




A Multilingual Corpus for Panic and Worry in Code-Mixed Tweets by VADER Sentiment Analysis

Razailin Abdul Rashid¹ , Siti Hafizah Ab Hamid^{2,*} , Faisal Fahmi³ 

^{1,2}*Department of Software Engineering, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia*

³*Departemen Ilmu Informasi dan Perpustakaan, Fakultas Ilmu Sosial and Ilmu Politik, Universitas Airlangga, Kampus B. Jl. Dharmawangsa Dalam, Surabaya 60286, Jawa Timur, Indonesia*

(Received: June 3, 2024; Revised: June 22, 2024; Accepted: July 5, 2024; Available online: July 27, 2024)

Abstract

The phenomenon of code-mixing in online discourse, on platforms such as X, offers an interesting setting to detect preliminary markers of anxiety within diverse linguistic expressions. The usage of more than one language within a single text or tweet necessitates the creation of a multilingual corpus to identify initial indicators of anxiety in code-mixed texts or tweets, contributing to a comprehensive understanding of mental health in the digital age. Existing research on code-mixed textual context primarily centres on code-mixed language of English with Spanish or Hindi, leaving a gap in our comprehension of other code-mixed languages, in particular; English with Malay or Indonesian language. Thus, our study focuses on anxiety-related linguistic expressions in Malay and Indonesian languages, such as ‘bimbang’, ‘bingung’, ‘panik’, ‘gelisah’, ‘cemas’, ‘takut’, ‘kacau’, ‘gemetar’, ‘gugup’, ‘teror’ and occasionally the usage of slangs such as ‘neves’, ‘gabra’, and ‘cape bgt’. In this paper, we introduce CORPUS4PANWO, an annotated sentiment-driven multilingual corpus for panic and worry detection in tweets. To experiment the corpus, we applied a corpus-based sentiment analysis utilizing VADER on diverse events, achieving accuracy of between 76.6% - 88.0% when used on tweets in negative circumstances. The corpus is a valuable resource for Southeast Asian linguistics, enabling exploration of emotional expression in diverse contexts.

Keywords: Anxiety Expression, Code-Mixing, Corpus Development, Panic and Worry Emotion VADER

1. Introduction

The linguistic phenomenon of code-mixing, prevalent on social platforms like X, involves the use of multiple languages within the same conversational context. Referred to interchangeably as mixed language, code-mixing, or code-switching, these terms describe distinct language behaviours. Code-mixing blends two languages or codes seamlessly without changing the topic, showcasing linguistic fluidity. In contrast, code-switching refers to shifting between languages or language varieties within discourse, illustrating dynamic language use in diverse social interactions [1], [2]. This practice extends beyond individual words to include nouns, verbs, adjectives, and adverbs, occurring both within sentences and across different linguistic contexts, highlighting its adaptability and complexity [3], [4]. Hence, recognizing the prevalence of code-mixed languages, creating a dedicated corpus becomes crucial. Such a corpus would facilitate the identification of initial indicators of anxiety in the diverse linguistic expressions encountered on social media platforms, contributing to a more comprehensive understanding of mental health in the digital age.

According to [3], a corpus is described as a “collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety”. However, most corpora are typically monolingual, focusing on one singular language. References [4], [5], [6], [7], [8] collectively contribute to advancing sentiment analysis in multilingual contexts, emphasizing the significance of a multilingual corpus in understanding emotions across diverse cultures and languages. In the context of our multilingual corpus, existing research on sentiment analysis within code-mixing contexts by [4], [5], [6] has explored multilingual environments but often overlooks the prevalent usage of slang languages, especially in code-mixing involving English

*Corresponding author: Siti Hafizah Ab Hamid (sitihafizah@um.edu.my)

 DOI: <https://doi.org/10.47738/jads.v5i3.259>

This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights

with Malay or Indonesian languages. Hence, there is a need to investigate these linguistic phenomena comprehensively to capture the full spectrum of emotional expression in diverse cultural and linguistic settings.

Our study centres on two key objectives. First, identifying tweets, formal and informal sentences that contains panic and worry instances or nuances, gaining insight into how emotions manifest in real-time communication. Second, experimenting the corpus across ten events to investigate the ability of the corpus to identify panic and worry sentiments in diverse events. In this study, our primary contribution lies in the creation of a multilingual corpus designed to facilitate the identification of panic and worry sentiments within the context of code-mixed text posts. The corpus is crafted to address the unique linguistic challenges posed by the code-mixing phenomenon, with a focus on English, Malay, and Indonesian languages.

The contributions of this paper are significant in advancing the field of sentiment analysis in code-mixed languages. Firstly, we have developed a multilingual corpus that is designed for the identification of panic and worry sentiments in code-mixed tweets. This corpus includes a variety of words and phrases from three different languages; which are English, Malay, and Indonesian languages. Secondly, we have validated words that, while not synonymous with panic and worry, still convey these emotions in a textual context. This validation process ensures that the corpus accurately reflects the nuances of emotional expression in code-mixed text, making it a valuable tool for researchers and practitioners in the field.

The remainder of the paper is structured as follows: Section 2 provides a concise overview of existing works in code-mixing multilingual sentiment analysis, including approaches to analysing panic and worry in social media content, advancements in corpus development, and associated challenges; Section 3 details the process involved in gathering the dataset for the corpus development; Section 4 showcases the structure of the corpus; Section 5 is the experimentation of the corpus; Section 6 focuses on the data analysis and potential applications of the corpus and finally; Section 7 and 8, the limitations of the study and conclusion of this research.

2. Related Works

Significant advancements have been made in the field of sentiment analysis and natural language processing, particularly in addressing challenges associated with code-mixed languages and under-resourced languages. Multilingual corpus-based sentiment analysis is a method used to analyse sentiment in multiple languages by utilizing annotated corpora. This approach involves creating datasets in different languages and using them to train sentiment analysis models. The goal is to develop models that can accurately classify subjective information as positive, negative, or neutral in various languages. Earlier research has given us valuable insights into sentiment analysis in situations where people mix different languages.

Reference [4] proposed a novel approach called Sentiment Analysis of Code-Mixed Text (SACMT) to classify sentences into positive, negative, or neutral sentiments using contrastive learning. Their method leverages shared parameters of Siamese networks to map sentences from code-mixed and standard languages into a common sentiment space, resulting in improved accuracy and F-score compared to existing approaches. Similarly, [5] conducted a systematic review of multilingual sentiment analysis techniques for under-resourced languages. They evaluated over 35 studies and highlighted the importance of developing models for languages with limited resources. Their work contributes to theoretical literature reviews and emphasizes the need for appropriate strategies for sentiment analysis in diverse linguistic contexts. Additionally, [6] addressed the detection of panic potential in social media messages during disaster situations. They proposed methods for quantifying panic potential and rumour potential in tweets, aiming to improve crisis communication and management. Their research underscores the importance of understanding and mitigating the impact of emotionally laden information on public sentiment during emergencies.

However, none of these studies included an element present in our work: the identification of singular or compound words that, while not synonymous with panic and worry terms, express these emotions. Although these studies offer valuable insights, our research goes further by incorporating this aspect, thereby enhancing the scope of multilingual sentiment analysis methodologies. The linguistic phenomenon of code-mixing on social platforms such as X, encompasses the use of more than one language in the same conversational event. It is referred to by various terms, such as mixed language, code-mixing, and code-switching. Code-mixing involves the mixing of two codes or languages

without a change of topic, while code-switching entails a change from one language or language variety to another [1], [2]. This phenomenon can involve mixing nouns, verbs, adjectives, and adverbs, and can occur both inter-sentential and intra-lexically, indicating its flexibility and complexity [7], [8]. Hence, recognizing the prevalence of code-mixed languages, creating a dedicated corpus becomes crucial. Such a corpus would facilitate the identification of initial indicators of anxiety in the diverse linguistic expressions encountered on social media platforms, contributing to a more comprehensive understanding of mental health in the digital age.

3. Methodology

In this study, we present an annotated sentiment-driven multilingual corpus designed for detecting panic and worry sentiments in tweets. Our methodology involved gathering synonymous terms from authoritative sources and incorporating slang from local movie dialogues to enrich expressions. Inclusion of sample sentences from dictionaries further increased textual representations and we also targeted student-specific tweets using academic-context keywords. Notably, the resulting dataset showcased diverse linguistic expressions, including code-mixed styles. Additionally, we expanded the corpus to include tweets expressing panic and worry sentiments without the use of words synonymous with panic and worry terms, leveraging manual annotation and expert supervision. In the next subsection, we further detailed the processes done during the corpus development, as depicted in figure 1.

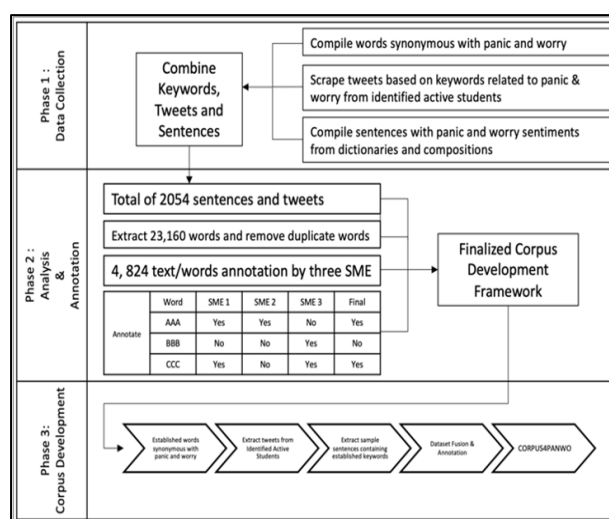


Figure 1. Corpus Development

3.1. Combine panic and worry keywords, tweets, and sentences

As shown in figure 1, we began with synonymous terms of panic and worry in the Malay language by accessing the authoritative 'Kamus Dewan' database hosted on the language portal of Dewan Bahasa dan Pustaka (DBP) [9]. For English keywords, synonyms were sourced from Dictionary.com [10], while Indonesian keywords were obtained from 'KBBI Daring' [11] and 'Tesaurus Tematis Bahasa Indonesia' [12], the official word search page for the 'Kamus Besar Bahasa Indonesia' (KBBI). We also included slang words conveying panic and worry sentiments from local Malaysian movie dialogues [13]. This approach recognizes that movies often portray how people truly converse in everyday situations, incorporating informal expressions and language mixtures. So, an inspection of conversations in movies aids more interest in examining the two emotions. Additionally, we expanded our list of keywords by prompting ChatGPT by Open AI. Table 1 displays the sample of synonymous words.

Table 1. Sample of Words Synonymous with Panic and Worry Emotions

	Malay language	Indonesian language	English language
Panic keywords	bimbang, cemas, cuak, gabra, gelisah, gila-gila, glabah, gugup, huru-hara, kalut	baper, begadang, bete, bingung jadi satu, cemas, cemas berat, deg-degan, galau	afraid, agitation, alarm, aghast, anxious, become hysterical, chicken out, dread, fear

Worry keywords	bingung, dalam keadaan panik, gelabah, gelisah, kacau, khuatir, neves, risau	cemas, khawatir, kekhawatiran, gugup, kegugupan, ketakjuban, ketidaktenangan	apprehension, aroused, bad news, beside oneself, bothered, vexation	apprehensive,
----------------	--	--	---	---------------

In addition to collecting words, we gathered sample sentences provided by the dictionaries, as they are established to convey the emotions of panic and worry in textual form as indicated in [figure 2](#). Both the compiled keywords and sentences served as the foundation for developing the corpus related to panic and worry, ensuring a general representation of panic and worry expressions in textual format.

*Apa yg dirisaukan tentang perkara ini?
Apa yg kamu bimbangkan sekarang?
Apabila berita itu disiarkan, rasa panik melanda bandar-bandar besar negara itu.
He became more and more perturbed as the hours went by.
He bolted from the chair on hearing the scream.
He had a momentary feeling of panic.*

Figure 2. Sample of dictionary sentences

3.2. Student-specific tweet collection process

The main purpose of the creation of our multilingual corpus is to identify initial indicators of anxiety in diverse linguistic expressions. Hence, following up the previous process, we need the corpus to include data relating to how students express their panic and worry emotions in textual conversation, mainly using the social platform X as their outlet.

To construct this targeted dataset, we initiated our data collection by identifying X users that are students, within a specified timeframe, ranging from January 1, 2018, to December 4, 2022. This identification process was crucial for ensuring the relevance and specificity of the collected tweets consists of students actively engaged in academic discourse on X. We employed a set of keywords associated with academic contexts. These keywords included 'kursus', 'pensyarah', 'markah', assignment, 'projek', exam, test, 'universiti', and 'pelajar Malaysia'. This process yielded a result of 91,497 tweets. Next, we filter these 91,497 tweets by using the established keywords that we have compiled in the previous phase.

[Figure 3](#) displays the pseudocode of how we performed this filtering process using the Python language. The pseudocode begins by importing the necessary libraries, pandas for handling DataFrames, and re for regular expressions. It then defines the list of keywords to check, referred to as `word_list`. Following this, the code loads the dataset containing tweets into a pandas DataFrame called `df`. Two new columns are created in the DataFrame: 'Contains_PanWo' is initialized with 'No' and will track if a tweet contains any word from the list, and 'PanWo_Word' is initialized as an empty string to store the specific word from the list found in a tweet.

```
1. Import necessary libraries (pandas and re).
2. Define the list of words to check (word_list).
3. Load the dataset into a pandas DataFrame (df).
4. Create two new columns in the DataFrame:
   - 'Contains_PanWo' and initialize it with 'No'
   - 'PanWo_Word' and initialize it with an empty string
5. Iterate through each tweet in the DataFrame:
   a. Iterate through each word in the word_list:
      i. Check if the word is present in the current tweet as a whole word using regular expressions.
      ii. If the word is found:
          - Update 'Contains_PanWo' to 'Yes'
          - Update 'PanWo_Word' to the current word
          - Break the inner loop to avoid updating with multiple words
6. Save the original first column along with the updated DataFrame to a new CSV file.
7. End.
```

Figure 3. Pseudocode for keyword matching in tweets

The code then iterates through each tweet in the DataFrame, and for each tweet, it iterates through the `word_list`. Using regular expressions, it checks if the current word from the list is present in the tweet as a whole word. If a match is found, it updates 'Contains_PanWo' to 'Yes' and 'PanWo_Word' to the matched word. The inner loop is broken to prevent updating with multiple words for a single tweet. After processing all tweets, the pseudocode saves the original

first column along with the updated DataFrame to a new CSV file. To provide clarity and aid reproducibility, we included the following Python code snippet in [figure 4](#) that corresponds to the pseudocode described.

```
import pandas as pd
import re

# Define keywords list
word_list = ['keyword1', 'keyword2', 'keyword3'] # Add actual keywords here

# Load dataset
df = pd.read_csv('path_to_tweet_dataset.csv')

# Initialize new columns
df['Contains_PanWo'] = 'No'
df['PanWo_Word'] = ''

# Iterate through tweets and keywords
for i, tweet in df.iterrows():
    for word in word_list:
        if re.search(r'\b' + re.escape(word) + r'\b', tweet['tweet_column'], re.IGNORECASE):
            df.at[i, 'Contains_PanWo'] = 'Yes'
            df.at[i, 'PanWo_Word'] = word
            break # Stop after the first match to avoid multiple keywords per tweet

# Save updated DataFrame
df.to_csv('path_to_updated_dataset.csv', index=False)
```

Figure 4. Code snippet in Python

This process yielded 2054 rows of results, matching 134 keywords out of 216 keywords used to filter the tweets. Table 7 in Appendix A displays the breakdown of the keywords. These tweets are included as well within the corpus. Notably, because these tweets are written by local students, they are often written in code-mixed style. This also adds to the varied data that we have now compiled within the corpus.

Now, the corpus contains singular word, compound words, sentences and tweets that contains at least one of the synonymous words of panic and worry that were the foundation of the corpus at the beginning of the corpus development. But sometimes, humans tend to subtly hint on their emotions without using words that are synonymous with the emotion. And so, we further expanded our corpus to include tweets that do not contain any of the synonymous words yet still emits the emotions of panic and worry.

To do this, we engaged a Subject Matter Expert (SME) in this process. The SME, a local Malaysian student with bilingual proficiency, and familiarity with social messaging, provided an additional layer of validity to the corpus. Their role included the manual annotation of panic or worry sentiment of randomly selected 1000 tweets using Google Form. The annotation effort took one week, and the tweets are incorporated as well in the corpus.

For the Indonesian students' tweets, they were already identified prior to extraction via 'menfess', as groups of students will often anonymously post their tweets for confidentially purposes. Under the supervision of Dr Faisal Fahmi from the Department of Information Science and Library at Universitas Airlangga (UNAIR), a group of undergraduate students conducted a systematic data collection process. This process involved not only employing targeted keyword searches but also extracting data from accounts utilizing 'menfess' (short for mention and confess).

3.3. Non-synonymous words of panic and worry

To identify words conveying sentiments of panic and worry but not synonymous with "panic" and "worry," we conducted a detailed annotation process on a dataset initially comprising 23,160 words, resulting in 4,824 unique singular or compound entries after removing duplicates. This step aimed to pinpoint words that encapsulate these emotional states without directly mirroring their typical expressions.

The identification and validation of these non-synonymous words were crucial for enhancing the credibility of our corpus development. We implemented an Inter-Annotator Agreement (IAA) process, following methodologies similar to those described in [14], [15] involving three Subject Matter Experts (SMEs). These SMEs, local postgraduate students from Malaysian universities with expertise in machine learning, software testing, and natural language processing, contributed their bilingual proficiency and cultural awareness to refine the annotation process. Their familiarity with both English and Malay languages and their understanding of social messaging nuances were instrumental in this refinement.

Ensuring annotator reliability was crucial for maintaining the accuracy and consistency of sentiment annotations in our corpus. Each annotator, selected from local Malaysian postgraduate university students, brought a deep understanding of social media dynamics and informal code-mixing practices from their own daily interactions. Regular

communications were maintained throughout the annotation process to clarify guidelines and ensure alignment in sentiment identification within the allocated timeframe. This approach facilitated a shared understanding of emotional nuances in texts, contributing to the reliability and validity of our sentiment analysis. These efforts collectively aimed to enhance the credibility of our corpus for research applications.

The IAA process resulted in a 19.3% inter-annotator agreement percentage, reflecting consensus on 936 words out of the total 4,824 annotated entries. This approach ensured consistent identification of emotional nuances related to panic and worry across annotators, thereby supporting the reliability and relevance of our corpus for sentiment analysis in multilingual contexts.

To quantify the level of agreement beyond chance, Cohen's Kappa coefficient [16] was calculated. The calculated Cohen's Kappa coefficient is approximately 0.28. This indicates a slight level of agreement beyond chance, which can be attributed to several factors. Firstly, the subjective nature of sentiment analysis often leads to varying interpretations of the same text, especially when dealing with nuanced emotions like panic and worry. Secondly, the relatively small number of annotators also contribute to a lower Kappa value. Despite the lower Kappa value, the annotation process still produced a panic and worry lexicon, consisting of 936 curated words or phrases, capturing the nuanced expressions of panic and worry sentiments in the academic context within university students' tweets, which are also added to the corpus.

These metrics highlight challenges in achieving high agreement levels but emphasize the strength of our annotation approach in capturing various expressions of panic and worry in university students' tweets, despite financial and time constraints. To improve annotation consistency, future efforts could involve increasing annotator numbers, refining guidelines, and adding training sessions focused on interpreting nuanced emotional cues. These steps aim to enhance the reliability and relevance of our corpus for sentiment analysis in multilingual contexts.

4. Results and Discussion

4.1. Result: CORPUS4PANWO

The finalized dataset contains a total of 3,495 rows of data. It includes a mix of languages such as English, Indonesian, and Malay, with some texts containing a blend of languages. Figure 5 shows how the data in the corpus are structured. The texts fall into different textual categories, including compound words, sentences, single words, and tweets. The data comes from various sources, including an AI tool, dictionary references, and contributions from a Subject Matter Expert (SME). This corpus provides a diverse and comprehensive dataset for studying emotions, particularly focusing on panic and worry.

PanWo	Language	Category	Source
dalam keadaan panik	Malay	compound	Dictionary
huru-hara	Malay	compound	Dictionary
kelam kabut	Malay	compound	Dictionary
gabra	Malay	single	Dictionary
menyimpan dalam hati	Malay	compound	Dictionary
sakit kepala	Malay	compound	Dictionary
glabah	Malay	single	Dictionary
susah hati	Malay	compound	Dictionary
tak dapat tidur	Malay	compound	Dictionary
tak sedap hati	Malay	compound	Dictionary
tak tentu arah	Malay	compound	Dictionary
kecut	Malay	single	Dictionary
terkinja-kinja	Malay	compound	Dictionary

Figure 5. Structure of CORPUS4PANWO

This diversity is reflected in table 2, which shows that the texts predominantly consist of Indonesian (1,216 entries), English (1,181 entries), and Malay (896 entries), with a smaller portion being Mixed (198 entries) and Blended (4 entries) languages. The dataset also includes different types of texts, as outlined in table 3. These range from single words (1,127 entries) and compound words (159 entries) to full sentences (201 entries) and tweets (2,008 entries). The sources of these texts are varied, as detailed in table 4, with a significant portion manually annotated by a Subject Matter Expert (SME) (2,943 entries), alongside texts generated by AI tools (111 entries) and dictionary references (441 entries). This structured and detailed corpus allows researchers to explore emotional expressions related to panic and

worry across different languages and text types, making it a valuable resource for sentiment analysis studies. The breakdown of each metadata and its description within the corpus is shared in [table 2](#), [table 3](#), and [table 4](#) respectively.

Table 2. Language distribution

Category	Description	Amount
Blended	Texts that exhibit a blend of multiple languages	4
English	Texts predominantly in the English language	1,181
Indonesian	Texts predominantly in the Indonesian language	1,216
Malay	Texts predominantly in the Malay language	896
Mixed	Texts that incorporate a mixture of languages	198

Table 3. Text category breakdown

Category	Description	Amount
Compound	Texts consisting of compound words.	159
Sentence	Complete sentences expressing sentiments.	201
Single	Texts comprised of single words conveying emotions	1127
Tweet	Short, concise messages typically found on social media platforms	2008

Table 4. Source composition

Category	Description	Amount
AI Tool	Texts generated through automated tools leveraging artificial intelligence.	111
Dictionary	Texts sourced from lexical references or predefined lists.	441
SME	Texts manually annotated and evaluated by a Subject Matter Expert (SME), ensuring human context and interpretation	2,943

The structured nature of this corpus extends its utility to researchers exploring diverse facets of emotion analysis. Its organization allows researchers to tailor their investigations by selecting specific languages, opting for types of sentences, or utilizing the corpus. This flexibility empowers researchers to navigate and manipulate the corpus according to the unique requirements of their studies, enhancing the applicability of the dataset across various research endeavours.

4.1.1. Experimentation of the Corpus

We conducted a sentiment analysis using CORPUS4PANWO on a variety of events. These events were chosen based on their significance and the level of public discourse they generated. Each event was carefully selected to reflect a wide range of emotions and situations that people discuss online, ensuring a comprehensive analysis of different sentiment expressions in various contents.

For example, we looked at things like cyberattacks, sports events, and protests, as well as political events and health crises. We also explored how people feel about social issues like climate change and movements like #MeToo. Additionally, we examined sentiments around holidays like Christmas and New Year, as well as the emotions tied to Father's Day. We hypothesized that events characterized by positive circumstances have lower levels of concern detected in their tweets.

By selecting this diverse set of events, our aim was to ensure a comprehensive examination of emotional tones in tweets across various subjects. We sought to capture a broad spectrum of human experiences and reactions, allowing us to understand how different events influence public sentiment. This approach enabled us to identify patterns and variations in emotional responses, providing valuable insights into how people express their feelings about a wide array of topics on social media platforms. [Table 5](#) displays the event name and the description of each of the event in detail.

Table 5. Event name and description

Event	Description
BSI Ransomware	The BSI Ransomware incident refers to a cyberattack that targeted Bank Syariah Indonesia (BSI), the largest Islamic bank in Indonesia, in May 2023.
FIFA World Cup 2022	The 2022 FIFA World Cup, held in Qatar from November 20 to December 18, 2022
Iran Protests 2022	The Iran protests of 2022, also known as the Mahsa Amini protests, were a series of widespread demonstrations that erupted in Iran in September 2022 and continued into 2023
US-Afghan War	The US-Afghan War, also known as the War in Afghanistan or the 2001 invasion of Afghanistan, was the longest war in US history, spanning over two decades from 2001 to 2021
Australian 2019 Polls	The federal election was held on Saturday May 18 2019.
Covid-19	COVID-19, caused by the SARS-CoV-2 virus, is a highly contagious respiratory illness that emerged in late 2019.
Climate Change	Users voicing out their concern about the effects of climate change
Me Too Movement	The #MeToo movement is a powerful social movement and awareness campaign against sexual abuse, sexual harassment, and rape culture
Sri Lanka Crisis	Sri Lanka faces a crippling economic and humanitarian crisis due to mismanagement, pandemic impact, and global factors, seeking international support while navigating a complex path to recovery.
Xenophobia	Xenophobia is the intense dislike or fear of people from other countries or cultures.
Christmas 2022	Christmas is a festival that is celebrated on 25th of December every year across the world.
New Year 2021	Celebrated on January 1st, marks the beginning of a new calendar year. People often gather with friends and family to countdown to midnight, enjoy fireworks displays, and toast to new beginnings.
Father's Day	Father's Day is a holiday of honouring fatherhood and paternal bonds, as well as the influence of fathers in society.

We conducted a sentiment analysis on X datasets using our corpus and VADER [17], to recognize the emotional nuances within various contemporary events. VADER, which stands for Valence Aware Dictionary and sEntiment Reasoner, is a pre-built, rule-based sentiment analysis tool specifically designed for social media text. Reference [18] in their work used VADER to automatically determine whether tweets are positive, negative, or neutral. This tool helped them label tweets with these sentiments.

After labelling the tweets, they used machine learning algorithms like Support Vector Machines and Random Forests to train models that can predict the sentiment of new tweets. They compare the performance of these models by looking at their F1 scores to find the best one. Whilst in our sentiment analysis, we aimed to assign sentiment scores to words and phrases within the X data without the usage of any machine learning algorithms, reflecting the intensity of panic and worry sentiments. This scoring process was twofold, involving both our corpus and the VADER sentiment analysis tool.

Firstly, the corpus-based scoring involved using a list of words and phrases specifically associated with panic and worry sentiments. For each word or phrase identified in the X data, if it matched an entry in our corpus, we assigned a sentiment score of -1 to indicate a negative sentiment, reflecting the intensity of panic and worry. Secondly, VADER was used to assign sentiment scores to entire sentences or texts, evaluating the overall sentiment. The tool's positive scoring specifically helps identify and emphasize positive sentiments within the tweets analysed.

The sentiment analysis results, as illustrated in figure 6, underscores the importance of corpus-based sentiment analysis in uncovering emotional expressions in X data. The detailed breakdown of sentiments by events in table 6 highlights the flexibility of the corpus in capturing emotional nuances across diverse events, reinforcing its usefulness in understanding the emotional currents in contemporary discourse.

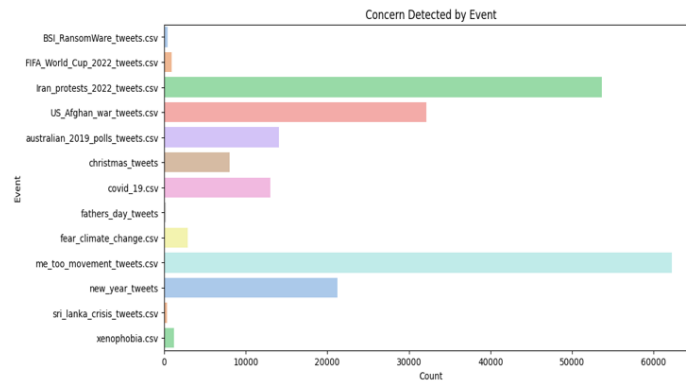


Figure 6. Corpus-based sentiment analysis across events

Table 6. Breakdown of sentiments by events

Event	Concern Detected	Total Tweets	Percentage
BSI Ransomware	456	938	48.6%
FIFA World Cup 2022	939	2021	46.5%
Iran Protests 2022	53612	61476	87.2%
US-Afghan War	32146	49614	64.8%
Australian 2019 Polls	14089	22422	62.8%
Covid-19	13008	14781	88.0%
Climate Change	2902	3934	73.8%
Me Too Movement	62253	93072	66.9%
Sri Lanka Crisis	362	500	72.4%
Xenophobia	1182	1543	76.6%
Christmas 2022	8026	20000	40.1%
New Year 2021	21216	109978	19.3%
Father's Day	258	990	26.1%

The level of 'Concern Detected' in different events reflects the extent of worry, panic, or unease among people as depicted on [figure 7](#). Events like the Covid-19 pandemic and the Iran Protests of 2022 caused high concern due to their serious impact on society. They raised alarms about health and human rights, making people anxious. On the other hand, events like the Father's Day and New Year celebrations typically evoke less concern, as they're seen as fun and enjoyable occasions. The level of concern also depends on the ongoing issues in society. For example, events like the US-Afghan War, Climate Change, and Xenophobia, marked by conflict and instability, naturally cause more concern due to their broader implications.

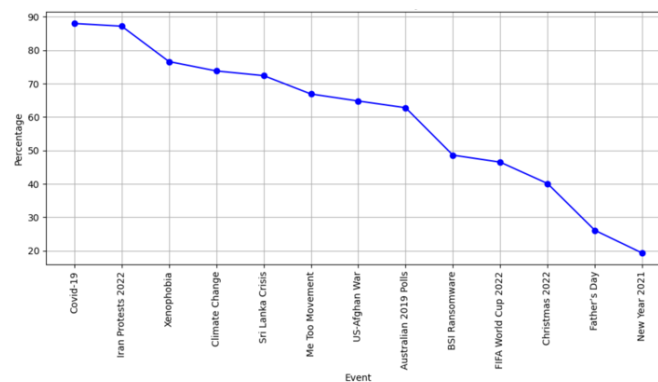


Figure 7. Trend of concern detected tweets across events

Overall, the varying levels of 'Concern Detected' across different events, underscore the complex interplay of factors shaping public sentiment, ranging from immediate threats to long-term societal challenges, and reflect the diverse array of issues that capture the attention and concern of individuals and communities worldwide. We also proved our hypothesis that events characterized by positive circumstances exhibit diminished levels of detected concern within related tweets.

The breakdown of sentiments across specific events, as outlined in the provided Table 6, reflects the corpus's adaptability in capturing emotional nuances within each dataset. There is room for performance improvement, and future refinements to the corpus will enhance its accuracy in identifying and distinguishing various emotional tones. While our approach identified sentiments using the VADER tool and a custom corpus, it did not involve machine learning algorithms like those employed by [18]. Despite this, our method captured nuanced emotional expressions within tweets. The main distinction lies in scalability; machine learning models can offer more flexibility and adaptability for much larger datasets. However, for the specific needs of our study, our approach proved sufficient. Moreover, in our work, despite using only VADER without combining with other machine learning algorithms, in comparison with existing works, such as [4], [15], [19], we achieved significant results in par with their results as displayed in table 7. This finding underscores the potential of our research in contributing to the field of sentiment analysis, particularly in the challenging context of code-switching.

Table 7. Model Performance

Model	Accuracy Metric
Our Solution	76.6% - 88.0% (Negative Circumstances)
BiLSTM RNN[4]	78.0% (with pre-processing)
Subword-LSTM[15]	69.7%
SVM[19]	40%-60%

For instance, the sentiment analysis of major events using VADER, combined with our custom corpus, yielded accurate results, as demonstrated by the high percentages of tweets with detected concerns: 87.2% for Iran Protests 2022, 88.0% for Covid-19, 76.6% for Xenophobia, and 76.6% for Xenophobia. Even for less dominant topics, such as the Sri Lanka Crisis (72.4%) and 66.9% for the Me Too Movement, VADER effectively identified sentiments. These figures illustrate that even without additional machine learning algorithms, VADER, when paired with an emotion-based custom corpus, performed exceptionally well in capturing the emotional tone of tweets, highlighting the robustness and reliability of our chosen method. VADER, a lexicon-based sentiment analysis tool, has been adapted for use in Malay and Indonesian languages, showing promising results in detecting emotional cues in mixed-language environments [20].

4.2. Discussion

Panic and worry are the parts and parcels of our life. People try to voice and gesture them in many ways and expressions. In this study, we focus on the way students convey these emotions. Both the formal and informal conversations have an impact on these two emotions, but informal languages and conversations are indebted to emotions more than formal ones when used on social media. Thus, an inspection of textual context on the platform X aids more interest in examining the two emotions.

A total of 216 synonymous words were used to filter 91,497 students' tweets and out of this process, 134 synonyms of panic and worry were matched within 2,204 of those tweets. Figure 8 exhibits that the word **problem** is the most used word by students to express their anxiety. The word '*takut*' is added up to the second number in the list, **fear** third and so on. The list is figured also with words which seldom occurred. For instance, shaken and apprehension. The lesser the frequency the lesser would be its strength and usage by a student to use the word in their tweets to convey their emotion.



Figure 8. WordCloud based on frequency of words

CORPUS4PANWO: The corpus serves as a valuable resource for delving deeper into the comprehension and experiences of anxiety among individuals in various contexts and stages of life. Researchers can explore nuanced aspects such as situational influences, individual variations, and developmental perspectives.

Our study delves into three key areas of analysis to understand the expression of panic and worry in the corpus. First, we conduct a cross-cultural analysis to explore how these emotions differ across diverse cultural contexts represented in the data. We seek to identify both commonalities and distinctive patterns in the way panic and worry are expressed by individuals from different cultural backgrounds.

Secondly, we examine demographic variances to determine if there are discernible differences in the articulation of anxiety based on factors such as age, gender, or educational background. This aspect of the study aims to reveal how demographic characteristics influence emotional expression.

Lastly, we consider contextual factors, investigating how specific environments, such as academic settings or online social interactions, shape the manifestation of panic and worry in the corpus. By analysing these three dimensions, we aim to gain a comprehensive understanding of the factors influencing the expression of these emotions in various contexts.

4.3. Limitations

While this study provides valuable insights into anxiety-related linguistic expressions in the English, Malay, and Indonesian languages, it is not without limitations. The corpus may not encompass the full spectrum of linguistic diversity within both languages. Additionally, the study focuses on X discourse, which may not fully capture the entirety of anxiety-related expressions in social-media related communities. Furthermore, the analysis is conducted within the confines of the available dataset and machine learning models, which may introduce certain biases. Future research could explore these factors, considering linguistic nuances, data quality, and other potential contributors to the observed variations in classification accuracy.

While our study provides valuable insights into anxiety-related expressions across English, Malay, and Indonesian languages, it is important to acknowledge potential biases inherent in the platform's demographic and usage patterns. The platform, focusing on X discourse, may skew towards specific demographic groups or usage behaviours, which could influence the generalizability of our findings. These biases may limit the representation of broader societal sentiments related to panic and worry. Future research should explore how these demographic and usage patterns impact emotion expression on different social media platforms to enhance the applicability of our conclusions across diverse user populations.

In the context of our research, it is crucial to acknowledge potential ambiguities arising from words with shared spellings across different languages. This linguistic complexity introduces variations in the semantic meanings associated with identical terms. For instance, consider the word '*jam*'. While in English, it conveys meanings aligned with panic or worry, in Malay and Indonesian languages, '*jam*' denotes a clock or the act of indicating time, as exemplified by the phrase '*Jam 10 pagi di Malaysia*' (10 AM in Malaysia).

This linguistic divergence underscores the need for consideration when implementing keyword-based methodologies for sentiment analysis or emotion detection. Failure to account for language-specific interpretations may lead to mislabelling and inaccuracies in the identification of emotional expressions. To mitigate such challenges, it is

recommended to preclude keywords that exhibit multilingual ambiguity to enhance the precision of emotion classification methodologies.

Moreover, employing context-aware word disambiguation techniques such as Part-of-Speech Tagging (POS) and Named Entity Recognition (NER) can enhance accuracy. These techniques help clarify word meanings by identifying grammatical roles and detecting proper nouns, respectively. Additionally, leveraging advanced natural language processing (NLP) methods like contextual embeddings, which capture word meanings within broader sentence contexts, further improves the precision of emotion classification methodologies.

Although this study does not explicitly address ethical considerations and privacy concerns, it is crucial to acknowledge the potential ethical implications of emotion analysis in social media data. Analysing individuals' emotional states without explicit consent raises ethical questions and privacy issues, underscoring the need for responsible data handling practices.

While our study contributes valuable sentiment, corpus tailored for panic and worry detection, it is crucial to acknowledge the dynamic nature of language and sentiments. Words that are currently associated with the detection of early signs of anxiety may undergo semantic shifts over time. The evolving nature of societal perceptions and thinking patterns introduces a level of unpredictability to sentiment analysis.

Even though our study has successfully applied sentiment analysis techniques to multilingual datasets, it is important to acknowledge the inherent challenges and limitations in this approach. Conducting sentiment analysis across multiple languages introduces complexities such as translation accuracy, cultural nuances, and variations in linguistic expressions. These factors can impact the accuracy and reliability of emotional analysis results.

Additionally, the study addresses the challenge of language-dependent meanings in emotional analysis. Words may carry different emotional connotations across languages, affecting the interpretation of sentiment. To mitigate these challenges, strategies will be discussed, including refined keyword selection and validation techniques for sentiment lexicons tailored to each language within our corpus.

As language adapts to cultural shifts, some words within our corpus may lose their association with panic or worry sentiments, reflecting changes in societal norms and attitudes. Conversely, new words may emerge, capturing sentiments of panic and worry that were not previously recognized. This inherent fluidity poses a challenge to maintaining words or terms that remain indefinitely aligned with the ever-changing landscape of emotions.

It is essential for researchers and practitioners to recognize the temporal limitations of sentiment corpora and stay attuned to emerging linguistic nuances. Continuous updates and re-evaluations of corpora may be necessary to ensure their relevance in accurately identifying panic and worry sentiments within the evolving language of online discourse.

Moving forward, future research should aim to further refine these strategies and explore advancements in multilingual sentiment analysis methodologies. By addressing these limitations head-on, we can enhance the robustness and applicability of emotional analysis in diverse linguistic contexts.

5. Conclusion

This study significantly advances our comprehension of anxiety-related linguistic expressions in the code-switching of English with Malay or Indonesian languages, particularly within the context of students' experiences and their discourse. By constructing a multilingual corpus and conducting linguistic analyses, we have gained insights into the nuanced ways in which anxiety is expressed within linguistic communities. The findings underscore the importance of considering linguistic specificities in cross-cultural emotional studies, highlighting the uniqueness of emotional expression within diverse linguistic contexts.

The corpus developed in this study stand as valuable resources for scholars in the field of Southeast Asian linguistics, providing a foundation for further exploration and understanding of emotional expression in diverse linguistic and cultural contexts. As this research lays the groundwork for future studies, we anticipate that ongoing research will continue to enrich our comprehension of emotional expression in a variety of linguistic and cultural contexts.

6. Declarations

6.1. Author Contributions

Conceptualization: R.A.R., S.H.A.H., and F.F.; Methodology: S.H.A.H.; Software: R.A.R.; Validation: R.A.R., S.H.A.H., and F.F.; Formal Analysis: R.A.R., S.H.A.H., and F.F.; Investigation: R.A.R.; Resources: S.H.A.H.; Data Curation: S.H.A.H.; Writing Original Draft Preparation: R.A.R.; Writing Review and Editing: S.H.A.H., R.A.R., and F.F.; Visualization: R.A.R.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the main author.

6.3. Funding

This work was supported in part by the University of Malaya Research under Grant ST014-2022 for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] K. Bhuvanagirir and S. K. Kopparapu, "Mixed language speech recognition without explicit identification of language," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 92-97, 2012.
- [2] C. Nilep, "Code-Mixing and Code-Switching," *In The International Encyclopedia of Linguistic Anthropology*, J. Stanlaw (Ed.), Wiley Online Library, vol. 2020, no. Nov., pp. 1-11, 2020.
- [3] H. Harisal, N. P. Somawati, W. Dyah, and K. Kanah, "Code-Mixing in Student Interaction of Japan UKM Members in State Polytechnic of Bali," *IZUMI*, vol. 10, no. 2, pp. 267-277, 2021.
- [4] A. I. Wibowo, Z. Ramdhani, and R. Rahayuningsih, "Code Mixing Usage in Imperfect: Karier, Cinta & Timbangan Movie Directed by Ernest Prakarsa," *Journal of Pragmatics Research*, vol. 4, no. 1, pp. 60-72, 2022.
- [5] R. Susanti, H. Haryanto, I. Pranawukir, M. Safar, and I. Tjahyadi, "The use of code-mixing and code-switching: Challenge identification in language online mass media," *IJOTL-TL: Indonesian Journal of Language Teaching and Linguistics*, vol. 9, no. 1, pp. 32-43, 2024..
- [6] A. Balahur and M. Turchi, "Multilingual sentiment analysis using machine translation?" *in Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, vol. 3, no. July, pp. 52-60, 2012.
- [7] W. Becker, J. Wehrmann, H. E. L. Cagnini, and R. C. Barros, "An efficient deep neural architecture for multilingual sentiment analysis in twitter," *Proceedings of the 30th FLAIRS*, vol. 30, no. May, pp. 246-251, May 2017.
- [8] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. A. Hawalah, A. Gelbukh & Q. Zhou, "Multilingual sentiment analysis: state of the art and independent comparison of techniques," *Cognitive computation*, vol. 8, no. 3, pp. 757-771, June 2016.
- [9] F. M. P. Del Arco, C. Strapparava, L. A. U. Lopez, and M. T. Martín-Valdivia, "EmoEvent: A multilingual emotion corpus based on different events," *in Proceedings of the Twelfth Language Resources and Evaluation Conference*, vol. 2020, no. May, pp. 1492-1498, 2020.
- [10] J. Deriu, A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hofmann, and M. Jaggi, "Leveraging large amounts of weakly supervised data for multi-language sentiment classification", *In Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of

Geneva, CHE., pp. 1045-1052, April. 2017.

- [11] N. Choudhary, R. Singh, I. Bindlish, and M. Shrivastava, "Sentiment analysis of code-mixed languages leveraging resource rich languages," in *Computational Linguistics and Intelligent Text Processing, 19th International Conference, CICLing 2018*, A. Gelbukh, Ed., Lecture Notes in Computer Science, vol. 13397, Cham: Springer, pp. 104-114, Mar. 2018.
- [12] A. Joshi, A. Prabhu, M. Shrivastava, and V. Varma, "Towards sub-word level compositions for sentiment analysis of hindi-english code-mixed text," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, vol. 26, no. Dec., pp. 2482-2491, 2016.
- [13] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "Sentiment analysis on monolingual, multilingual and code-switching twitter corpora," in *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment, and social media analysis*, vol. 6, no. Sept., pp. 2-8, 2015.
- [14] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," *IEEE Access*, vol. 11, no. Nov., pp. 1-1, 2022.
- [15] A. Hariharan, V. Dorner, C. Weinhardt, and G. W. Alpers., "Detecting panic potential in social media tweets," in *Proceedings of the 25th European Conference on Information Systems (ECIS)*, Guimarães, Portugal, vol. 2016, no. June, Jun. 5-10, 2017, pp. 3181-3190, Research-in-Progress Papers. ISBN: 978-0-9915567-0-0.
- [16] Dewan Bahasa dan Pustaka, "Kamus Dewan," PRPM, 14-Dec-2023. [Online]. Available: <https://prpm.dbp.gov.my/>. [Accessed: 14-Jul-2023].
- [17] Dictionary.com. "Dictionary.com." <https://www.dictionary.com/> (accessed November, 20, 2023).
- [18] Kemendikbud. "Kamus Besar Indonesia(KBBI)." <https://kbbi.kemdikbud.go.id/Beranda> (accessed December 14, 2023).
- [19] Kemendikbud. "Tesaurus Tematis Bahasa Indonesia." <https://tesaurus.kemdikbud.go.id/tematis/> (accessed December, 14, 2023).
- [20] K.F. Yusob and M. Z. Zakaria, "Penggunaan bahasa slanga dalam filem tempatan: satu kajian terhadap filem cereka aksi," in *KONAKA Konferensi Akademik 2016 Perkongsian Ilmu Dari Perspektif Islam*, UiTM Pahang, Nov.30, vol. 2016, no. Nov., pp. 451-456, 2016.
- [21] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37-46, 1960.
- [22] C. Hutto and E. Gilbert, "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, no. 1, pp. 216-225, 2014.
- [23] D. D. Prasad, P. Guttula, T. S. R. Manasa, V. B. Sri, T. Prashanth, and P. N. Sai, "Emotion Analysis of Tweets," in *2023 International Conference on Computer Communication and Informatics (ICCCI), January 2023: IEEE*, vol. 2023, no. Jan., pp. 1-6, 2023.
- [24] P. Dhanalakshmi, G. A. Kumar, B. S. Satwik, K. Sreeranga, A. T. Sai, and G. Jashwanth, "Sentiment Analysis Using VADER and Logistic Regression Techniques," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, vol. 2023, no. Feb., pp. 139-144, doi: 10.1109/ICISCoIS56541.2023.10100565.