since regression involves predicting numerical values, the evaluation metrics used for classification cannot be applied.

3.6.1 Mean Squared Error

Mean Squared Error (MSE) is a commonly used metric that measures the average squared difference between the predicted values and the actual values in a regression model. The MSE metric provides a straightforward way to gauge how well the model's predictions align with actual values.

The formula for MSE is as follows:

$$\text{MSE} = \frac{1}{n} \sum (\hat{y} - y)^2$$

To calculate MSE, the difference between each predicted value (\hat{y}) and the actual value (y) is computed and then squared to ensure all errors are positive and to emphasize larger deviations. These squared differences are then summed across all instances in the dataset, with the final step being to divide by the total number of observations (n) to obtain the mean squared error.

A low MSE value indicates that the model's predictions are close to the actual values. Since MSE is sensitive to larger errors due to the squaring of differences, a high MSE can suggest that the model's predictions contain significant deviations from the true values.

3.6.2 Root Mean Squared Error

Root Mean Squared Error (RMSE) is one of the evaluation metrics used to evaluate regression tasks. It measures the average distance between the predicted and actual values of the model, where it minimizes the effect of outliers. RMSE is particularly useful when large errors are not specifically needed for the model, as it penalizes large errors heavily.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum (\hat{y} - y)^2}$$

The equation starts by calculating the difference between the predicted and actual values and then squaring the result to calculate the RMSE value. Next, the squared differences are summed and averaged by the total number of instances. Finally, the square root is applied to the averaged squared error value.

A low RMSE value means that the model's predictions are close to the actual values, which implies good model performance. The lower the value, the better

the model at predicting instances. Like MSE, RMSE is sensitive to outliers because it calculates the squared difference between the predicted and actual values. However, once the square root is applied, the values are returned to their original value, which minimizes the effect of outliers.

3.6.3 Coefficient of Determination

The coefficient of Determination, also known as R^2 , measures how much the features explain the target variable. In other words, this evaluates the performance of regression models in predicting how well the model fits the data.

$$R^2 = 1 - \frac{\sum(\hat{y} - y)^2}{\sum(y - \bar{y})^2}, \text{ where } \bar{y} \text{ is the mean of the labels}$$

To compute the coefficient of determination, we must first compute the total sum of squares and the sum of squared residuals, which are denoted as TSS and RSS, respectively.

$$\text{TSS} = \sum (y - \bar{y})^2$$

$$\text{RSS} = \sum (\hat{y} - y)^2$$

To obtain the total sum of squares (TSS), calculate the difference between the actual values and the mean of all the data points. For the residual sum of squares (RSS), calculate the difference between the predicted values and the actual values of the data points. Finally, substitute the values of TSS and RSS into the R^2 formula and subtract the result from 1 to get the coefficient of determination.

The value of R^2 can range from less than 0 to 1, where 1 indicates the model fits the data without any differences between the predicted value and ground truth value. However, when R^2 is lower than 0, it indicates that the model did not capture or learn any relationship between the features and the target variable. A negative R^2 is possible if the ratio of TSS and RSS is greater than 1.

3.6.4 Spearman's Rank Correlation

Spearman's rank correlation is a correlation metric that is used for nonparametric measurement between two data samples (Dodge, 2008, p. 502). The Spearman's correlation formula takes the values of X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_n in two sample sets X with sizes n and takes into account the ranks of each value ($R[X_i]$) and ($R[Y_i]$). The general formula of Spearman's correlation (ρ) is as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\text{where } d_i = R[X_i] - R[Y_i]$$

The scores of Spearman's correlation coefficients range from -1 to +1. Values closer to -1 or +1 represent stronger relationships than values closer to 0. A value of 0 indicates no association. A coefficient of +1 represents a strong positive association, meaning both scores increase together, while a coefficient of -1 indicates a strong negative association, meaning as one score increases, the other decreases.

Chapter 4

Methodology

The methodology details the specific procedures necessary to achieve the objectives of this study. Figure 4.1 presents the overall research pipeline that will be followed throughout the research process.

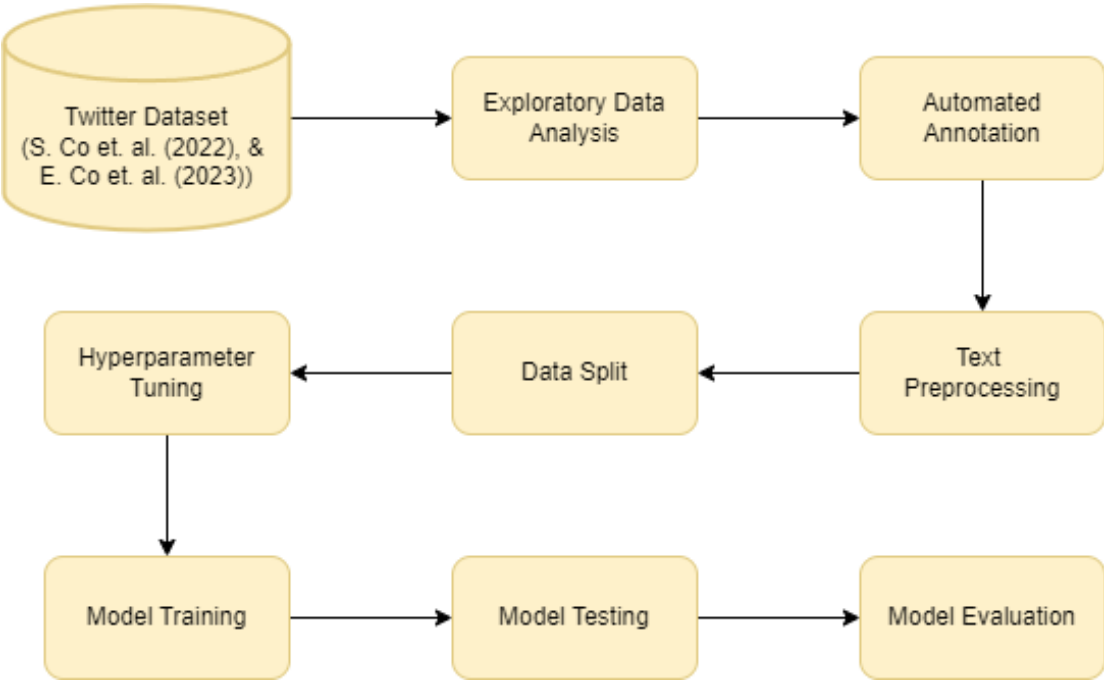


Figure 4.1: Diagram of Research Pipeline

The dataset used in this study is a collection of tweets from the Philippines, collected at two different periods. An exploratory data analysis (EDA) will be

performed to better understand the dataset’s characteristics, including the number of tweets containing emojis and the count of unique tweets. Following the EDA, we will apply an annotation scheme that leverages information from an emoji-based sentiment lexicon to assign sentiment scores to tweets containing emojis found within the lexicon. Subsequently, the data will undergo relevant text preprocessing strategies for transformer-based models. Once text preprocessing is complete, the data will be divided into three sets: training, validation, and testing. Each pre-trained transformer model will utilize the training set for prediction, the validation set for fine-tuning, and the testing set for evaluating the model’s performance. These steps will be elaborated upon in the following sections.

4.1 Data Source

The datasets used in this study are sourced from two private collections of social media posts from X (formerly Twitter). The first dataset was collected between March and December 2020 as part of the research by S. Co et al. (2022). The second dataset was gathered from December 2021 to October 2022 for the study by E. Co et al. (2023). Both datasets consist of tweets from users in the Philippines, sourced through the Twitter API. The geographical boundaries for data collection were specified using a bounding box with coordinates 117.17427453, 5.58100332277, 126.537423944, and 18.5052273625, ensuring that all tweets originated within the country.

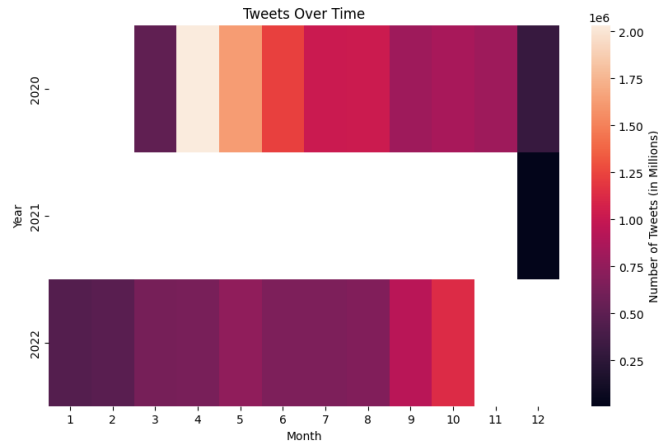


Figure 4.2: Temporal Distribution of Collected Tweets from March 2020 to October 2022

The datasets include 17 million tweets, of which 6.2 million contain emojis. There are 3,455 unique emojis in the dataset. Additionally, each entry in the

dataset includes a unique identifier for the tweet (Tweet ID), along with meta-data such as the date and time the tweet was posted, the language of the tweet (Automatically Generated by Twitter), and the unique identifier associated with the tweet’s author (Author ID).

Table 4.1: Language Distribution in Datasets by S. Co et al. (2022) and E. Co et al. (2023)

Language	Count	Percentage(%)
Tagalog	9,277,800	54.00%
English	5,049,291	29.39%
Undefined	1,174,028	6.83%
Others	1,055,124	6.15%
Indonesian	479,601	2.79%
Spanish	145,027	0.84%

The datasets includes 66 distinct languages, including Tagalog, English, Indonesian, Spanish, and Italian. The specific language distributions are presented in Table 4.1. The data indicates that Tagalog and English are the dominant languages, accounting for a combined total of 83.39%. Additionally, the tweets classified under undefined languages are those the Twitter API could not identify. This may occur if the tweets consist solely of numerical values or elements (e.g., links) that do not carry distinct linguistic meaning.

Figure 4.2 shows the number of tweets collected over time from March 2020 to October 2022, ranging from 0.25, representing the lowest amount of data collected, to 2.00, representing the highest. April 2020 had the highest data collection of all the months, likely due to the impact of the COVID-19 pandemic, during which social media was widely used.

It is important to note that social media data is inherently noisy and sensitive to external factors such as significant events or trends. Tweets collected across a wide timeframe may reflect varying patterns of activity and sentiment which are influenced by the context of their respective periods. Additionally, the data collection process is dependent on the Twitter API, which dictates the available tweets during the specified collection periods.

Regarding ethical considerations, direct consent from users was not obtained, and the use of the official Twitter API for data collection aligned with Twitter’s developer agreement and policy, which permitted data gathering from the platform under these terms. For additional information on the ethical considerations related to the usage of the datasets, please consult Appendix A.

4.2 Automated Annotation

For this study, the Multidimensional Lexicon of Emojis (MLE) by Godard and Holtzman (2022) will be utilized to automate the labeling of sentiment scores. MLE consists of 359 commonly used emojis, each rated across ten affect scores, including eight emotional dimensions (Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, and Trust) and two sentiment dimensions (Positive and Negative). This study will only focus on the two sentiment dimensions for automatic sentiment annotation.

The annotation process involves calculating sentiment scores from emojis present in tweets. Only tweets containing at least one emoji that is a part of the MLE are included in this automatic annotation. Tweets lacking at least one emoji from the MLE will be discarded. For each tweet, positive and negative sentiment scores are determined by identifying its emojis and referencing their respective positive and negative sentiment scores from the MLE. If a tweet contains multiple emojis, the individual positive and negative scores are averaged separately, resulting in two distinct final scores, one positive and one negative, that represent the sentiment of a tweet. This procedure is applied across all tweets in the datasets.

Algorithm 1 Automatic Annotation with MLE Sentiment Scores

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  $\leftarrow$  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  $\leftarrow$  positive_sum / emoji_count
    negative_score  $\leftarrow$  negative_sum / emoji_count
    assign positive_score and negative_score to tweet
  else
    discard tweet
  end if
end for
```

4.3 Text Preprocessing

Text preprocessing will be applied to the datasets in preparation for model training. Raw dataset entries are simplified by removing non-essential metadata, leaving only the tweet text. Tweets without a text or lacking at least one emoji from the MLE will also be discarded. Duplicate text entries are also discarded. After these processes, tweets were reduced from 17 million to 5.46 million.

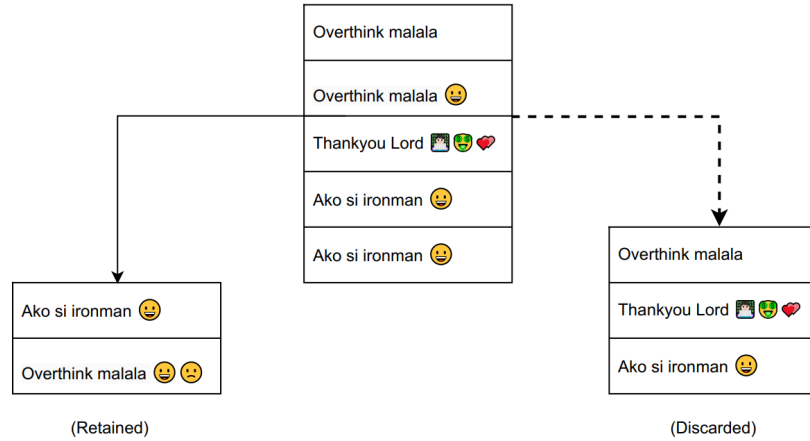


Figure 4.3: Text preprocessing on the retention and exclusion of tweets. Excluded tweets include tweets with no emojis, tweets with unrecognized emojis (based on the MLE), and duplicate tweets.

For the remaining data, excess whitespace, including extra spaces, multiple newline characters, and repeated tab spaces, will be reduced to single instances. Context preservation is important for transformer-based models. Hence, casing, punctuation, and other Unicode symbols will be retained.

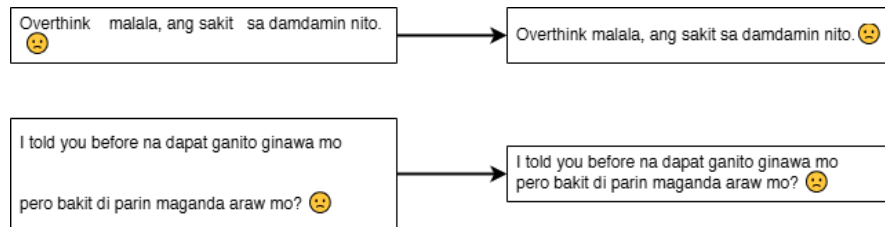


Figure 4.4: Text preprocessing on reducing excess whitespace. Examples include the reduction of multiple spaces to a single instance (top) and the reduction of multiple newline character sequences to a single instance (bottom)

Additionally, token collapsing will be employed. The purpose of this approach

is twofold: to handle the most common noisy social media elements while preserving their structural role in the text and to address ethical considerations by anonymizing sensitive information. Following Cruz and Cheng (2020), links will be replaced with a [LINK] token, mentions (tokens starting with "@"") will be collapsed into a [MENTION] token, and hashtags (tokens starting with "#") will be collapsed into a [HASHTAG] token.

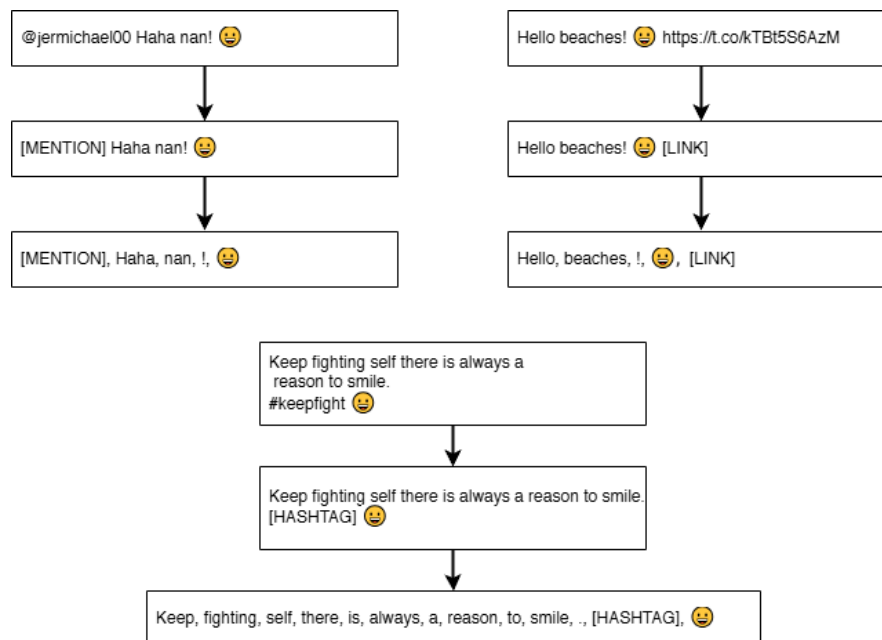


Figure 4.5: Text preprocessing on token collapsing. Examples include user mentions or handles replaced with the [MENTION] token (top left), links replaced with the [LINK] token (top right), and hashtags replaced with the [HASHTAG] token (bottom)

We will be utilizing tokenizers provided by the different transformer models. Tokenizers may differ in their use of tokenization algorithms. As such, the specific tokenizers that will be used in this study are detailed in section 4.5.

4.4 Neural Network Architecture

The neural network architecture for the sentiment regression task is based on the fine-tuning of pre-trained transformer models. Following the tokenization process, tokenized inputs are converted into embedding vectors, including positional, token, and segment embeddings, before passing through the transformer layers.

At the output level, transformers generate contextualized embeddings for each token in the sequence. The final hidden state of the [CLS] token embedding, which represents the entire sequence for downstream sentiment analysis task, is extracted as the final representation of the sequence. Following the transformer architecture, our implementation will add a regression model on top of the [CLS] token embedding the proposed transformers to perform sentiment regression. Figure 4.6 illustrates the proposed sentiment regression model.

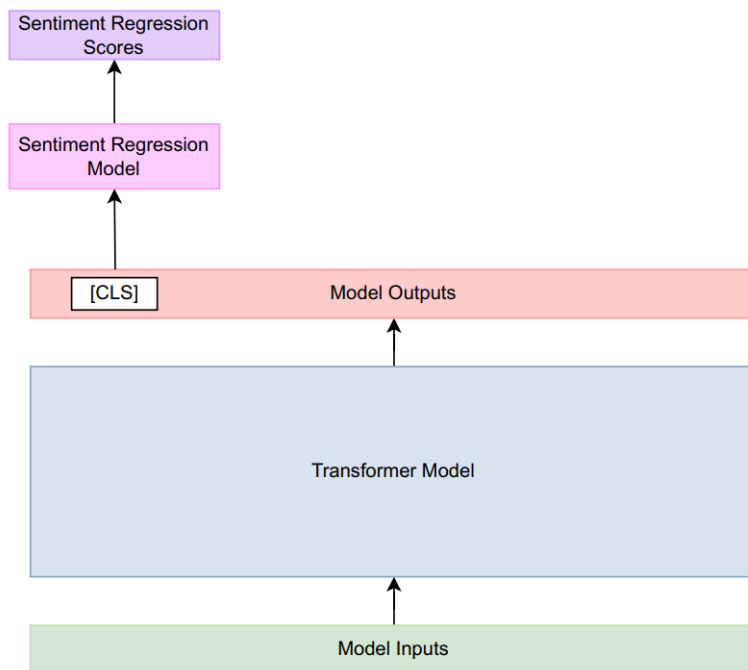


Figure 4.6: **Proposed Neural Network Architecture for Sentiment Regression**

Specific transformer models that will be used in this study are introduced in section 4.5.

4.5 Experimental Setup

This experiment explores the application of transformer models for sentiment regression. Our approach involves testing these models in their pre-trained form and with additional fine-tuning specifically for sentiment prediction. Experiments also include hyperparameter tuning of the neural network to arrive at the best neural network configuration.

Fine-tuning is employed in place of full pre-training to adapt these models for

sentiment regression. Fine-tuning is employed as it offers a computationally inexpensive alternative to pre-training, which requires extensive data and resources (Devlin et al., 2019).

4.5.1 Data Split

The final data split strategy will depend on the usable data after preprocessing. However, we are currently exploring two possible configurations: a 90-10 split or an 80-10-10 split among training, validation, and testing subsets.

Once automatic annotation is performed, an additional exploratory data analysis (EDA) will be conducted to examine the distribution of sentiment scores/labels in the corpus. If the analysis reveals a significant imbalance or skewness in the data, data balancing techniques, such as oversampling or undersampling, will be employed.

4.5.2 Transformer Models

The study plans to make use of the following transformer models:

- **RoBERTa-Tagalog**, developed by Cruz and Cheng (2022), is a transformer optimized for the Filipino language using the RoBERTa pretraining approach. The model was trained on the TLUnified dataset, a diverse corpus constructed to improve upon existing Filipino language resources. This dataset combines text from bilingual sources, news, and other web-crawled Filipino content.
- **SiEBERT**, or “Sentiment in English BERT,” (Hartmann et al., 2023) is a pre-trained language model developed for binary sentiment analysis. SiEBERT is a fine-tuned checkpoint of RoBERTa across a diverse range of English-language text types, including reviews and social media posts. SiEBERT was fine-tuned on 15 sentiment-labeled datasets from various sources through a leave-one-out evaluation (14 datasets for training; 1 for testing), which only resulted in a minor performance drop of 3 percentage points on average.
- **mBERT**, or multilingual BERT (Devlin et al., 2019), is a BERT derivative pre-trained on 104 languages using Wikipedia data. Its architecture mirrors the original BERT base model, with 12 transformer layers, 12 attention heads, and a hidden layer size of 768.

- **Tagalog-DistilBERT** (Cruz & Cheng, 2020) is a distilled version Tagalog-BERT transformer model, which is pre-trained on the WikiTextTL-39 corpus for the Tagalog language. Tagalog-BERT has 12 transformer layers, 12 attention heads, and a hidden layer size of 768.
- **TwHIN-BERT** (Zhang et al., 2023) is a socially-enriched, multilingual language model tailored to handle noisy, user-generated content on Twitter. TwHIN-BERT is trained on 7 billion tweets in over 100 languages. The model employs a Twitter Heterogeneous Information Network (TwHIN), which integrates engagement data such as likes, replies, and retweets.

The following tokenizers (AutoTokenizers from HuggingFace) will be utilized:

- For **mBERT** and **Tagalog-DistilBERT**, the **BERT tokenizer**, which is based on WordPiece, is utilized. Key tokens used by the BERT tokenizer include the [CLS] token at the beginning of each sequence, which captures the sequence-level embedding, and the [SEP] token for separating multiple segments within an input.
- The **RoBERTa** tokenizer is applied in both **RoBERTa** and **SiEBERT**. It makes use of Byte-Pair Encoding (BPE). The RoBERTa tokenizer does not include the [CLS] token. Instead, it uses <s> as a beginning-of-sequence token and </s> as both the end-of-sequence token and the segment separator.
- The **TwHIN-BERT** model employs the **XLM-RoBERTa tokenizer**, based on SentencePiece. The XLM-RoBERTa tokenizer is particularly suited for multilingual applications, as it uses a pre-built vocabulary that captures various linguistic structures across languages. Similar to RoBERTa, this tokenizer uses <s> and </s> tokens for sequence delimitations.

4.5.3 Sentiment Prediction Module

After passing through the transformer module, the output undergoes further processing through a fully connected linear layer. The output of the sentiment prediction module directly aligns with the sentiment scores from the Multidimensional Lexicon of Emojis (MLE) without additional transformations or scaling.

For the prediction architecture, we consider two potential configurations. The first option employs a transformer model with a multi-output setup. This architecture predicts two values simultaneously: one for the positive sentiment score

and one for the negative sentiment score. Alternatively, the second option involves training two separate instances of the same transformer model architecture: one dedicated to predicting the positive sentiment score and another for the negative sentiment score. Both configurations will be explored to identify the most effective approach for this sentiment regression task.

4.5.4 Hyperparameter Tuning

This experiment uses hyperparameter tuning to identify the optimal values for the sentiment regression task. Building on Devlin et al. (2019) recommendations for fine-tuning tasks, batch sizes of 16 and 32 will be tested alongside epoch counts of 2, 3, and 4. The learning rate is chosen following insights from Sun et al. (2019), which highlights the importance of a lower learning rate of $2e-5$ to mitigate catastrophic forgetting. Catastrophic forgetting is a common issue in transfer learning, where pre-trained knowledge is overwritten during training.

The feasibility of testing all combinations remains dependent on computational capacity and experimental timelines.

4.5.5 Model Evaluation

The evaluation metrics to be used for comparing the performances of the aforementioned models include Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2). These two metrics are often used when comparing and evaluating the performance and output of regression models (Tatachar, 2021). Additionally, the study will make use of the Mean Squared Error (MSE) as the loss function for sentiment regression tasks.

There will also be correlation measurement using Spearman’s Correlation performed with each pre-trained model with VADER (Hutto & Gilbert, 2014), a Rule-based Model for Sentiment Analysis of Social Media Text found within the NLTK Library, to validate the effectiveness of sentiment prediction.

4.5.6 Experiment Summary

This study will experiment with both finetuned and non-finetuned regression models. Non-fine-tuned models will utilize transformer models in their pre-trained form, where the transformer layers are frozen, and only the regression head is

updated through backpropagation. On the other hand, fine-tuned models will have unfrozen transformer layers. This allows the transformer and regression head weights to be updated during training.

Both approaches will use the same configurations for the sentiment prediction module (single-output and multi-output) setups. The models developed will undergo hyperparameter tuning, where we will look at the most appropriate batch size, epoch count, and learning rate.

Appendix A

Research Ethics Documents

A.1 Research Ethics Clearance Form

RESEARCH ETHICS CLEARANCE FORM For Thesis Proposals ¹	
Names of student researcher/s :	CALALANG, Hosea ESTANOL, Miguel JACINTO, Jon LOPEZ, Mauries
College:	College of Computer Studies
Department:	Software Technology Department
Research Title:	Exploring Transformer-based Approaches in Sentiment Prediction of English-Filipino Text Data
Course:	BS Computer Science
Expected duration of project:	from: AY24-25 Term 1 to: AY24-25 Term 3
<p>Ethical considerations</p> <p>The study will continue using a pre-existing dataset, similar to the studies of S. Co et al. (2022) and E. Co et al. (2023). The dataset was collected from Twitter via the official Twitter API by Mr. Edward P. Tighe, in adherence to the developer agreement and policy during the time of collection. The focus of the data collection was primarily on the tweets rather than the users, with the users being the authors of the tweets. Consent was not directly obtained from the users. However, in accordance with Twitter's developer agreement and policy during the time of collection, data collection from users on the platform was not a problem when using the official Twitter API.</p> <p>Furthermore, the dataset is not publicly accessible. While it can be accessed through the official Twitter API, users must create an account and submit a request for API access before being able to retrieve the data. For further understanding of tweet availability, Twitter's privacy policy is available for reference.</p> <p>The dataset consists of 17 million tweets generated within the Philippines, identified using a bounding box to confirm their geolocation within the country. This dataset will be used exclusively for research purposes, specifically for sentiment prediction using pre-trained transformer models, and will not be made publicly available. However, the best-performing sentiment prediction model will be released to the public to ensure transparency within the research community, allowing researchers to study the model. The public, however, is not advised to use it as a tool for sentiment prediction. The model that is going to be release will be in an experimental phase and is not suitable for reliable sentiment analysis. The data used in the transformer models for sentiment prediction will remain confidential and will only be utilized within the scope of this research.</p> <p>No sensitive data will be disclosed in the paper, and data anonymization will be applied where possible, such as withholding User IDs and handles. Additionally, no new data will be collected. Furthermore, permission to use the dataset has been granted by Mr. Edward P. Tighe.</p>	

To the best of our knowledge, the ethical issues listed above have been addressed in the research.

TIGHE, Edward P.

Name and signature of adviser/mentor

Date:

Name and signature of panelist

Date:

Name and signature of panelist

Date:

A.2 General Research Ethics Checklist

 De La Salle University	<h2>Research Ethics Review Committee</h2> <p>Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2(D)
		Version No.: 1
		Version Date: July 2016

DE LA SALLE UNIVERSITY General Research Ethics Checklist
<p><i>This checklist is to ensure that the research conducted by the faculty members and students of De La Salle University is carried out according to the guiding principles outlined in the Code of Research Ethics of the University. The investigator is advised to refer to the <u>De La Salle University Code of Research Ethics and Guide to Responsible Conduct of Research</u> before completing this checklist. Statements pertinent to ethical issues in research should be addressed below. The checklist will help the researcher/s and advisers/readers/evaluators determine whether procedures should be undertaken during the course of the research to maintain ethical standards. The University's <u>Guide to the Responsible Conduct of Research</u> provides details on these appropriate procedures.</i></p>

Researcher Details	
Students	CALALANG, Hosea ESTANOL, Miguel JACINTO, Jon LOPEZ, Mauries
Thesis Adviser	Mr. Edward P. Tighe
Department/College	Department of Software Technology / College of Computer Studies
Proposed Title of the Research	Exploring Transformer-based Approaches in Sentiment Prediction of English-Filipino Text Data
Term(s) and academic year in which research project is to be undertaken	Term 1 of AY 24-25 to Term 3 of AY 24-25

<p><i>This checklist must be completed AFTER the De La Salle University Code of Ethics has been read and BEFORE gathering data.</i></p>		
Questions	Yes	No
1. Does your research involve human participants (this includes new data gathered or using pre-existing data)? If your answer is yes , please answer Checklist A (Human Participants) . Please specify if the kind of research you will be conducting falls under any of the following Human Participants sub-categories:	✓	
1.A. Will you be conducting Action Research in an existing business, company, or school? If your answer is yes , please answer Checklist F (Action Research) .		✓

 De La Salle University	<h2>Research Ethics Review Committee</h2> <p>Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2(D)
		Version No.: 1
		Version Date: July 2016

1.B. Does your research involve online communities (this includes culling data from social media platforms, online forums and blogs)? If your answer is yes , please answer Checklist G (Internet Research) .	✓	
1.C. Does your research involve human participants who are situated in a community and may necessitate permission to acquire access to them? If your answer is yes , please answer Checklist H (Community Research) .		✓
2. Will your research make use of documents which are not in the public domain and, thus, require permission for use from the custodian of such documents? If YES, please provide certification that permission from the custodian of the data was sought and granted.		✓
3. Will your research make use of secondary data (e.g., surveys, inventories, plans, official documents, etc.) from an institution, organization, or agency, which are not in the public domain and, thus, require permission for use from the custodian of such documents? If YES, please provide certification that permission to use the data was sought from the institution, organization, or agency and approval was granted.		✓
4. Does your research involve animals (non-human subjects)? If your answer is yes , please answer Checklist B (Animal Subjects) .		✓
5. Does your research involve Wildlife? If your answer is yes , please answer Checklist C (Wildlife) .		✓
6. Does your research involve microorganisms that are infectious, disease causing or harmful to health? If your answer is yes , please answer Checklist D (Infectious Agents) .		✓
7. Does your research involve toxic/chemicals/ substances/materials? If your answer is yes , please answer Checklist E (Toxic Agents) .		✓

Research with Ethical Issues to address:

If you have a YES answer to any of the above categories, you will be required to complete a detailed checklist for that particular category. A YES answer does not mean the disapproval of your research proposal. By providing you with a more detailed checklist, we ensure that the ethical concerns are identified so these can be addressed in adherence to the University Code of Ethics.

 De La Salle University	<h2 style="text-align: center;">Research Ethics Review Committee</h2> <p style="text-align: center;">Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2(D)
		Version No.: 1
		Version Date: July 2016

Declaration of Conflict of Interest

☒ 1. I do not have a conflict of interest in any form (personal, financial, proprietary, or professional) with the sponsor/grant-giving organization, the study, the co-investigators/personnel, or the site.

☐ 2. I do have a conflict of interest, specifically:

☐ A. I have a personal/family or professional interest in the results of the study (family members who are co-proponents or personnel in the study, membership in relevant professional associations/organizations).

Please describe the personal/family or professional interest:

☐ B. I have proprietary interest vested in this proposal (with the intent to apply for a patent, trademark, copyright, or license)

Please describe proprietary interest:

☐ C. I have significant financial interest vested in this proposal (remuneration that exceeds P250,000.00 each year or equity interest in the form of stock, stock options or other ownership interests).

Please describe financial interest:

A.3 Research Checklist A - Human Participants

 De La Salle University	Research Ethics Review Committee Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513	SOP No.: 2
		Form No.: 2.03
		Version No.: 1
		Effectivity Date: July 2016

DE LA SALLE UNIVERSITY Checklist A Research Ethics Checklist for Investigations involving Human Participants

This checklist must be completed AFTER the De La Salle University Code of Research Ethics and Guide to Responsible Conduct of Research has been read and BEFORE gathering data. The University Code of Research Ethics is available at http://www.dlsu.edu.ph/offices/urco/forms/URCO-Code-of-Research-Ethics_August2011.pdf

NOTE: This checklist is completed after the research proponent fills out the General Checklist Form.

Only answer this Checklist if you answered YES on question 1 of the General Checklist.

Researcher Details	
Students	CALALANG, Hosea ESTANOL, Miguel JACINTO, Jon LOPEZ, Mauries
Thesis Adviser	Mr. Edward P. Tighe
Department/College	Department of Software Technology / College of Computer Studies
Proposed Title of the Research	Exploring Transformer-based Approaches in Sentiment Prediction of English-Filipino Text Data
Term(s) and academic year in which research project is to be undertaken	Term 1 of AY 24-25 to Term 3 of AY 24-25

Provide a brief description of the data collection procedure to be undertaken in the research:

The tweets were gathered using Python with the official Twitter API, utilizing a bounding box to ensure that they were generated within the Philippines. The data was scraped between March 2020 and December 2020 and stored as unstructured data in CSV files. Each tweet includes metadata such as the Tweet ID, date of the tweet, location, language, and User ID of the tweet's author. These details will not be disclosed or released to the public, and data anonymization will be applied where possible.

 De La Salle University	<h2>Research Ethics Review Committee</h2> <p>Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2.03
		Version No.: 1
		Effectivity Date: July 2016

The following should be attached to the checklist:

- A copy of the informed consent form to be used in the study.
- A copy of the instrument/tool that will be administered to the participants.
- If applicable, a copy of the letter seeking permission to collect data from participants who are under the supervision of an agency, institution, department, or office.
- If applicable, a copy of the parental consent form for participants below 18 years old.

The following items refer to important ethical considerations in the conduct of research with human participants. Provide a check for the appropriate answer to each question.

Source of data

Please check all that apply:

✓	1. New data will be collected from human participants	
	If you checked this item, how will the new data be gathered? Please check all that apply.	
	After answering this question, please proceed to page 3	
	<input type="checkbox"/>	Experimental Procedures/Intervention/ Treatments
	<input type="checkbox"/>	Focus Group
	<input type="checkbox"/>	Personal Interviews
	<input type="checkbox"/>	Self-administered Questionnaire
	<input type="checkbox"/>	Researcher-administered Questionnaire
	<input type="checkbox"/>	Internet survey
<input type="checkbox"/>	Observation	
<input type="checkbox"/>	Telephone survey	
	✓	Others, please specify: Official Twitter API
✓	2. Pre-existing data from human participants, i.e., from a dataset	
If you checked this item, please proceed to page 7		

If both options are checked (both new data and pre-existing data), answer all of the questions in this document.

Only answer if new data will be collected (item 1 above)

Sampling Details

Number of Participants/Subjects	
Location where the participants will be recruited/ where subjects will be obtained?	Online and through recommended Experts
How long will the data collection take place?	

 De La Salle University	<h2>Research Ethics Review Committee</h2> <p>Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2.03
		Version No.: 1
		Effectivity Date: July 2016

Who will perform the data collection?	
Location(s) where data collection will take place	
What procedures will be employed to ensure voluntary consent from participants?	
Data Retention	
How long will data with participant identifiers be kept after the publication of the first paper from the project?	
How long will anonymized data be kept after the publication of the first paper from the project?	
Procedure for Informed Consent	
How will informed consent be recorded? (check all that applies)	<input type="checkbox"/> Written Consent <input type="checkbox"/> Audio-recorded Consent <input type="checkbox"/> Online/Email recorded Consent <input type="checkbox"/> Others, please specify:
Reminder: please attach informed consent that will be used in the study	

If you will not obtain a recorded informed consent, answer the questions that follow:

Why does the waiver of informed consent not pose a threat to the welfare and rights of the participants?

Why is recording an informed consent not practical for the proposed study?

	Yes	No	Not Applicable
1. Will the research involve students who will be receiving course credits for their participation? If YES, please attach a copy of the consent form and a summary of the debriefing process that will help participants			

 De La Salle University	<h2>Research Ethics Review Committee</h2> <p>Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2.03
		Version No.: 1
		Effectivity Date: July 2016

<p>understand how their participation in the research has provided a relevant learning experience to the crediting course.</p>			
<p>2. Does the study involve participants below 18 years old or those who are unable to give their informed consent?</p> <p>If YES, please attach a copy of the parental consent form.</p>			
<p>3. Is there a possibility that the research can induce physical and/or psychological harm to the participants? Will they experience pain or some discomfort as a result from their participation in the research?</p> <p>If YES, please attach an acceptable argument that outlines the benefits of doing the research and how they outweigh the cost of harming the participants.</p>			
<p>4. Will the participants be deliberately falsely informed or made unaware that they are being observed? Will they be misled in a way that they will possibly object to or show unease when told of the real purpose of the study?</p> <p>YES, please attach an acceptable argument that outlines the benefits of doing the research and how they outweigh the cost of harming the participants.</p>			
<p>5. Will the research involve the discussion of, or questions on, sensitive topics (e.g. sexual activity, substance abuse, or mental health)?</p> <p>If YES, please make sure that the informed consent form explicitly states that sensitive questions will be posed and that you will safeguard the anonymity of the participants and ensure confidentiality. Please attach a copy of your informed consent form and your instrument.</p>			
	Yes	No	Not Applicable

 De La Salle University	<h2>Research Ethics Review Committee</h2> <p>Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2.03
		Version No.: 1
		Effectivity Date: July 2016

<p>6. Will the research involve the administration of drugs, or other substances to the participants?</p> <p>YES, please attach an acceptable argument that outlines the benefits of doing the research and how they outweigh the cost of harming the participants.</p> <p>Please also attach a description of the procedure that will ensure that the participants will be brought back to their physical and psychological states prior to their participation in the research.</p>			
<p>7. Will biological samples (e.g. blood, saliva, urine) be obtained from the participants?</p> <p>If YES, will this involve invasive procedures? Please attach a description of these procedures.</p>			
<p>8. Will genetic materials be obtained from the biological samples?</p> <p>If YES, please attach a description of the procedures that will ensure confidentiality. Please attach the informed consent form.</p>			
<p>9. Will financial inducements (other than reasonable expenses, like transportation or meal allowances) be offered to the participants for their participation in their research?</p> <p>If YES, the researcher(s) should be mindful of how the inducements can influence the participants' responses or behaviors during the research. Indicate the financial inducements offered to the participants:</p> <p>_____</p>			
<p>10. Is there a possibility for groups or communities to be harmed by the dissemination of the research findings?</p> <p>If YES, please attach a description of procedures to ensure the anonymity and confidentiality of the research findings.</p>			

 De La Salle University	Research Ethics Review Committee Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513	SOP No.: 2
		Form No.: 2.03
		Version No.: 1
		Effectivity Date: July 2016

11. Will the results of this study have a commercial value? If yes, do you intend to apply for a patent for the output of this research? Please check: _____ Yes _____ No			
--	--	--	--


FOR PROPONENTS WHO WILL GATHER NEW DATA ONLY, PLEASE STOP ANSWERING.

Use of Pre-existing Data collected from Human Participants		
Indicate the dataset from which the data for the study will be sourced	The dataset from the improvements of S. Co et al. (2022) and E. Co et al. (2023)	
Is the data publicly available, i.e., the access to which does not necessitate an approval process?		Yes Please indicate where the dataset is available:
	✓	No Please indicate/attach the approval authority for access:
Was the original dataset originally collected for the present study's purpose?	✓	Yes Please attach the Consent Form used in the original study.
		No Please attach the Information Collection Statement (i.e., the statement given to informants providing them with the rationale for the collection of specific information).
Does the original data set contain sensitive data, that is information that an individual would not likely want to be disclosed publicly, e.g., data on sexual activities, substance use?	✓	Yes Please describe the type of sensitive data to be used in the present research: (1) Personalized Tweets (2) Username
		No
		No (This means that neither the researcher nor the participant provided any personal identifiers)

 De La Salle University	<h2>Research Ethics Review Committee</h2> <p>Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2.03
		Version No.: 1
		Effectivity Date: July 2016

Does the original dataset have personal identifiers?	✓	Yes, specifically: _____ Direct (<i>i.e., the participant provided personal details like name and address</i>) ___ ✓ ___ Indirect (<i>i.e., the participant was given a respondent code to make the participant identifiable</i>)
Will new data be collected and analyzed along with data from the existing dataset?		Yes Please answer questions on page 3-5.
	✓	No

A.4 Research Checklist G - Internet

 De La Salle University	Research Ethics Review Committee Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513	SOP No.: 2
		Form No.: 2(K)
		Version No.: 1
		Version Date: March 2017

DE LA SALLE UNIVERSITY

Checklist G

Research Ethics Checklist for Investigators conducting Internet Research

This checklist must be completed AFTER the De La Salle University Code of Research Ethics and Guide to Responsible Conduct of Research has been read and BEFORE gathering data. The University Code of Research Ethics is available at http://www.dlsu.edu.ph/offices/urco/forms/URCO-Code-of-Research-Ethics_August2011.pdf

NOTE: This checklist is completed after the research proponent fills out the General Checklist Form and Checklist A.

Only answer this Checklist if you are conducting Internet Research (research on online communities).

Which of the following online data will you be using in your research? Check all that apply:


- ☒ Social Media Platform (e.g. Twitter, Facebook, 9gag)
- ☐ Blogs & Forum including Comments (e.g. BoredPanda, Reddit)
- ☐ E-mails & Chats
- ☐ Video Blogs (e.g. YouTube)
- ☐ Collaborative (e.g. Wikipedia)
- ☐ Websites (e.g. News, Media, Company, E-Commerce, Government, etc.)
- ☐ Online Recruitment Platform (e.g. Mechanical Turk, Freelancer, E-lance, O-Desk)
- ☐ Open Source Websites (e.g. GitHub)
- ☐ Others: _____

Where will you source your data?

- ☐ New
- ☐ Existing
- ☐ Combination

What type of data will be collected?


- ☒ Text
- ☐ Audio
- ☐ Video/Film
- ☐ Photo
- ☒ Metadata (e.g. Profile, Geographic Location, Tags)

 De La Salle University	Research Ethics Review Committee Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513	SOP No.: 2
		Form No.: 2(K)
		Version No.: 1
		Version Date: March 2017

<input type="checkbox"/> Presentations (e.g. downloaded PowerPoint or Keynote presentations) <input type="checkbox"/> Contents of an application such as input, output, log files for analysis software, simulation software, schemas <input type="checkbox"/> Correspondence, including electronic mail
--


<p>What is the period coverage of data collection? (indicate in years and months)</p> <p>_____</p> <p>How many participants will you collect data from?</p> <p>_____</p> <p>What are all the websites you will source your data from? Please list all URLs: x.com (<i>Previously known as "Twitter"</i>)</p> <p style="text-align: center;"><i>Please see "Revised Research Ethics Clearance Form" for further details.</i></p>

<p>How will informed consent be obtained? Please check.</p> <p> <input type="checkbox"/> Electronic Information Sheet with "check box" for consent <input type="checkbox"/> E-mail with name <input type="checkbox"/> Consent implied through submitting information <input type="checkbox"/> Written Consent <input type="checkbox"/> Audio-recorded Consent <input checked="" type="checkbox"/> Others: _____ </p> <p>How will the participants obtain a copy of the informed consent form? Please check.</p> <p> <input type="checkbox"/> Hard copy <input type="checkbox"/> Online copy (If online copies will be given, the users should be able to click on the Print button incorporated in the information page, or an e-mail should be sent to the participant containing the agreement form.) </p> <p style="text-align: center;"><i>Please see "Revised Research Ethics Clearance Form" for further details.</i></p>


 De La Salle University	<h2 style="text-align: center;">Research Ethics Review Committee</h2> <p style="text-align: center;">Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2(K)
		Version No.: 1
		Version Date: March 2017

If a participant wishes to leave the study and wishes to withdraw the submitted information, describe the steps a participant must take to withdraw from the study in the research proposal and in the informed consent form.

	Yes	No	Not Applicable
1. Is the data you are planning to gather publicly available? <i>Please see "Revised Research Ethics Clearance Form" for further details.</i> If NO...attach a letter of support from the website or server owner/moderator indicating approval to use this for data gathering		✓	
2. Will the participants be compensated for participating? If YES...indicate the type of compensation to be provided and provide information on how appropriate and just compensation (proportionate to the contribution in the research, research budget, and local conditions) will be provided.		✓	
3. Will you have minors as participants in your study? Minors are individuals under the age of 18 years old. If YES...Obtain parental/guardian consent and participant assent to participate in your study. Attach the parental consent and assent forms to your proposal. The consent forms should indicate the measures you will undertake to ensure confidentiality and protect the participants. The electronic versions of the consent forms are acceptable.			✓
4. Will data collection involve students? If YES... let the participant sign an informed consent form that assures that his/her academic status will not be affected by participation or non-participation in the study.			✓

 De La Salle University	<h2 style="text-align: center;">Research Ethics Review Committee</h2> <p style="text-align: center;">Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513</p>	SOP No.: 2
		Form No.: 2(K)
		Version No.: 1
		Version Date: March 2017

5. Will data collection involve persons who belong to a vulnerable group (PWDs, minorities, abuse victims, students, etc.) If YES... ensure that the participants are briefed about the study and the data gathering protocols, and will be provided special assistance should they require it. Please submit a corresponding detailed research protocol.			✓
	Yes	No	Not Applicable
6. Will data be collected using an automated system? (e.g. hardware, software, or meta search engines) <i>Please see "Revised Research Ethics Clearance Form" for further details.</i> If YES... provide details on the automation system (web crawler, etc.)	✓		
7. Will there be information about the study that will be deliberately withheld from the participants? If YES... what information will be withheld, what are the benefits for doing so, and how will debriefing be done?		✓	
8. Will your data be stored after this study? If YES, please indicate in the research proposal how long data will be stored, and how it will be stored. Include this information in the informed consent form and provide a plan for data confidentiality.		✓	
9. Will the data be made available for future research? If YES... indicate in the informed consent form that the results of the study will be used for future studies. Specify that the identity of the participant will be anonymized in future studies. Provide participants with the option to allow the results to be used for further studies and the option to withdraw their results from future studies.		✓	
10. Will you anonymize the participants in your data?	✓		

 De La Salle University	<h2 style="text-align: center;">Research Ethics Review Committee</h2> <p style="text-align: center;"> Research Ethics Office, 3F Henry Sy Sr. Hall De La Salle University Manila 2401 Taft Avenue, Manila 1004, Philippines REO@dlsu.edu.ph (632) 524-4611 loc. 513 </p>	SOP No.: 2
		Form No.: 2(K)
		Version No.: 1
		Version Date: March 2017

<p><i>Please see "Revised Research Ethics Clearance Form" for further details.</i></p> <p>If YES...provide details on how you will assure the confidentiality and anonymity of the participants.</p>			
<p>11. Will the results of this study have a commercial value?</p> <p>If yes, do you intend to apply for a patent for the output of this research? Please check:</p> <p>_____ Yes</p> <p>_____ No</p>		✓	

References

- Abimbola, B., Tan, Q., & De, E. A. (2024, Jun). Sentiment analysis of canadian maritime case law: a sentiment case law and deep learning approach. *International journal of information technology*. Retrieved from <https://link.springer.com/article/10.1007/s41870-024-01820-2> doi: <https://doi.org/10.1007/s41870-024-01820-2>
- 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).
- Aliyu, Y., bt Sarlan, A., Danyaro, K. U., Rahman, A. S. B. A., & Abdullahi, M. (2024). Sentiment analysis in low-resource settings: A comprehensive review of approaches, languages, and data sources. *IEEE Access*, *12*, 66883–66909. Retrieved from <https://api.semanticscholar.org/CorpusID:269673346>
- Aliyu, Y., Sarlan, A., Danyaro, K. U., & Rahman, A. S. (2024). Comparative analysis of transformer models for sentiment analysis in low-resource languages. *International Journal of Advanced Computer Science & Applications*, *15*(4).
- Alqaryouti, O., Siyam, N., Abdel Monem, A., & Shaalan, K. (2020, Jul). Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics*, *20*(1/2), 142–161. Retrieved from <https://www.emerald.com/insight/content/doi/10.1016/j.aci.2019.11.003/full/pdf?title=aspect-based-sentiment-analysis-using-smart-government-review-data> doi: <https://doi.org/10.1016/j.aci.2019.11.003>
- Azhar, A. N., & Khodra, M. L. (2020). Fine-tuning pretrained multilingual bert model for indonesian aspect-based sentiment analysis. In *2020 7th international conference on advance informatics: Concepts, theory and applications (icaicta)* (pp. 1–6).

- Bischi, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... others (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2), e1484.
- Boquiren, A., Garcia, R., Hungria, C., & de Goma, J. (2022). Tagalog sentiment analysis using deep learning approach with backward slang inclusion. In *Proceedings of the international conference on industrial engineering and operations management nsukka*.
- Chen, Y., Yuan, J., You, Q., & Luo, J. (2018). Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In *Proceedings of the 26th acm international conference on multimedia* (pp. 117–125).
- Co, E., Noblefranca, J., & Gan, J. M. (2023, August). *Affect regression of tweets from the philippines using emoji-based labels*. doi: 10.13140/RG.2.2.23937.84325
- Co, S., Custodio, N., Dela Cruz, A., & Sanchez, M. (2022). *Sentiment classification of tweets from the philippines with emoji-based data annotation* (Unpublished Undergraduate Thesis). De La Salle University Manila, 2401 Taft Avenue, Malate 0922 Manila, Philippines.
- Contreras, J. O., Ballera, M. A., Lagman, A. C., & Raviz, J. G. (2018). Lexicon-based sentiment analysis with pattern matching application using regular expression in automata. In *Proceedings of the 6th international conference on information technology: Iot and smart city* (pp. 31–36).
- Cosme, C. J., & De Leon, M. M. (2024). Sentiment analysis of code-switched filipino-english product and service reviews using transformers-based large language models. In A. Iglesias, J. Shin, B. Patel, & A. Joshi (Eds.), *Proceedings of world conference on information systems for business management* (pp. 123–135). Singapore: Springer Nature Singapore.
- Cruz, J. C. B., & Cheng, C. (2022, June). Improving large-scale language models and resources for Filipino. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 6548–6555). Marseille, France: European Language Resources Association. Retrieved from <https://aclanthology.org/2022.lrec-1.703>
- Cruz, J. C. B., & Cheng, C. K. (2020). Establishing baselines for text classification in low-resource languages. *ArXiv, abs/2005.02068*. Retrieved from <https://api.semanticscholar.org/CorpusID:218502456>
- Cureg, M. Q., De La Cruz, J. A. D., Solomon, J. C. A., Saharkhiz, A. T., Balan, A. K. D., & Samonte, M. J. C. (2019). Sentiment analysis on tweets with

- punctuations, emoticons, and negations. In *Proceedings of the 2nd international conference on information science and systems* (pp. 266–270).
- Delizo, J. P. D., Abisado, M. B., & De Los Trinos, M. I. P. (2020). Philippine twitter sentiments during covid-19 pandemic using multinomial naïve-bayes. *International Journal*, 9(1.3).
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/N19-1423> doi: 10.18653/v1/N19-1423
- Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer New York. Retrieved from <https://books.google.com.ph/books?id=k2zklGOBRDwC>
- Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161, 707–714.
- Frias, R. R., Medina, R. P., & Sison, A. S. (2023). Attention-based bilateral lstm-cnn for the sentiment analysis of code-mixed filipino-english social media texts. *2023 International Conference on Digital Applications, Transformation & Economy (ICDATE)*, 1-5. doi: 10.1109/ICDATE58146.2023.10248926
- Ghafoor, A., Imran, A. S., Daudpota, S. M., Kastrati, Z., Batra, R., Wani, M. A., et al. (2021). The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9, 124478–124490.
- 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.
- Gondaliya, C., Patel, A., & Shah, T. (2021, Jan). Sentiment analysis and prediction of indian stock market amid covid-19 pandemic. *IOP Conference Series: Materials Science and Engineering*, 1020(1), 012023. Retrieved from <https://iopscience.iop.org/article/10.1088/1757-899X/1020/1/012023/pdf> doi: <https://doi.org/10.1088/1757-899x/1020/1/012023>
- Google. (2018). *The wordpiece algorithm in open source bert*. Retrieved from <https://github.com/google-research/bert/blob/master/>

tokenization.py#L300-L359

- Gupta, V., Jain, N., Shubham, S., Madan, A., Chaudhary, A., & Xin, Q. (2021). Toward integrated cnn-based sentiment analysis of tweets for scarce-resource language—hindi. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5), 1–23.
- Hakami, S. A. A., Hendley, R. J., & Smith, P. (2022). Emoji sentiment roles for sentiment analysis: A case study in arabic texts. In *Proceedings of the seventh arabic natural language processing workshop (wanlp)* (pp. 346–355).
- Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1), 75-87. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0167811622000477> doi: <https://doi.org/10.1016/j.ijresmar.2022.05.005>
- Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international aaai conference on web and social media* (Vol. 8, pp. 216–225).
- Imperial, J. M., Orosco, J., Mazo, S. M., & Maceda, L. (2019). *Sentiment analysis of typhoon related tweets using standard and bidirectional recurrent neural networks*. Retrieved from <https://arxiv.org/abs/1908.01765>
- Jaiswal, S. (2024). *Multilayer perceptrons in machine learning: A comprehensive guide*. Retrieved from <https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning>
- Jurafsky, D., & Martin, J. (2024). *Speech and language processing*. Retrieved from https://web.stanford.edu/~jurafsky/slp3/ed3bookaug20_2024.pdf
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020, December). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 757–770). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from <https://aclanthology.org/2020.coling-main.66> doi: 10.18653/v1/2020.coling-main.66
- Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015, 12). Sentiment of emojis. *PLOS ONE*, 10(12), 1-22. Retrieved from <https://doi.org/10.1371/journal.pone.0144296> doi: 10.1371/journal.pone.0144296
- Kudo, T., & Richardson, J. (2018, November). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In E. Blanco & W. Lu (Eds.), *Proceedings of the 2018 conference*

- on empirical methods in natural language processing: *System demonstrations* (pp. 66–71). Brussels, Belgium: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/D18-2012> doi: 10.18653/v1/D18-2012
- Li, M., Ch’ng, E., Chong, A. Y. L., & See, S. (2018). Multi-class twitter sentiment classification with emojis. *Industrial Management & Data Systems*, 118(9), 1804–1820.
- Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). Wangchanberta: Pretraining transformer-based thai language models. *arXiv preprint arXiv:2101.09635*.
- 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).
- Mateo, J. (2024). *Pinoys still top consumers of online video content*. <https://www.philstar.com/headlines/2024/02/07/2331549/pinoys-still-top-consumers-online-video-content>.
- Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE access*, 8, 131662–131682.
- Mohammad, S., & Turney, P. (2013). Crowdsourcing a word-emotion association lexicon. *National Research Council Canada*.
- Muhammad, S., Abdulmumin, I., Ayele, A., Ousidhoum, N., Adelani, D., Yimam, S., ... Arthur, S. (2023, December). AfriSenti: A Twitter sentiment analysis benchmark for African languages. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 13968–13981). Singapore: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2023.emnlp-main.862> doi: 10.18653/v1/2023.emnlp-main.862
- Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.
- Ogueji, K. (2022). *Afriberta: Towards viable multilingual language models for low-resource languages* (Unpublished master’s thesis). University of Waterloo.
- Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). *Sentiment analysis in social networks*. Elsevier.
- Raychawdhary, N., Hughes, N., Bhattacharya, S., Dozier, G. V., & Seals, C. D.

- (2023). A transformer-based language model for sentiment classification and cross-linguistic generalization: Empowering low-resource african languages. *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, 1-5. Retrieved from <https://api.semanticscholar.org/CorpusID:264816018>
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. Retrieved from <https://aclanthology.org/2020.tacl-1.54> doi: 10.1162/tacl_a_00349
- Sabri, N., Edalat, A., & Bahrak, B. (2021). Sentiment analysis of persian-english code-mixed texts. In *2021 26th international computer conference, computer society of iran (csicc)* (pp. 1–4).
- Sennrich, R., Haddow, B., & Birch, A. (2016, August). Neural machine translation of rare words with subword units. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1715–1725). Berlin, Germany: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/P16-1162> doi: 10.18653/v1/P16-1162
- Shehu, H. A., Majikumna, K. U., Suleiman, A. B., Luka, S., Sharif, M. H., Ramadan, R. A., & Kusetogullari, H. (2024). Unveiling sentiments: A deep dive into sentiment analysis for low-resource languages—a case study on hausa texts. *IEEE Access*.
- Shiha, M., & Ayvaz, S. (2017). The effects of emoji in sentiment analysis. *Int. J. Comput. Electr. Eng.(IJCEE.)*, 9(1), 360–369.
- Stets, J. E. (2003). Emotions and sentiments. In *Handbook of social psychology* (pp. 309–335). Springer.
- Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th china national conference, ccl 2019, kunming, china, october 18–20, 2019, proceedings 18* (pp. 194–206).
- Taboy, C. (2023, Feb). A sentiment analysis of “filipinx” on twitter using a multinomial naïve bayes classification model. *Dissertations, Theses, and Capstone Projects*. Retrieved from https://academicworks.cuny.edu/gc_etds/5234/
- Tatachar, A. V. (2021). Comparative assessment of regression models based on model evaluation metrics. *International Journal of Innovative Technology and Exploring Engineering*, 8(9), 853–860.

- Ullah, F., Chen, X., Shah, S. B. H., Mahfoudh, S., Hassan, M. A., & Saeed, N. (2022). A novel approach for emotion detection and sentiment analysis for low resource urdu language based on cnn-lstm. *Electronics*, 11(24), 4096.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., & El-Kishky, A. (2023). Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th acm sigkdd conference on knowledge discovery and data mining* (pp. 5597–5607).
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43–76.