big data and cognitive computing

MDPI

Article

A Mixed Malay–English Language COVID-19 Twitter Dataset: A Sentiment Analysis

Jeffery T. H. Kong [1], Filbert H. Juwono [2], Ik Ying Ngu [3], I. Gde Dharma Nugraha [4,*], Yan Maraden [4] and W. K. Wong [2]

[1] Department of Electrical and Computer Engineering, Curtin University Malaysia, Miri 98009, Malaysia
[2] Computer Science Program, University of Southampton Malaysia, Iskandar Puteri 79100, Malaysia
[3] Department of Media and Communication, Curtin University Malaysia, Miri 98009, Malaysia
[4] Department of Electrical Engineering, Universitas Indonesia, Depok 16424, Indonesia
* Correspondence: i.gde@ui.ac.id

Abstract: Social media has evolved into a platform for the dissemination of information, including fake news. There is a lot of false information about the current situation of the Coronavirus Disease 2019 (COVID-19) pandemic, such as false information regarding vaccination. In this paper, we focus on sentiment analysis for Malaysian COVID-19-related news on social media such as Twitter. Tweets in Malaysia are often a combination of Malay, English, and Chinese with plenty of short forms, symbols, emojis, and emoticons within the maximum length of a tweet. The contributions of this paper are twofold. Firstly, we built a multilingual COVID-19 Twitter dataset, comprising tweets written from 1 September 2021 to 12 December 2021. In particular, we collected 108,246 tweets, with over 67% in Malay language, 27% in English, 2% in Chinese, and 4% in other languages. We then manually annotated and assigned the sentiment of 11,568 tweets into three-class sentiments (positive, negative, and neutral) to develop a Malay-language sentiment analysis tool. For this purpose, we applied a data compression method using Byte-Pair Encoding (BPE) on the texts and used two deep learning approaches, i.e., the Multilingual Bidirectional Encoder Representation for Transformer (M-BERT) and convolutional neural network (CNN). BPE tokenization is used to encode rare and unknown words into smaller meaningful subwords. With the CNN, we converted the labeled tweets into image files. Our experiments explored different BPE vocabulary sizes with our BPE-Text-to-Image-CNN and BPE-M-BERT models. The results show that the optimal vocabulary size for BPE is 12,000; any values beyond that would not contribute much to the F1-score. Overall, our results show that BPE-M-BERT slightly outperforms the CNN model, thereby showing that the pre-trained M-BERT network has the advantage for our multilingual dataset.

Keywords: BPE; CNN; COVID-19; fake news; M-BERT; Malaysia; sentiment analysis

check for updates

Citation: Kong, J.T.H.; Juwono, F.H.; Ngu, I.Y.; Nugraha, I.G.D.; Maraden, Y.; Wong, W.K. A Mixed Malay–English Language COVID-19 Twitter Dataset: A Sentiment Analysis. *Big Data Cogn. Comput.* **2023**, *7*, 61. https://doi.org/10.3390/bdcc7020061

Academic Editor: Carson K. Leung, Giancarlo Fortino, Fei Hao and Xiaokang Zhou

Received: 16 January 2023
Revised: 7 March 2023
Accepted: 10 March 2023
Published: 27 March 2023

1. Introduction

The insurgence of the Coronavirus disease 2019 (COVID-19) pandemic has had a heavy impact on the global information environment by changing people's everyday habits and the way they interact with one another. When governments implemented lockdowns, people were forced to stay at home and relied on online social networks to connect with others, share information, and express their emotions and frustrations online [1]. In order for governments to successfully implement effective health policies and movement control order decisions, public opinions posted on social media need to be analyzed [2]. Sentiment analysis is an essential approach, usually used for gauging public sentiment on a certain subject, particularly on highly polarizing topics of discussion. This study uses sentiment analysis to gauge public sentiments on highly polarizing COVID-19-related topics on social media, as negative interactions could lead to misinterpretations of the pandemic [3]. The sentiment analysis is used because it has gained popularity in developing algorithms for knowledge discovery from opinionated content online [4]. This method involves advanced statistical and machine learning approaches [5].

Most of the research on sentiment analysis focuses on single-language expressions, English in particular, and less on multilingual expressions that mix more than two languages or dialects. Therefore, this study would like to fill this gap by developing a sentiment analysis method that could cater to the multilingual or mixed-language context in a multi-ethnic country such as Malaysia. In particular, people in Malaysia use the Malay language as the primary language, followed by English and Chinese. Often, Malaysians mix Malay and English, creating a ubiquitous Creole-like language in expressing their opinions on social media platforms. As such, we will further expand our understanding of people's interactions on social media in framing their perceptions or misperceptions about COVID-19 in multiple languages and dialects.

The mixing of languages is not uncommon, and sometimes, online users tend to mix English with Chinese on Twitter [6]. The multilingual sentiment analysis in Malaysia is more complex with the existence of Malay dialects in each state. To give a few examples, Sarawak state has 'Bahasa Sarawak' [7], Sabah state has 'Bahasa Sabah' [8], and Kelantan state has 'Bahasa Kelantan' [9], where 'Bahasa' is the word for 'language' in English. The written texts by the locals of different states are often direct translations from the pronunciations into Latin characters. The frequent use of short forms of these dialect words complicates the sentiment analysis even further. To understand the sentiment of the mixed language or multilingual form, one would need to develop a sentiment analysis method catering to the multilingual setting in Malaysia.

A well-known deep learning framework for natural language processing (NLP) is the Multilingual Bidirectional Encoder Representations from Transformers (M-BERT), a variation of BERT [10] that is pre-trained on 104 languages using a Wikipedia corpus with shared vocabulary across many languages that include Malay and English. One way to deal with low-resource languages is to translate the text into English and perform sentiment using a readily available English sentiment analysis model. The authors of [11] proposed word-to-word translation of the lack-of-resource language to a language that has a sentiment dictionary using machine translation with Google Translate. In [12], a neural machine translation (NMT) system was used to translate multilingual text into English. However, the machine translation, as noted by the authors, has difficulties in translating rarely occurring words. In the Malaysian context, automatic machine translation from Malaysian dialects to English is currently not available. The frequent use of short forms in tweets further complicates the translation accuracy. Therefore, we applied a popular data compression method used by the state-of-the-art NLP models, i.e., Byte-Pair Encoding (BPE), to perform subword tokenization on the texts. The experiments were carried out with different vocabulary sizes to find the optimal vocabulary size for the BPE.

Due to the unique multilingual setting in Malaysia, in this paper, we present a multilingual Twitter dataset obtained in Malaysia from 1 September 2021 to 12 December 2021. The dataset will be used for sentiment analysis using the M-BERT model. We further use a text-to-image method with deep learning approach to perform sentiment analysis on the images of the BPE tokens. In particular, this paper aims to cover the sentiment analysis during the reopening of the economy, administration of third doses to health front liners and elderly, vaccination for adolescents aged between 12 and 17 years old, and the emergence of the new variant (Omicron) discovery in Malaysia.

1.1. Existing Work

A few datasets for sentiment analysis regarding the COVID-19 pandemic have been published in the literature. The authors in both [13,14] collected tweets in the early phase of the pandemic, between March 2020 and June 2020, to study sentiment analysis during the lockdown in India. During a similar timeframe, the authors of [15] collected data on social media platforms to track emotional expressions during COVID-19 in Austria. Meanwhile, the authors of [16] combined three COVID-19 datasets from Kaggle.com with a total of 412,721 tweets collected from February 2020 to April 2021 to perform sentiment analysis on COVID-19 vaccination. Similarly, Ref. [17] conducted sentiment analysis towards COVID-19 vaccination and vaccine types from Twitter users in the USA, the UK, Canada, Turkey, France, Germany, Spain, and Italy. They managed to collect a total of 928,402 vaccination-related tweets in English and Turkish from November 2020 to March 2021. In addition, the authors of [18] collected tweets in the early stage of vaccinations in Japan, the US, and the UK from December 2020 to June 2021.

In general, there are four main types of approaches to sentiment analysis, a lexicon-based approach, a machine learning approach, a deep learning approach, and a hybrid approach [19]. Sentiment analysis using a lexicon-based approach builds a dictionary of words that are labeled with either positive or negative sentiment to determine the text sentiment polarity. A lexicon-based approach was performed in [13], which used TextBlob, a library of python that carries a pre-defined dictionary of positive and negative words to predict the tweets' sentiments. On the other hand, a machine learning approach has been explored, for example, [16], in which the authors performed 14 machine learning algorithms on a labeled dataset and found that logistic regression achieved the highest accuracy using count vectorizer and TF-IDF as feature extraction techniques. In the field of deep learning approaches, Ref. [20] demonstrated how CNNs process text and found that the filters in each layer may use several activation patterns to capture various different semantic classes of ngrams. Contrary to the traditional CNN approach, the authors of [21] proposed input text as images and applied a 2D CNN to extract the local and global semantics of the sentences from different word visual patterns on Chinese text classification. In more recent studies, a hybrid approach, a combination of two or more approaches to sentiment analysis, is introduced. In this study [22], the authors combined Naïve Bayes and Random Forest and achieved better accuracy. In agreement with the previous study [23], the authors concurred that the hybrid approach can improve performance accuracy in sentiment analysis, while a lexical approach has more consistency in the performance. The majority of these four approaches for sentiment analysis, however, used English datasets in their studies. In the most recent systematic review work published by [24] in early 2020, low-resource languages such as Malay are falling behind, with only one sentiment analysis study [25] performed in Malay language in 2018. However, the authors did not publish their Malay sentiment dataset. Furthermore, most sentiment analysis studies in the Malay language evolved around lexicon-based approaches [26–29]. Since 2019, sentiment analysis research completed in the Malay language using machine learning and deep learning approaches has been very much lacking, with only one study [30] comparing the results of the Decision Tree, Support Vector Machine (SVM), and Naïve Bayes (NB) methods.

In solving multilingual problems using a deep learning approach, Ref. [31] developed a deep learning approach using XLM-RoBERTa, Bidirectional Recurrent Neural Networks (Bi-RNNs), and Bidirectional Long Short-Term Memory (LSTM) to perform multi-label emotion classification on 100 types of languages without detecting its language. The authors of [12] used Global Vectors (GloVe) as word embeddings and input them into RNN-LSTM to create a basic model on an English dataset and to convert other languages into English using Neural Machine Translation (NMT) before the model could perform sentiment analysis on multilingual texts. The accuracy of the translation greatly depends on NMTl and with noises such as short forms, and misspelled words in tweets [24], highly accurate translation to English will be difficult. Using a similar strategy to translate another language corpus into English, the authors of [32] translated a sentiment-labeled dataset in the Bengali language into English language using Google Translate and used LSTM to predict sentiment analysis. However, the authors only achieved 59% accuracy for the Bengali language and 72.2% accuracy for English using a deep learning approach.

In 2019, BERT was introduced by [9] as a new language representation model that achieves state-of-the-art results in eleven NLP tasks including sentiment analysis. BERT has two main components in its implementation, which are pre-training and fine-tuning. To train BERT in an unsupervised way, two unique training approaches, the Masked Language

Model (MLM) and Next Sentence Prediction (NSP), are used on BooksCorpus and English Wikipedia. In fine-tuning, BERT uses the self-attention mechanism in the transformer, and the process only requires a simple classification layer added to the pre-trained model. Shortly after the introduction of BERT, the same authors, Jacob Devlin and his colleagues from Google, released the multilingual version (M-BERT). The M-BERT model is a single-language model pre-trained on a Wikipedia corpus of 104 languages [33]. The model also supports non-Latin languages, such as Chinese, Tamil, and Hindi, and aims to tackle low-resource languages such as Malay.

We noted that a few researchers have utilized M-BERT to predict sentiment analysis on low-resource languages, such as [34], which achieved an accuracy of 60% on a three-class manually tagged dataset for sentiment analysis in Bengali language using M-BERT. For sentiment analysis on India-specific English tweets, the authors of [14] used the BERT model and A Light BERT (ALBERT) model to achieve 65% and 61% accuracy, respectively, with higher accuracy on the positive class as compared to the neutral class. The authors in [35] used the BERT model to perform sentiment analysis on the Persian and English datasets and achieved 66.17% accuracy. Sentiment analysis accuracy with BERT was 66.7% in the Indonesian language dataset, with higher accuracy on the positive class and lower accuracy on the negative class [36]. Similarly, the authors of [37] applied the M-BERT cased model to perform sentiment analysis on the Vietnamese dataset of three classes and achieved 65% accuracy.

1.2. Motivation and Contributions

The absence of advancements in Malay sentiment analysis, coupled with the intricacy of multilingual sentiment analysis, motivated us to explore the feasibility of utilizing a deep learning approach for multilingual sentiment analysis in Malaysia. As outlined in the section on existing work, it has been observed that the average accuracy of M-BERT model on low-resource languages (such as Bengali, Persian, Indonesian, and Vietnamese) using a three-class sentiment dataset ranged between 60% and 70%.

Our research presents a novel approach to sentiment analysis for the Malay language, a low-resource language with limited annotated datasets. We address the challenge of representing the many emojis, rare words, and unknown words in Malay tweets by incorporating Byte Pair Encoding (BPE) tokens and converting them into subword units. In addition, we introduce a new modality for multilingual sentiment analysis by converting tweet texts into fixed-size images and using a convolutional neural network (CNN) to classify sentiment. To our knowledge, this is the first application of image-based sentiment analysis to Malay tweets using BPE tokens. We compare our image-based approach (BPE-Text-to-Image-CNN) with a BPE token-based model using the multilingual BERT (MBERT) architecture, providing insight into the relative performance of these two approaches. In summary, our research makes the following contributions:

- We collected 108,246 tweets to provide a multilingual dataset on COVID-19-related tweets posted in Malaysia. The dataset has been published on Github [38] and made publicly available for further research work.
- We manually annotated sentiments on 11,568 tweets in terms of three classes of sentiments (positive, negative, and neutral) for two different languages: Malay and English.
- This study contributes to the field of sentiment analysis by demonstrating the effectiveness of incorporating BPE tokens into MBERT and text-to-image CNN models for sentiment analysis in low-resource languages such as Malay.

The rest of the paper is structured as follows. In Section 2, we describe the methodology of our dataset collection and the propose a new sentiment analysis method in a multilingual setting. In Section 3, we outline the experiment settings, present the results, and analyze a comparison with a state-of-the-art method. Finally, we draw a conclusion and consider possible future work in Section 4.

2. Methodology

We introduce our dataset collection method, manual annotation method, and BPE-Text-to-Image-CNN methods for the purpose of the sentiment analysis task in this section.

2.1. Dataset Collection Method

Twitter offers a free standard product track to researchers and students to access their platform. We used Tweepy, an open-source Python library, to connect to Twitter's Application Programming Interface (API) for data collection. We scheduled the data collection daily at 12 noon from 1 September 2021 to 12 December 2021. The timeframe under consideration in this study encompasses several key events in Malaysia, including the reopening of the economy, the administration of third vaccine doses to healthcare workers and the elderly, the vaccination of adolescents aged 12–17 years old, and the discovery of the new Omicron variant.

We defined the search location and search keywords in the search query to limit the search within Malaysia and retrieved only COVID-19-related tweets. As the three main races in Malaysia are Malays, Chinese, and Indians, we translated the search keywords from English into Malay, Chinese, and Tamil using Google Neural Machine Translation (GNMT). Table 1 shows the search keywords used in the data collection. Some keywords were added at a later date. For example, the keyword "Omicron", a new variant of concern, was officially named Omicron on 26 November 2021 by the World Health Organization (WHO). The tracing date was recorded for each keyword to track its starting date used in the search.

Table 1. The list of search keywords used to collect the tweets.

English	Malay	Chinese	Tamil	Tracing Date
vaccination	vaksinasi	接种	தடுப்பூசி	2021-09-01
vaccine	vaksin	疫苗	தடுப்பூசி	2021-09-01
delta variant	varian delta	增量变体	டெல்டா மாறுபாடு	2021-09-01
booster	penggalak	助推器	பூஸ்டர்	2021-09-01
covid	covid	新冠	கோவிட்	2021-09-01
mask	topeng	口罩	மாஸ்க்	2021-09-01
quarantine	kuarantin	隔离	தனிமைப்படுத்துதல்	2021-09-01
Movement control order	Perintah kawalan pergerakan	行动管制令	இயக்க கட்டுப்பாட்டு ஒழுங்கு	2021-09-01
mkn, jkjav, kitajagakita, mysejahtera, icu, pcr, mco, pkp, az				2021-09-01
endemic	endemik	地方病	உள்நாட்டு	2021-09-06
oximeter	oksimeter	血氧计	ஆக்சிமீட்டர்	2021-09-08
hospital	hospital	医院	மருத்துவமனை	2021-09-08
pandemic	pandemik	大流行	சர்வதேச பரவல்	2021-09-10
Astrazeneca	Astrazeneca	阿斯利康	அஸ்ட்ராசெனெகா	2021-09-10
Pfizer	Pfizer	辉瑞	பைசர்	2021-09-10
Sinovac	Sinovac	华兴	சினோவாக	2021-09-10
test kit	kit ujian	测试套件	சோதனை கிட்	2021-09-15
pneumonia	pneumonia	肺炎	நிமோனியா	2021-10-06
ivermectin	ivermektin	伊维菌素	ஐவர்மெக்டின்	2021-10-22
wuhan	wuhan	武汉	வுஹான்	2021-10-22
comorbidity	komorbiditi	合并症	கொமொர்பிடிட்டி	2021-10-22
comirnaty	comirnaty	共同体	கொமர்னாடி	2021-10-22
panadol	panadol	帕纳多	பனடோல்	2021-10-31
PICK (Program Imunisasi COVID-19 Kebangsaan), CITF (COVID-19 Immunisation Special Task Force)				2021-11-04
TRIIS (Test, Report, Isolate, Inform, Seek)				2021-11-12
Omicron	Omicron	奥米克戎	ஓமிக்ரான்	2021-11-29

In order to ensure the collected tweets were the ones posted within Malaysia, we added geographic coordinates (geocode) into the search parameters. The latitude and longitude of each state's capital city and other cities with large populations were plotted on a map of Malaysia with a radius range of 3 km to 70 km, as shown in Table 2. Variations of the radius circles method were inspired by the theory of circles used to collect the tweets in Great Britain [39]. As discussed in the paper, the method using highly overlapping circles and larger circles would cause a lot of duplicate tweets, which is known as the 'circle coverage problem'. As we aimed to reduce both the overlaps between circles and their numbers as much as possible and to solve the problem of retrieving duplicate tweets, the following control measure was applied before saving the tweets. The data collection program would refer to a master file containing all the tweet IDs that had been retrieved. If the new tweet ID did not appear inside the master file list, then the program would proceed to save the tweet content. In the case that the retrieved tweet ID was found inside the master file list, the program would stop processing the particular tweet ID and proceed to the next tweet ID.

Table 2. Geocodes used in data collection.

Geocodes (Latitude,Longitude)	Radius	City	State
1.8548,102.9325	30 km	Batu Pahat	Johor
1.4655,103.7578	3 km	Johor Bahru	Johor
1.6006,103.6419	30 km	Senai	Johor
6.12104,100.36014	50 km	Alor Setar	Kedah
6.13328,102.2386	50 km	Kota Bahru	Kelantan
3.1412,101.68653	70 km	Kuala Lumpur	Federal Territories
2.196,102.2405	50 km	Malacca	Malacca
3.8077,103.326	50 km	Kuantan	Pahang
5.41123,100.33543	30 km	George Town	Penang
4.5841,101.0829	50 km	Ipoh	Perak
5.9749,116.0724	50 km	Kota Kinabalu	Sabah
5.8402,118.1179	50 km	Sandakan	Sabah
4.24482,117.89115	50 km	Tawau	Sabah
3.16667,113.03333	50 km	Bintulu	Sarawak
1.55,110.33333	50 km	Kuching	Sarawak
4.4148,114.0089	20 km	Miri	Sarawak
2.3,111.81667	50 km	Sibu	Sarawak
5.3302,103.1408	50 km	Kuala Terengganu	Terengganu

2.2. Dataset Description (MyCovid-Senti)

We collected a total of 108,246 tweets over the period of 103 days. The average number of daily tweets on COVID-19 in Malaysia during this period was 1050 tweets. The bar chart in Figure 1 clearly illustrates that tweets about COVID-19 concerns did not slow down over that time period. Fewer counts were seen in the first two weeks due to a lower number of keywords recorded.

Overall, Twitter detected 40 distinct languages in the collected data. For the language visualization, all languages other than English, Malay, Chinese, and Tamil were placed under the 'other' category. The percentage distribution of the languages was recorded as follows: 67% of the tweets gathered were in Malay, 27% in English, 2% in Chinese, less than 1% in Tamil, and roughly 4% in other languages. Tweets were seen to be a combination of Malay and English, as well as some Chinese and Tamil.

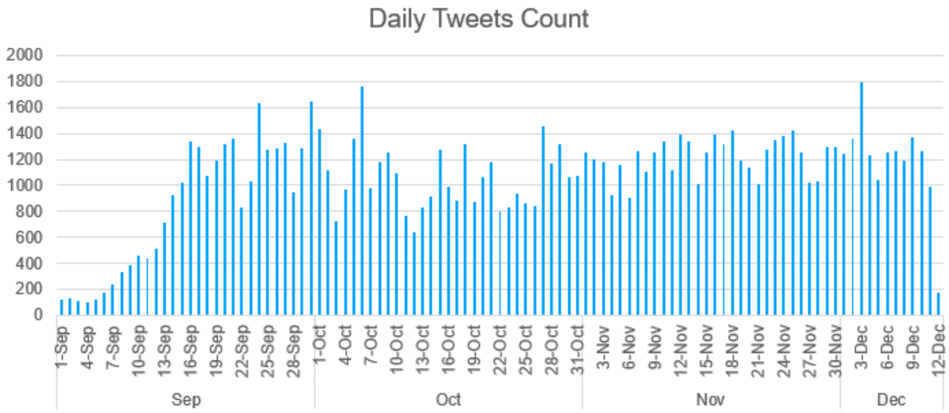


Figure 1. Timescale of Tweets acquired on COVID-19.

Of the 108,246 tweets collected, we self-annotated 11,568 tweets randomly in accordance with the guidelines outlined in the methodology and named as *MyCovid-Senti* dataset. We labeled 5655 tweets as negative, 2728 as neutral, and 3185 as positive. Table 3 shows the total samples for the three-class sentiment analysis. We set up another experiment for a two-class sentiment analysis (positive and negative). The two-class sentiment would give us a balanced dataset, whereby we changed a neutral label into a positive label to combine neutral and positive classes together. In order to have an overview of MyCovid-Senti dataset, we display the most used words in the dataset in a word cloud image, shown in Figure 2. Our dataset is made available on the GitHub page.

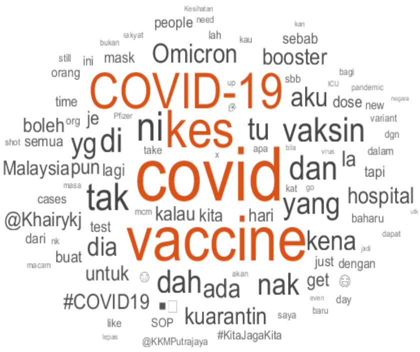


Figure 2. Word cloud image of the MyCovid dataset.

Table 3. MyCovid-Senti Classes.

MyCovid-Senti	3-Class	2-Class
Negative	5655	5655
Neutral	2728	-
Positive	3185	5913
Total	11,568	11,568

2.3. Manual Sentiment Annotation Method

In the manual annotation task, we appointed three independent annotators to label the tweets into three classes of sentiments: positive, negative, and neutral. Manual annotation is a seemingly easy task, but in reality, tagging text with sentiment labels is highly subjective and influenced by personal beliefs. For example, the annotator being a pro-vaccine or anti-vaccine activist would very much influence their sentiment label selection judgment of a vaccine-related post. Thus, a clear and straightforward guideline is needed for the long hours and mental focus required for the task. Our approach to manual sentiment annotation took the intuition from the proposed semantic role-based sentiment questionnaire by [40], that is, to determine the speakers' emotional state and identify the Primary Target of Opinion (PTO) using four questions. There are times when the speaker's emotional state is absent in the text. As such, we can look at the PTO's emotional state. A PTO is an entity that might be a person, object, organization, group of people, or other similar entity. For instance, the PTO is the person or group being targeted if the text condemns the behavior or opinions of that person (or group of people). Another example of a PTO is 'those who do not believe in evolution' if the text makes fun of those who reject evolution. Otherwise, the the PTO is 'evolution' if the text criticises or challenges evolution. The semantic-role-based sentiment questions are as follows:

(Q1) What best describes the speaker's emotional state? (The following emotional states are used in the following questions as well).

(a) *positive state*: there is an explicit or implicit clue in the text suggesting that the speaker is in a positive state, i.e., happy, excited, task completion, festive greetings, hope for better, advise, recovering, taken positive actions (e.g., booster shots done), good intention, and make plans, etc.

(b) *negative state*: there is an explicit or implicit clue in the text suggesting that the speaker is in a negative state, i.e., sad, angry, disappointed, demanding, questioning, doubt, worry, forcing, ill intention, impatience, etc.

(c) *neutral state*: there is no explicit or implicit indicator of the speaker's emotional state, i.e., news that purely reports about daily statistics on COVID-19 cases, notices of meeting/webinars date time, describing guidelines, and information.

(Q2) When speaker's emotional state is absence, identify the Primary Target of Opinion (PTO) attitude, it can be towards a person, group, object, events, or actions. If there are more than one opinions, select the stronger sentiment of opinions.

(Q3) If the entire text is a quote from another person (the original author) and the speaker's attitude is not clear, then select the original author as the speaker.

(Q4) What best describes how the majority of individuals feel or public opinion about the PTO?

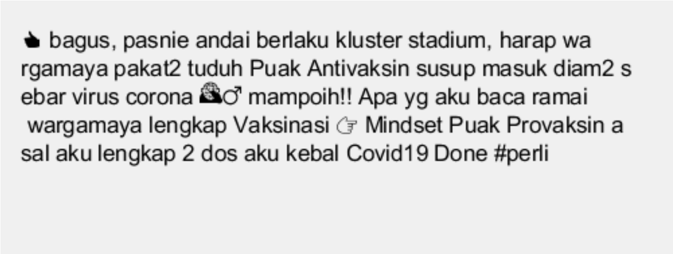
Note that Q4 is needed when none of the first three questions is able to categorize sentiment of the tweet. The last question considers the public sentiments or majority sentiments towards the PTO or the PTO's actions.

2.4. BPE-Text-to-Image-CNN Method

To fit the 280-character limitation imposed by the Twitter platform, the language used in the tweet can often be informal and can contain short forms, emojis, emoticons, symbols, and misspelled words. In a recent study [21], the authors demonstrated that their 2D CNN method was able to extract semantically significant features from images containing text without the need for sequential processing pipelines and optical character recognition. Therefore, a streamlined technique of converting text into images is reasonably used in order to collect all of this information and see how the deep learning network attempts to capture the features surrounding the text and learn. In four tasks, we demonstrate the detailed implementation of our BPE-Text-to-Image-CNN method incorporating Byte-Pair Encoding (BPE) tokens. We ran the experiments 10 times and obtain the mean F1-scores for each three-class and two-class sentiment classification setup on the BPE-Text-to-Image-CNN method.

1. **Task 1: Text Pre-processing.** We applied case-folding to lowercase words and the removal of stop words, white spaces, @mentions, and URLs from the tweets. The list of stop words was obtained in English from scikit-learn, a python library. The list had 317 English words that were converted into Malay words to build Malay stop words for the removal. We removed the duplicates as well as the retweets with the same wordings. However, we kept emojis and emoticons, as the model might learn from these features.

2. **Task 2: BPE Tokenization.** We created another set of text using the tokens generated by BPE. BPE is a data-compression method that selects the most occurring pair of characters and replaced them with a character that does not exist within the data [41]. BPE tokenization was chosen, as the algorithm deals with the unknown word problem, which is very common with the usage of short forms on text postings on online social media. BPE can also reduce or increase the dataset's vocabulary size by changing the value of the maximum vocabulary size during the BPE tokenization process. In a very recent study [4], the authors found that in a 10k opinion dataset, when the vocabulary size increased beyond 8000, the accuracy score dropped. Another study [42] concurred that the best performance was achieved with small (30K) to medium (1.3M) data sizes, at an 8000 vocabulary size. Our dataset has around 11K tweets and is a comparable dataset size to the previous research findings in terms of optimal vocabulary size. The total count of unique tokens of the MyCovid-Senti dataset was 19,185. In our experiments, we tested a few vocabulary sizes for BPE tokens by setting it to 1000, 2000, 4000, 8000, 12,000, 16,000, the original token size (19,185), and 24,000.
3. **Task 3: Text-to-image Conversion.** With the tweet text limit at a maximum of 280 characters, we reshaped texts from one-row vector into a matrix size of 5 rows × 56 columns. In other words, we arranged the texts on an image with only 56 characters in a row. Then, the next characters were moved to a new line. In the final step of the conversion, we used the print function in Matlab to export the matrix into image form, as shown in Figure 3.



👍 bagus, pasnie andai berlaku kluster stadium, harap wa
rgamaya pakat2 tuduh Puak Antivaksin susup masuk diam2 s
ebar virus corona 🤦‍♂ mampoih!! Apa yg aku baca ramai
 wargamaya lengkap Vaksinasi ☞ Mindset Puak Provaksin a
sal aku lengkap 2 dos aku kebal Covid19 Done #perli

Figure 3. Tweet in image form.

4. **Task 4: CNN.** We fed the images as features input into a deep-learning neural architecture with 32 layers. The images were augmented to reduce their size by half, i.e., from 188 × 500 to 94 × 250. In our image pre-processing phase, we converted the images from a Red, Green, Blue (RGB) components format to grayscale (where each pixel contains only one data point with a value ranging from 0 to 255). According to [43], gray-scaling is performed so that the number of data that can be represented or need to be processed in each pixel is lower in comparison with a colored image (where each pixel contains three data components for the RGB format). Thus, with the reduced data in each pixel, it naturally reduces the processing power and time required. For the CNN experiments, the dataset was split randomly by 80–20%, with 80% for training and 20% for testing. The CNN model used in the experiment contains seven sets of convolutional layers, a batch normalization layer, and a Rectified Linear Unit (ReLU) layer, as shown in Figure 4. In between the seven sets, there were 2-D max pooling layers before the convolution layer and after the ReLU layer to divide the input by half. After learning features in the seven sets of layers, the CNN architecture shifted to classification. A dropout layer with a dropout probability of 0.5 was applied before the fully connected layer that outputs a vector of K dimensions, where K is the number of classes that the network predicts. Finally, a softmax function with the classification layer was used as the final layer. For the training options, we used the Stochastic Gradient Descent with Momentum (SGDM) optimizer and set the initial learning rate to 0.001. Then, we reduced the learning rate by a factor of 0.2 every five epochs. The training was run for a maximum of 15 epochs with a mini-batch of 64 observations at each iteration.

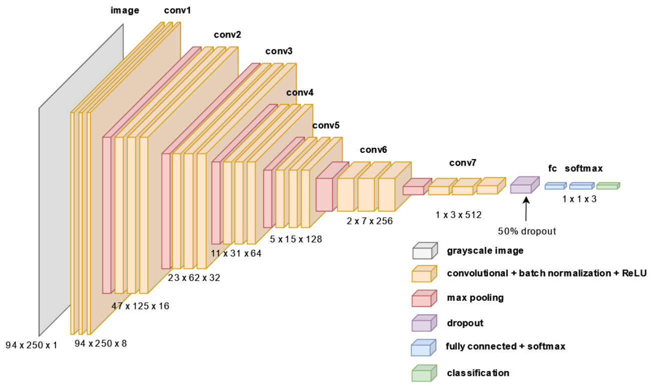


Figure 4. BPE-Text-to-Image-CNN (32-layer) model architecture for sentiment analysis.

2.5. *BPE-M-BERT*

The M-BERT model was released by [10] as a single language model and pre-trained on monolingual Wikipedia in 104 languages [44], which include the Malay language. The model contains 12 layers of transformers, where each layer contains an embedding length of 768 and 12 attention heads per transformer layer, with a total of 110M trainable parameters. In the BPE-M-BERT model, we also tokenized the texts into BPE tokens. The BPE process is similar to that described in the BPE Tokenization step of the BPE-Text-to-Image-CNN method. Then, we utilized the tokenizer provided by the M-BERT model that uses 110k shared WordPiece vocabulary. Besides performing the basic tokenization (lower casing, punctuation splitting, and whitespace tokenization), the tokenization process encodes text into sequences of integers that represent the details of padding, start, separator, and mask tokens. The sequences of BERT model tokens are then converted to an N-by-D array of feature vectors, where N is the number of training observations and D is the dimension of the BERT embedding, which is 768 columns of weights in decimals.

M-BERT is particularly good at zero-shot learning, in which the training is performed in one language to fine-tune the model and then evaluated in another language. According to the research performed by [33], M-BERT's performance improved with a similarity between languages, as the similarity makes it easier for M-BERT to map linguistic structures. In their findings, the authors also noted that M-BERT could generalize well from monolingual inputs to code-switching text. In a recent study [45], the authors witnessed that code-switching was common on social media with the multilingual community, in particular mixing a low-resource language with high-resource language in the same text. We note that in the collected tweets, mixing of English and Malay words within a single text was a common phenomenon. Thus, M-BERT is well suited for this comparison study because Malaysians often code-switch between English and Malay words in their online texts.

In our M-BERT experiments, the dataset was split into 80% train data and 20% test data. We ran the M-BERT experiment 10 times by using the M-BERT model as described in [10] to obtain the mean F1-scores.

3. Results and Analysis

We compared the performance of our proposed BPE-Text-to-Image-CNN method with the latest M-BERT model coded by [46], specifically for a multilingual model that fixes normalization issues in languages with both Latin and non-Latin alphabets, incorporating our BPE tokenization. In the experiments, both models were trained on the MyCovid-Senti dataset that we had annotated.

The performance for multi-class classification is commonly evaluated by using the F1-score, which is also known as F-measure. The F1-score can be defined as the harmonic mean of the precision and recall, with the best value being 1 and the lowest being 0. In particular, we adopted the averaging methods for F1-score calculation, namely F1-micro, F1-macro, and F1-weighted to evaluate the performance of each method.

1. **F1-micro**. Count the total of true positive samples (TP), false negative samples (FN), and false positive samples (FP) to determine the F1-score globally. The expression is given by

$$\text{F1-micro} = \frac{2 \times \text{Pr-micro} \times \text{Re-micro}}{\text{Pr-micro} + \text{Re-micro}}, \tag{1}$$

where

$$\text{Pr-micro} = \frac{\sum_{i=1}^{|\mathcal{C}|} TP_i}{\sum_{i=1}^{|\mathcal{C}|} TP_i + FP_i}, \tag{2}$$

$$\text{Re-micro} = \frac{\sum_{i=1}^{|\mathcal{C}|} TP_i}{\sum_{i=1}^{|\mathcal{C}|} TP_i + FN_i}, \tag{3}$$

and $|\mathcal{C}|$ is the cardinality of the class $\mathcal{C}$.

2. **F1-macro**. Calculate F1-score for each class, and find the mean of all F1-score per class. However, this metric disregards the class imbalance. The expression is given by

$$\text{F1-macro} = \frac{2 \times \text{Pr-macro} \times \text{Re-macro}}{\text{Pr-macro} + \text{Re-macro}}, \tag{4}$$

where

$$\text{Pr-macro} = \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \frac{TP_i}{TP_i + FP_i}, \tag{5}$$

$$\text{Re-macro} = \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} \frac{TP_i}{TP_i + FN_i}. \tag{6}$$

3. **F1-weighted**. Calculate the F1-score for each class, and find the mean of all F1-score per class while considering each class weight. The weight is proportional to the number of samples in each class. This allows 'macro' to account for class imbalance. The expression is given by

$$\text{F1-weighted} = \sum_{i=1}^{|\mathcal{C}|} w_i \times F1_i, \tag{7}$$

where

$$F1_i = \frac{2 \times Pr_i \times Re_i}{Pr_i + Re_i}, \tag{8}$$

$$Pr_i = \frac{TP_i}{TP_i + FP_i}, \tag{9}$$

$$Re_i = \frac{TP_i}{TP_i + FN_i}, \tag{10}$$

and w_i is the weight for class i.

Table 4 presents the training results of two proposed BPE-models, BPE-Text-to-Image-CNN and BPE-M-BERT, on our MyCovid-Senti dataset with three-class and two-class labels (where the neutral label is treated as positive in two-class). The F1-score results for micro, macro, and weighted averages of 10 trials are presented for both label types.

It can be observed that BPE-M-BERT outperforms BPE-Text-to-Image-CNN in all scenarios. Specifically, BPE-M-BERT achieves the best results on the three-class label dataset with a BPE vocabulary size of 12,000, where it achieves F1-micro, F1-macro, and F1-weighted scores of 0.6645, 0.6308, and 0.6517, respectively. For the two-class label dataset, the best results were obtained with the original BPE vocabulary size, where the scores were 0.7170, 0.7165, and 0.7165, respectively.

On the other hand, for the three-class labels, BPE-Text-to-Image-CNN achieved better results with the BPE vocabulary size, ranging from 12,000 to 19,185 for F1-micro, F1-

macro, and F1-weighted scores of 0.5823, 0.5110, and 0.5452, respectively. For the two-class labels, the corresponding scores were 0.6268, 0.6299, and 0.6294, respectively, with the BPE vocabulary size ranging from 12,000 to 24,000.

This section examines the impact of BPE vocabulary size on model performance. Our experiments revealed that lowering the BPE vocabulary size to 8000 and below resulted in decreased F1-score performance. However, increasing the vocabulary size above 8000 using both methods maintained the F1-score within ±1% range. These results contradict the assertion made in [4], which claimed that increasing the vocabulary size above 8000 resulted in reduced accuracy. Notably, reducing the vocabulary size to 1000 led to a significant drop in F1-scores for both methods. In summary, the findings demonstrate that it is possible to maintain high F1-scores while reducing the BPE vocabulary size or the number of unique tokens representing the dataset to 12,000.

Overall, we observed a significant improvement in F1-scores for both BPE-models when performing two-class sentiment classification. Specifically, the BPE-M-BERT model achieved the highest F1-macro of 0.7165, while the BPE-Text-to-Image-CNN model achieved an F1-macro of 0.6299. In contrast, for the three-class sentiment setting, the BPE-M-BERT model achieved the highest F1-macro of 0.6308, whereas the BPE-Text-to-Image-CNN model only scored an F1-macro of 0.5110. We also compared our results with the work by [34] on low-resource Bengali language and found that our model achieved a similar F1-micro of 58% for three-class sentiment and a slightly lower F1-micro of 62% for two-class sentiment. Notably, the highest accuracy for their model CNN-BERT-three-class was 58%, while that of CNN-BERT-two-class was 67%.

Table 4. Training results comparison between BPE-Text-to-Image-CNN and BPE-M-BERT model on three-class and two-class labels.

Methods	BPE Vocabulary Size	Dataset 3-Class(neg., pos., Neutral)			Dataset 2-Class(neg., pos.+Neutral)		
		F1-Micro	F1-Macro	F1-Weighted	F1-Micro	F1-Macro	F1-Weighted
BPE-Text-to -Image-CNN	24,000	0.5752	0.5047	0.5383	**0.6268**	0.6265	0.6264
	Original Text (19,185)	**0.5823**	0.5103	**0.5452**	0.6250	0.6247	0.6247
	16,000	0.5755	0.5053	0.5388	0.6222	0.6218	0.6217
	12,000	0.5771	**0.5110**	0.5437	0.6295	**0.6299**	**0.6294**
	8000	0.5713	0.5065	0.5391	0.6236	0.6233	0.6232
	4000	0.5684	0.4950	0.5298	0.6259	0.6254	0.6251
	2000	0.5664	0.4931	0.5281	0.6078	0.6075	0.6075
	1000	0.5608	0.4831	0.5201	0.6086	0.6082	0.6081
BPE-M-BERT	24,000	0.6595	0.6250	0.6466	0.7104	0.7094	0.7093
	Original Text (19,185)	0.6517	0.6161	0.6391	**0.7170**	**0.7165**	**0.7165**
	16,000	0.6557	0.6177	0.6407	0.7118	0.7108	0.7110
	12,000	**0.6645**	**0.6308**	**0.6517**	0.7053	0.7040	0.7042
	8000	0.6536	0.6163	0.6390	0.6992	0.6988	0.6988
	4000	0.6411	0.6020	0.6255	0.6945	0.6940	0.6938
	2000	0.6300	0.5831	0.6095	0.6825	0.6804	0.6807
	1000	0.6068	0.5629	0.5880	0.6633	0.6627	0.6628

4. Conclusions

In this paper, we have provided a manually annotated COVID-19 sentiment dataset in three-class (positive, negative, or neutral) and two-class (positive and negative) sentiment labels, consisting of mixed Malay and English language. This study has practical implications for researchers and practitioners who work with sentiment analysis in low-resource languages. The study shows that the BPE tokenization method can effectively represent rare and unknown words, as well as emojis, in tweets, which can improve the performance of sentiment analysis models. Additionally, the study suggests that the use of a different modality, such as converting tweet text into a fixed-size image, can also be effective for multilingual sentiment analysis in two-class labels. In particular, we have proposed BPE-M-BERT and BPE-Text-to-Image-CNN methods for sentiment analysis on the low-resource Malay language. In the BPE-Text-to-Image-CNN method, our model captures multiple languages' linguistic features and text styles in BPE tokens, which are then converted into images. We then used a deep learning CNN architecture to analyze the text images for the sentiment classification task. As elaborated in the existing work, there is a lack of studies into the low resource language of Malay since 2018. Thus, a baseline performance of the F1-score for three-class and two-class datasets is established using BPE models on sentiment analysis in the Malay language. In summary, our results indicate that the BPE-M-BERT model is more effective than the BPE-Text-to-Image-CNN model for sentiment analysis on our dataset. Additionally, our findings suggest that maintaining a BPE vocabulary size of 12,000 or more is necessary to achieve optimal performance for both models. Furthermore, the performance of our two BPE models is comparable to that of other M-BERT models tested on low-resource languages in previous research. Based on our findings, we recommend further research to explore the potential of BPE models in other low-resource languages. We believe that this approach could be particularly beneficial for languages with limited resources and data, as it allows for effective modeling of sub-word units and can improve the quality of image-based processing in these languages.

Author Contributions: Conceptualization, J.T.H.K. and W.K.W.; methodology, J.T.H.K. and F.H.J.; software, J.T.H.K.; validation, I.G.D.N., Y.M. and W.K.W.; formal analysis, J.T.H.K. and I.Y.N.; investigation, I.Y.N.; resources, J.T.H.K.; data curation, J.T.H.K.; writing—original draft preparation, J.T.H.K., W.K.W. and F.H.J.; writing—review and editing, J.T.H.K., F.H.J., I.Y.N., I.G.D.N. and Y.M.; visualization, J.T.H.K.; supervision, F.H.J., I.Y.N. and W.K.W.; project administration, W.K.W. and I.G.D.N.; funding acquisition, W.K.W. and I.G.D.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Fundamental Research Grant Scheme (FRGS) under grant FRGS/1/2020/ICT06/CURTIN/02/1 and in part by Universitas Indonesia's International Indexed Publication (PUTI) Q1 Grant, year 2020, contract number NKB-1409/UN2.RST/HKP.05.00/2020.

Data Availability Statement: The data presented in this study are openly available in Zenodo at DOI: 10.5281/zenodo.7737457.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Saud, M.; Mashud, M.; Ida, R. Usage of social media during the pandemic: Seeking support and awareness about COVID-19 through social media platforms. *J. Public Aff.* **2020**, *20*, e2417. [CrossRef]
2. Samuel, J.; Rahman, M.M.; Ali, G.M.N.; Samuel, Y.; Pelaez, A.; Chong, P.H.J.; Yakubov, M. Feeling positive about reopening? New normal scenarios from COVID-19 US reopen sentiment analytics. *IEEE Access* **2020**, *8*, 142173–142190. [CrossRef] [PubMed]
3. Mourad, A.; Srour, A.; Harmanani, H.; Jenainati, C.; Arafeh, M. Critical impact of social networks infodemic on defeating coronavirus COVID-19 pandemic: Twitter-based study and research directions. *IEEE Trans. Netw. Serv. Manag.* **2020**, *17*, 2145–2155. [CrossRef]
4. Agathangelou, P.; Katakis, I. Balancing between holistic and cumulative sentiment classification. *Online Soc. Netw. Media* **2022**, *29*, 100199. [CrossRef]
5. Hasan, A.; Moin, S.; Karim, A.; Shamshirband, S. Machine learning-based sentiment analysis for twitter accounts. *Math. Comput. Appl.* **2018**, *23*, 11. [CrossRef]
6. Mao, Y.; Menchen-Trevino, E. Global news-making practices on Twitter: Exploring English-Chinese language boundary spanning. *J. Int. Intercult. Commun.* **2019**, *12*, 248–266. [CrossRef]
7. Junaini, S.N.; Hwey, A.L.T.; Sidi, J.; Rahman, K.A. Development of Sarawak Malay local dialect online translation tool. In Proceedings of the 2009 International Conference on Computer Technology and Development, Kota Kinabalu, Malaysia, 13–15 November 2009; pp. 459–462.

8. Hijazi, M.H.A.; Libin, L.; Alfred, R.; Coenen, F. Bias aware lexicon-based Sentiment Analysis of Malay dialect on social media data: A study on the Sabah Language. In Proceedings of the 2016 2nd International Conference on Science in Information Technology (ICSITech), Balikpapan, Indonesia, 26–27 October 2016; pp. 356–361.
9. Khaw, Y.M.J.; Tan, T.P. Hybrid approach for aligning parallel sentences for languages without a written form using standard Malay and Malay dialects. In Proceedings of the 2014 International Conference on Asian Language Processing (IALP), Kuching, Malaysia, 20–22 October 2014; pp. 170–174.
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the AACL HLT 2019 Conference of the North American Chapter of the Association for Computational Linguistic Humanity Language Technology, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
11. Fujihira, K.; Horibe, N. Multilingual Sentiment Analysis for Web Text Based on Word to Word Translation. In Proceedings of the 9th International Congress on Advanced Applied Informatics (IIAI-AAI), Kitakyushu, Japan, 1–15 September 2020; pp. 74–79.
12. Baliyan, A.; Batra, A.; Singh, S.P. Multilingual sentiment analysis using RNN-LSTM and neural machine translation. In Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 17–19 March 2021; pp. 710–713.
13. Afroz, N.; Boral, M.; Sharma, V.; Gupta, M. Sentiment Analysis of COVID-19 nationwide lockdown effect in India. In Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 25–27 March 2021; pp. 710–713.
14. Marathe, A.; Mandke, A.; Sardeshmukh, S.; Sonawane, S. Leveraging Natural Language Processing Algorithms to Understand the Impact of the COVID-19 Pandemic and Related Policies on Public Sentiment in India. In Proceedings of the 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 25–27 June 2021; pp. 1–5.
15. Pellert, M.; Lasser, J.; Metzler, H.; Garcia, D. Dashboard of sentiment in Austrian social media during COVID-19. *Front. Big Data* **2020**, *3*, 32. [CrossRef] [PubMed]
16. Jayasurya, G.G.; Kumar, S.; Singh, B.K.; Kumar, V. Analysis of public sentiment on COVID-19 vaccination using twitter. *IEEE Trans. Comput. Soc. Syst.* **2021**, *9*, 1101–1111. [CrossRef]
17. Aygun, I.; Kaya, B.; Kaya, M. Aspect Based Twitter Sentiment Analysis on Vaccination and Vaccine Types in COVID-19 Pandemic with Deep Learning. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 2360–2369. [CrossRef] [PubMed]
18. Yang, X.; Sornlertlamvanich, V. Public Perception of COVID-19 Vaccine by Tweet Sentiment Analysis. In Proceedings of the 2021 International Electronics Symposium (IES), Surabaya, Indonesia, 29–30 September 2021; pp. 151–155.
19. Alharbi, A.; de Doncker, E. Twitter Sentiment Analysis with a Deep Neural Network: An Enhanced Approach using User Behavioral Information. *Cogn. Syst. Res.* **2018**, *54*, 50–61. [CrossRef]
20. Jacovi, A.; Shalom, O.S.; Goldberg, Y. Understanding convolutional neural networks for text classification. *arXiv* **2018**, arXiv:1809.08037.
21. Merdivan, E.; Vafeiadis, A.; Kalatzis, D.; Hanke, S.; Kroph, J.; Votis, K.; Giakoumis, D.; Tzovaras, D.; Chen, L.; Hamzaoui, R.; et al. Image-based Text Classification using 2D Convolutional Neural Networks. In Proceedings of the 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Leicester, UK, 19–23 August 2019; pp. 144–149.
22. Srivastava, A.; Singh, V.; Drall, G.S. Sentiment analysis of twitter data: A hybrid approach. *Int. J. Healthc. Inf. Syst. Inform. (IJHISI)* **2019**, *14*, 1–16. [CrossRef]
23. Suri, V.; Arora, B. A Review on Sentiment Analysis in Different Language. In Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; pp. 1–9.
24. Abu Bakar, M.F.R.; Idris, N.; Shuib, L.; Khamis, N. Sentiment Analysis of Noisy Malay Text: State of Art, Challenges and Future Work. *IEEE Access* **2020**, *8*, 24687–24696. [CrossRef]
25. Al-Saffar, A.; Awang, S.; Tao, H.; Omar, N.; Al-Saiagh, W.; Al-Bared, M. Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm. *PLoS ONE* **2018**, *13*, e0194852. [CrossRef] [PubMed]
26. Chekima, K.; Alfred, R. Sentiment analysis of Malay social media text. In Proceedings of the International Conference on Computational Science and Technology, Kuala Lumpur, Malaysia, 29–30 November 2017; pp. 205–219.
27. Zabha, N.I.; Ayop, Z.; Anawar, S.; Hamid, E.; Abidin, Z.Z. Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 346–351. [CrossRef]
28. Bakar, M.F.R.A.; Idris, N.; Shuib, L. An Enhancement of Malay Social Media Text Normalization for Lexicon-Based Sentiment Analysis. In Proceedings of the 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 15–17 November 2019; pp. 211–215.
29. bin Rodzman, S.B.; Rashid, M.H.; Ismail, N.K.; Abd Rahman, N.; Aljunid, S.A.; Abd Rahman, H. Experiment with Lexicon Based Techniques on Domain-Specific Malay Document Sentiment Analysis. In Proceedings of the 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Kota Kinabalu, Malaysia, 27–28 April 2019; pp. 330–334.
30. Nabiha, A.; Mutalib, S.; Ab Malik, A.M. Sentiment Analysis for Informal Malay Text in Social Commerce. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), Virtual, 8–9 September 2021; pp. 1–6.
31. Yilmaz, S.F.; Kaynak, E.B.; Koç, A.; Dibeklioğlu, H.; Kozat, S.S. Multi-Label Sentiment Analysis on 100 Languages With Dynamic Weighting for Label Imbalance. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *34*, 331–343. [CrossRef] [PubMed]

32. Sazzed, S.; Jayarathna, S. A Sentiment Classification in Bengali and Machine Translated English Corpus. In Proceedings of the 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), Los Angeles, CA, USA, 30 July–1 August 2019; pp. 107–114.
33. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual is Multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 4996–5001.
34. Islam, M.S.; Amin, M.R. Sentiment analysis in Bengali via transfer learning using multi-lingual BERT. In Proceedings of the 23rd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 19–21 December 2020; pp. 19–21.
35. Sabri, N.; Edalat, A.; Bahrak, B. Sentiment Analysis of Persian-English Code-mixed Texts. In Proceedings of the 2021 26th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, 3–4 March 2021; pp. 1–4.
36. Fimoza, D.; Amalia, A.; Harumy, T.H.F. Sentiment Analysis for Movie Review in Bahasa Indonesia Using BERT. In Proceedings of the 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Medan, Indonesia, 11–12 November 2021; pp. 27–34.
37. Le, A.P.; Vu Pham, T.; Le, T.V.; Huynh, D.V. Neural Transfer Learning For Vietnamese Sentiment Analysis Using Pre-trained Contextual Language Models. In Proceedings of the 2021 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), Soyapango, El Salvador, 16–17 December 2021; pp. 1–5.
38. Kong, J. MyCovid-Senti. 2022. Available online: https://github.com/z3fei/Malaysia-COVID-19-Tweet-ID/tree/main/MyCovid-Senti (accessed on 20 December 2022).
39. Schlosser, S.; Toninelli, D.; Cameletti, M. Comparing methods to collect and geolocate tweets in Great Britain. *J. Open Innov. Technol. Mark. Complex.* **2021**, *7*, 44. [CrossRef]
40. Mohammad, S. A practical guide to sentiment annotation: Challenges and solutions. In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, San Diego, CA, USA, 16 June 2016; pp. 174–179.
41. Gage, P. A new algorithm for data compression. *C Users J.* **1994**, *12*, 23–38.
42. Gowda, T.; May, J. Finding the optimal vocabulary size for neural machine translation. *arXiv* **2020**, arXiv:2004.02334.
43. Kumar, A.; Singh, T.; Vishwakarma, D.K. Intelligent Transport System: Classification of Traffic Signs Using Deep Neural Networks in Real Time. In *Advances in Manufacturing and Industrial Engineering*; Springer: Singapore, 2021; pp. 207–219.
44. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is multilingual BERT? *arXiv* **2019**, arXiv:1906.01502.
45. Jose, N.; Chakravarthi, B.R.; Suryawanshi, S.; Sherly, E.; McCrae, J.P. A survey of current datasets for code-switching research. In Proceedings of the 2020 6th international conference on advanced computing and communication systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 136–141.
46. Willingham, D. Transformer Models. Github. 2022. Available online: https://github.com/matlab-deep-learning/transformer-models/releases/tag/1.2 (accessed on 4 January 2022).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.