M-COVIDLex: The Construction of a Domain-Specific Mixed Code Sentiment Lexicon

Siti Noor Allia Noor Ariffin, Sabrina Tiun, Nazlia Omar
Center for Artificial Intelligence Technology-Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

***Abstract*—Sentiment lexicons serve as essential components in lexicon-based sentiment analysis models. Research on sentiment analysis based on the Malay lexicon indicates that most existing sentiment lexicons for this language are developed from official text corpora, general domain social media text corpora, or domain-specific social media text corpora. Nonetheless, none of the current sentiment lexicons adequately complement the corpus utilized in this study. The rationale is that words in established sentiment lexicons may convey different sentiments compared to those in this paper's corpus, as the strength and sentiment of words are context-dependent, influenced by varying terminology or jargon across domains, and words may not share the same sentiment across multiple domains. This paper proposes the construction of a domain-specific mixed-code sentiment lexicon, termed M-COVIDLex, through the integration of corpus-based and dictionary-based techniques, utilizing seven Malay part-of-speech tags, and enhancing Malay part-of-speech tagging for social media text by introducing a new tag: FOR-POS. The constructed M-COVIDLex is evaluated using two distinct domains of social media text corpus: the specific domain and the general domain. The performance indicates that M-COVIDLex is more appropriate as a sentiment lexicon for analyzing sentiment in a domain-specific social media text corpus, providing valuable insights to governments in assessing the sentiment level regarding the analyzed topic.**

Keywords*—Malay social media text; mixed-code sentiment lexicon; sentiment analysis; domain-specific; lexicon-based; informal Malay; Malay part-of-speech; public health emergencies; COVID-19 Malaysia

I. Introduction

Sentiment analysis is a critical domain within natural language processing [1-4], as sentiment constitutes a significant aspect of human communication [5,6], often articulated through ambiguous and creative language [1]. For example, text that incorporates slang is comprehensible only to particular groups [7], presenting significant challenges for analysis [1]. Analyzing sentiment is essential for understanding individuals' beliefs regarding various issues and their decision-making processes [8-11].

Rising demand for a model that can analyze the sentiment of mixed code (also known as multilingual or cross-language language) text particularly text from social media arisen recently [12,13]. It occurred on account of (i) limited sentiment resources needed for sentiment analysis for languages other than English [14-18], such as sentiment lexicon, language tools, corpora, sentiment analysis algorithm, and sentiment classification algorithm, and (ii) the escalation of social media posts exercising mixed code in multilingual low-resource societies [19-23]. Furthermore, mixed code sentiment analysis eases a better understanding of the sentiment enunciated in various languages [12]. Up to the present time, Malay is still considered as a low-resource language, specifically in this field [5,24] since it is less studied and resource-scarce [25,26].

Sentiment lexicons are a crucial tool in lexicon-based sentiment analysis model [27-31], for any language [32]. Lexicon-based approaches demand human involvement to build a dictionary that has positive and negative lexicons [33]. This approach is divided into dictionary-based and corpus-based [20]. The former relies on sentiment words existing in digital linguistic dictionaries such as WordNet [34] and the latter relies on sentiment words that exist in the study corpus [35]. The success of lexicon-based sentiment analysis depends entirely on a sentiment lexicon that can be constructed manually or semi-automatically [36]. Furthermore, sentiment lexicons, which are also known as lexical dictionaries, are linguistic resources that have lexicons of opinions [37,38] that are labelled as positive or negative according to their semantic orientation [37-40]. For instance, words such as *baik* (good), *bagus* (excellent), and *hebat* (great) are often labelled as words with positive sentiment, while words such as *kejam* (cruel), *jahat* (evil), dan *hodoh* (ugly) are usually labelled as word with negative sentiment. The main purpose of sentiment lexicons is to assign each word in a text a corresponding sentimental weight [41]. Sentiment lexicons are divided into two types: general domain (or all-purpose) sentiment lexicon and domain-specific sentiment lexicon. A general domain sentiment lexicon is a list of words that carry either positive or negative meanings in casual conversation [42], while a domain-specific sentiment lexicon is a list of words that carry either positive or negative meanings in a discussion about a specific domain [43,44].

Recent literature review discloses that most existing sentiment lexicons developed for the field of Malay sentiment analysis are produced using either official text corpora [45], general domain social media text corpora [46], or domain-specific social media text corpora, such as affordable housing projects [47], crisis management [48], and telecommunications [49-51]. It is possible that words in the existing sentiment lexicon do not have any sentiment or carry different sentiments than words in this paper's corpus [52,53] considering the strength and sentiment of words depend on the context of their use and the terminology or jargon differs between domains [36]. Furthermore, it is impossible for a word to have a single score or sentiment in several domains [53-58]. For example, in Malay, the term *kacau* in the food domain carries a positive sentiment

with a score of +1, if the context of its use is to stir or mix something, such as food and the term can also carry a negative sentiment with a score of -1 in the public health emergency domain, if the context of its use is to cause chaos or anxiety. Therefore, the sentiment lexicon built from this paper's corpus is more practical since the words in the sentiment lexicon reflect the sentiment in the context of the domain being studied [55].

Moreover, sentiment lexicons can be generated using two seed words expansion methods: manual and automatic based on words in a dictionary or corpus [56]. Recent literature review reveals that these seeds words can be produced by employing part-of-speech (POS) tags extraction [59], such as adjectives [60-62], verbs [50,51], adverbs [46], and nouns [63]. POS tagging is a feature in the sentiment analysis model that groups words based on their category or role in a sentence [64]. Originally in Malay, words can be classified using four POS tags: nouns, verbs, adjectives, and task words [65]. These POS tags are known to be more suitable for tagging standard Malay text than Malay social media text [64] due to the difference in the writing structure in both texts. However, the latest enhancements on Malay POS tags by [64] have made tagging social media text much effortless, since they created new POS tags especially for mixed code word. For example, FOR-NEG tag for tagging foreign words with negative meaning and NEG tag for tagging Malay words with negative meaning. Though the newly developed Malay POS tags by [64] can tag negative meanings, they lack a tag for tagging foreign words with positive meanings. Therefore, this paper will further improve the Malay POS tags by adding a new POS tag, namely FOR-POS, for tagging foreign words with positive meanings. This enhancement is made to aid in speeding sentiment analysis process by instantly distinguishing words that carry sentiment in the corpus.

This paper presented a new polyglot sentiment lexicon that specific to the latest public health emergency issue in Malaysia which is Coronavirus 2019 (COVID-19). This sentiment lexicon is known as M-COVIDLex (Malay Coronavirus Lexicon) and consist of lexicon of two main languages in Malaysia: Malay and English. The construction of M-COVIDLex utilized a combination of corpus-based and dictionary-based techniques and seven Malay POS tags: adjectives (KA), verbs (KK), adverbs (KAD), nouns (KN), FOR-NEG, FOR-POS, and NEG. The contributions of this paper are listed as follows:

- A new domain-specific social media text corpus focusing on the topic of "the impact of the implementation of government efforts to address public health emergencies in the daily routines of Malaysians". Until the data's copyright is enforced, this corpus will be inaccessible to the public or future research.
- Enriching existing Malay Normalizer by [66] through adding four new rule elements to its existent database.
- Normalizing the generated corpus using an improved Malay Normalizer.
- Enhancing existing Malay POS tags by [64] through adding one new POS tag for tagging foreign words with positive meanings: FOR-POS.
- Tagging all words in the generated corpus with the recently enhanced Malay POS tags.
- Annotated each post in the generated corpus with its proper sentiment polarity either positive, negative, or neutral.
- M-COVIDLex: a domain-specific mixed code sentiment lexicon where each lexicon has been classified as either positive or negative sentiment word.
- The performance evaluation proves that the sentiment lexicon built from this paper's corpus is more practical in analyzing sentiment from the same domain corpus than the general domain corpus.

This paper is structured as follows. The following section (Section II) reviews the related work. In Section III, the proposed method is introduced in detail. Section IV presents experiment and obtains results. Section V discussed the experiments and obtained results and finally, Section VI concludes how this paper can be expanded to further contribute to the Malay social media text sentiment analysis fields.

II. Related Works

A. Types of Sentiment Lexicon

Sentiment lexicons are divided into two types: general domain and domain specific. General domain sentiment lexicon is a list of words that carry either positive or negative meanings in casual conversation [42], while domain-specific sentiment lexicon is a list of words that carry either positive or negative meanings in conversation about a specific domain [43,44].

A general domain sentiment lexicon is suitable for development and implementation in a model that analyzes sentiment text that is not based on any domain for the list of words in the general domain sentiment lexicon is limited to words commonly used in daily conversations and its polarity score is not sensitive to any domain [67-70]. In addition, the study by [36], [47], [71], and [72] stated that domain-specific sentiment analysis models that implement the studied domain-specific sentiment lexicon have the potential to provide better sentiment analysis results and more accurate classification results compared to the general domain sentiment lexicon. The reason for this is that the sentiment polarity scores of words in the sentiment lexicon are obtained based on their meanings in the domain studied [73,74] and sentiment analysis research is sensitive to the domain analyzed. Opinion expressions that carry sentiments, whether positive or negative, differ between domains since each domain has its own language, terminology, or jargon [53-58]. Furthermore, it is impossible for an opinion expression to have the same sentiment polarity score in all domains because words and their sentiment polarity scores are closely related depending on the domain context in which they are used [36].

In conclusion, models developed to analyze sentiment in a specific domain are encouraged to use a sentiment lexicon specific to that domain rather than a general domain sentiment lexicon. The reason for this is that the use of a correct and proper sentiment lexicon can produce good analytical accuracy.

B. Techniques for Constructing Sentiment Lexicon

This paper supported by The Ministry of Higher Education Malaysia (https://www.mohe.gov.my/en) under grant code: FRGS/1/2020/ICT02/UKM/02/1.

Sentiment lexicons can be generated using two seed word expansion methods: manual and automatic based on words in a dictionary or corpus [56]. However, generating a comprehensive sentiment lexicon using manual expansion techniques is complicated, time-consuming, and prone to various errors [56]. Therefore, researchers prefer to use existing sentiment lexicons, such as General Inquirer [75], WordNet [34], Opinion Lexicon [76], Subjectivity Lexicon [77], SentiWordNet [78], SentiStrength [79], SenticNet [80], AFINN [81], Semantic Orientation CALculator (SO-CAL) [63], National Research Council Canada (NRC) Emotion Lexicon [82], Valence Aware Dictionary and sEntiment Reasoner (VADER) [83], LIWC [84], TextBlob [85], and Bing [86].

Dictionary-based expansion techniques require researchers to manually collect initial list of seed words and their polarities before expanding them by extracting synonyms and antonyms of each seed word from existing dictionaries [20]. Corpus-based expansion techniques also use the initial list of seed words to identify words that carry sentiment values and their polarities in the research corpus [56]. Sentiment lexicon generation using corpus-based expansion techniques is more suitable for domain-specific sentiment analysis research than dictionary-based and manual expansion techniques [9]. This is due to the technique is extremely sensitive to the domain, capable of generating sentiment lexicons specific to the domain [40], and capable of handling informal texts well, for instance social media texts [9]. However, sentiment lexicon construction using this corpus-based expansion technique is a lengthy process [87], inefficient in labelling texts with formal terms [9] and has limitations in distinguishing all opinion words compared to dictionary-based expansion techniques [9]. The reason for this is that this technique requires a large corpus to include all opinion words in the study language and large-scale corpus collection is strenuous to be prepared [9]. However, regardless of the sentiment lexicon production technique chosen by the researcher to develop or use, the overall quality of this sentiment lexicon is hard to measure [9]. A summary of sentiment lexicon construction techniques implemented by past research that analyzed Malay sentiment can be seen in Table I.

Based on Table I, the corpus-based sentiment lexicon construction technique can be implemented using the feature extraction method, where this method is done by extracting words that belong to certain POS tag such as adjectives, verbs, and adverbs. Additionally, this paper discovered that there is other POS tag that can be extracted together as they also have sentiment values, namely nouns [63]. The dictionary-based sentiment lexicon construction technique can be performed via the translation method, where the lexicon in the English dictionary is interpreted into Malay. Meanwhile, the manual sentiment lexicon production technique is executed by compounding both techniques, namely corpus-based and dictionary-based.

In conclusion, sentiment lexicons can be developed through various techniques such as corpus-based, dictionary-based, or manually. Choosing the right technique is the key to certify that the sentiment lexicon produced is of excellent quality and capable of providing good sentiment analysis.

TABLE I. SUMMARY OF SENTIMENT LEXICON CONSTRUCTION TECHNIQUES BY PREVIOUS STUDY ON MALAY SENTIMENT ANALYSIS

Techniques	Methods	Method Description	References
Corpus-based	Feature extraction	*POS tag*: adjective	[46,50,51,61]
		POS tag: verb	[46.50,51,61]
		POS tag: adverbs	[46]
		POS tag: negation	[46]
Dictionary-based	Translation	AFINN to Malay	[88]
Manual	Combination of corpus-based & dictionary-based techniques	*Corpus*: Malay Sabah lexicons *Dictionary*: multilingual lexicons	[89]
		Corpus: emoticon, neologism *Dictionary*: Malay lexicons (MySentiDic), English lexicons (MySentiDic translation)	[45]
		Corpus: feature extraction (*POS tag*: adjective) *Dictionary*: WordNet Bahasa & WordNet translation to Malay	[60]
		Corpus: feature extraction (*POS tag*: adjective) *Dictionary*: lexicons by Alexander & Omar (2017)	[62]

III. METHODOLOGY

As previously mentioned, this paper aims to construct domain-specific mixed code sentiment lexicons otherwise known as M-COVIDLex through a combination of corpus-based and dictionary-based techniques along with seven Malay POS tags. The proposed M-COVIDLex construction method entails five key phases: (i) data gathering, (ii) data preprocessing, (iii) construction of M-COVIDLex, (iv) sentiment analysis, and (v) sentiment classification. Each phase is enlightened in further detail below and it is important to emphasize that the data gathering and analysis presented in this paper adhere to the terms and conditions of social media platform, X.

A. Data Gathering

This paper gathered data manually from the social media platform, X (formerly Twitter), where the data must be composed of the combination of two languages, Malay and English, concerning "the impact of the implementation of government efforts to address public health emergencies in the daily routines of Malaysians". This sort of data is recognized as mixed code or code-switching or multilingual. To achieve this purpose, this paper employed keyword-driven data-gathering techniques [64,66,90]. The search was performed on X's advanced search functions [64,66,89] using a predefined list of thirty-three keywords of four affected sectors during COVID-19 in Malaysia: education, safety, health, and economy (see Table II). The keywords were obtained from data issued by [91] and [92]. Nevertheless, the keywords used are restricted to the Malay

language apart from acronyms such as SOP (Standard Operational Procedure) as well as the English loanword like "moratorium" and "internet". X's advanced search function permits users to refine search results based on several criteria including keywords, publication date, language type, and account type, which authorizes researchers to oversee the kind of posts suitable to be extracted. For a post to be included in the search, it must have at least one keyword and was posted between March 2020 and September 2021. The sectors and keywords were selected in such a way to boost the number of posts about the selected topic and limit any extraneous posts [90]. As a result, this paper achieved in gathering 16,898 related posts which were saved in textfiles (txt). The data gathering stage is deemed complete and adequate when all posts resulting from the keyword search have been extracted.

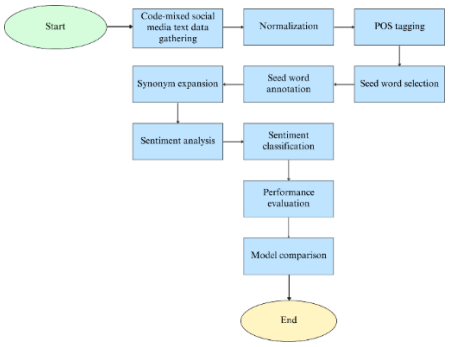

Fig. 1. M-COVIDLex flowchart.

TABLE II. DATA GATHERING KEYWORDS USED IN THIS PAPER

Keywords	Sectors	Overall Keywords	Total Post
harga barang	Economy	11	4,838
tiket pengangkutan			
bantuan kerajaan			
pemulihan ekonomi			
bayaran pinjaman			
inisiatif kerajaan			
pelancongan			
BDR			
moratorium			
wang simpanan			
golongan rakyat			
vaksin	Health	6	3,773
pencegahan			
kontak			
kesihatan			
kuarantin			
kawalan pergerakan			
kawalan	Safety	5	3,131
kebenaran			
rentas			
SOP			
sekatan			
bayaran belia	Education	11	5,156
bayaran pendidikan			
pengurangan yuran			
bantuan makanan			
bantuan pelajar			
internet			
kediaman pelajar			
pergerakan pelajar			
pelajar institusi			
pelajar sekolah			
PDPR			

B. Data Preprocessing

Data preprocessing plays a fundamental role in cleansing text data, specifically social media text, where typographical errors, slang terms, acronyms, and polyglot (code-mixed, code-switching, and multilingual) lexes are frequent [93]. According to [88], the existence of noise in the Malay text is not being managed at this phase [13] affirmed that various preprocessing procedures are important to alleviate noise, typographical divergences, and morphological intricacy, especially in a low-resources language, such as Malay. In this phase, the gathered data is overseen in two steps: normalization and POS tagging. Normalization is essential to diminish the presence of noise from the data that has the potential to interfere with the results of the sentiment analysis and to make the data more manageable [89]. POS tagging is needed to annotate each word in the data with a suitable POS tag.

1) Normalization: A recent study by [66] proposed a rule-based normalization application exclusively built to refine Malay social media text, where it achieves high accuracy in their analysis (97 percent). Therefore, in this paper, this Malay Normalizer was employed to normalize all lexicons in the gathered data as it incorporates diverse essential preprocessing procedure for Malay social media text. For instance, (i) removing noise, (ii) normalizing Malay, English, and Romanized Arabic words to their standard form, (iii) expanding abbreviations, contractions, and acronyms, and (iv) normalizing slang term, colloquialism, and dialects. Nonetheless, it is expected that the application will be inept at normalizing most of the words that exist in this paper's gathered data seeing that it was built using a smaller corpus size. Hence, it needs to be enriched with words from this paper's corpus. Albeit [94] exclaimed that the size of a corpus does not imply its quality and could have more noise, this enrichment is compulsory to ensure all lexicons in the corpus are normalized to their standard form. The enrichment entails adding four new rule elements to its existent database: (i) normalization of novel words, for example, slang term, dialects, and domain-specific terms, (ii) normalization of emoticons, (iii) exclusion of Malay and English question words, and (iv) elimination of English and selected Malay stop words. This supplementation effectively reduces the number of X posts from 16,898 posts to 16,600 posts, where 298 posts were removed from this paper's corpus due to noise.

2) POS tagging: The normalized gathered data initially annotated using Malay POS tags by [64]. These Malay POS tags was chosen given that (i) it achieved high accuracy in their analysis (95 percent) against Malay social media text, (ii) it has been thoroughly upgraded from the previous Malay POS tags by [95], (iii) the POS tags were tailored to be able to annotate each word in the Malay social media text, for example, any foreign language exist in the corpus will be tag with FOR tag, slang term will be tag with SL tag, and dialect terms will be tag with LD tag, and (iv) standard Malay POS tags absence proper tags to label all words in this paper's corpus since social media text are usually written using mixed code otherwise known as multilingual or code-switching [66]. Although [64] has

improved these Malay POS tags to accommodate words in Malay social media text, this paper discovered that it has yet to set up a proper tag for tagging foreign words with positive sentiments. Therefore, it needs to be improved by adding a new POS tag exclusively for tagging positive sentiment foreign words. This paper ruled out to name the newly created POS tag as FOR-POS (foreign positive). Words for the FOR-POS tag are found and extracted from the word that conveys positive meanings in the FOR-tag word list. This paper performed this improvement to aid in speeding the sentiment analysis phase by instantaneously distinguishing words that carries sentiment in the corpus.

C. Construction of M-COVIDLex

The proposed M-COVIDLex construction method entails three steps: (i) seed word selection, (ii) seed word annotation, and (iii) synonym expansion. In this phase, the proposed method will be enlightened in detail.

1) Seed word selection: A seed word is a lexicon that carries either positive or negative sentiment polarity. Based on the literature review in Section II, this paper discovered that these seed words can be generated using four types of POS tags: KA, KK, KAD, and KN. Therefore, in this step, the lexicon tagged with these four POS tags will be extracted as M-COVIDLex seed words. However, since this paper's corpus is made of mixed code social media text and each lexicon is tagged using the improvised Malay POS tags, this paper decided to add three more POS tags to produce M-COVIDLex seed words. The three additional POS tags are (i) NEG, a POS tag designed specifically for negative particles in the Malay language, (ii) FOR-NEG, a POS tag for foreign language words that carry negative sentiments, and (iii) FOR-POS, a POS tag for foreign language words that carry positive sentiments. The lexicons from these seven POS tags were extracted from the corpus using the POS tagging extraction technique. The technique used specific patterns to extract all related lexicons in the corpus, and its implementation produced a list of seed words for each POS tag, where the list functions as a lexical dictionary and is needed when constructing M-COVIDLex. Algorithm 1 shows how this seed word extraction was performed for the lexicon with the KA POS tag. Fig 2 presents the result of executing Algorithm 1 and Fig 3 summarizes the total number of seed words for each POS tag in the M-COVIDLex.

Algorithm 1: M-COVIDLex Seed Word Selection

Input: tagged_corpus K_G, post *S*, lexicon *L*, part_of_speech *G*
Output: seed_word *M-COVIDLex* = (*L*, *G*)
Start
  for all *S* in K_G do
    for all *L* in *S* do
      find *L* == *G* (KA)
        if *L* == *G* (KA)
          add *L* in *M-COVIDLex*
        end if
    end for
  end for
End

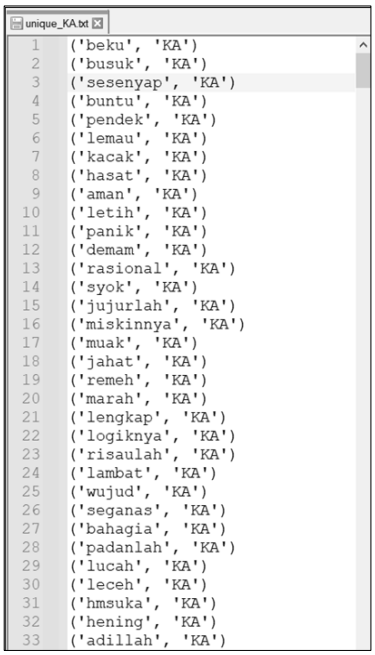

Fig. 2. Seed word selection result for lexicon with the KA POS tag.

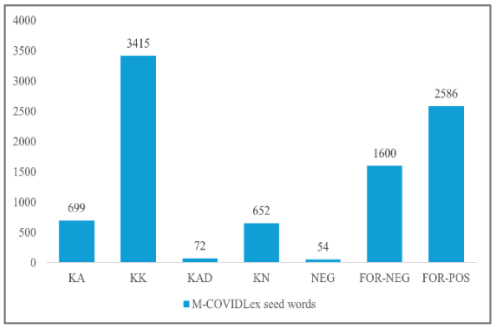

Fig. 3. Summary of M-COVIDLex.

2) Seed word annotation: In this step, seed words will be annotated with either positive or negative sentiment polarity. However, only seed words with KA, KK, KAD, and KN POS tags engage in this annotation process. Seed words with NEG, FOR-NEG, and FOR-POS POS tags are exempted since their sentiment polarity is clear based on the POS tag type. The annotation process in this paper was done by the hired annotators. According to [96], an annotator is an individual appointed to annotate data according to guidelines and time limits given. The following are the guidelines criteria for selecting annotator listed by [96] and followed in this paper:

- Researchers need to figure out the type of language the annotator needs to know or be fluent in before doing the annotation task.
- The researcher needs to set up the specific knowledge that the annotator needs to know about the annotation task.
- Researchers need to make practical considerations about several things, for instance, funds, time, data size, etc.

These guidelines are crucial for researchers to be able to find and appoint annotators who meet the qualifications and criteria [47]. Table III presents the criteria and qualifications of the appointed annotator in this paper along with its references.

TABLE III. THE NEWLY REVISED CRITERIA AND QUALIFICATION OF AN ANNOTATOR USED IN THIS PAPER

Criteria & Qualification of an Annotator	References
Language: Malay native speakers and fluent in English.	[47,96]
Knowledge: Have basic knowledge of the topic and field of study.	[47,96]
Practical Consideration: 1. Current academic graduate (highest): • Malaysian Certificate of Education or • Malaysian Higher Certificate of Education or • Matriculation or • Diploma or • Degree or • Master 2. Aware of the meaning and use of current social media language. 3. Able to give full commitment to annotating data within the specified period.	[96]

The appointed annotator was assigned to manually annotate the lexicons based on the search results on the *Pusat Rujukan Persuratan Melayu*, or PRPM for short [89]. PRPM is an online dictionary specifically for the Malay language developed by DBP. It can be reached at the following link: https://prpm.dbp.gov.my/ . The task of the annotator is to identify and label the sentiment polarity of words based on the definition issued by PRPM and the context of its use. Fig 4 presents some of the seed words for KA POS tag in the M-COVIDLex that have been annotated with their proper sentiment polarity.

3) Synonym expansion: In this step, all M-COVIDLex seed words will be expanded. This expansion aims to expand the coverage of M-COVIDLex seed word variations based on the search results in PRPM for Malay and WordNet [34] for English.

a) Malay lexicon synonyms: The expansion of the Malay language lexicon synonyms was conducted based on the lexicon search results in PRPM [89] and the technique was manual [97]. This paper aims to standardize the expansion of the Malay lexicon by limiting it to level one, given the variability in the number of synonyms across different Malay lexicons. Alternative approaches to this process include expanding the lexicon by accepting all possible synonyms and antonyms.

The first step in the expansion of the Malay lexicon synonyms was to conduct a lexicon search on the PRPM homepage. Fig 5 shows the search conducted on the PRPM homepage for the lexicon *lupa* and Fig 6 shows the search results for the lexicon, where there is word information consisting of its definition and thesaurus. The second step is to extract lexical synonyms up to level one only. Lexical synonyms can be obtained in the thesaurus section. Fig 7 shows the thesaurus for *lupa*, where the lexicon has synonyms up to level three: not remembering (level one), not aware (level two), and not arising in memory (level three). The third step is to add the level one synonym to the M-COVIDLex seed words. Table IV presents examples of synonym expansion for the *lupa*, *layak*, and *usaha* lexicon.

	A	B
1	Leksikon	Polariti
2	adil	Positive
3	adillah	Positive
4	adilnya	Positive
5	agam	Positive
6	aib	Negative
7	ajaib	Negative
8	aktif	Positive
9	alang	Positive
10	alpa	Negative
11	aman	Positive
12	aneh	Negative
13	angkuh	Negative
14	asing	Negative
15	asyik	Positive
16	atasi	Positive

Fig. 4. Annotation results from several seed words in the M-COVIDLex.

Fig. 5. Search for *lupa* on the PRPM homepage.

Fig. 6. Search results for *lupa* on the PRPM website.

TABLE IV. EXAMPLE OF SYNONYMS EXPANSION FOR MALAY LEXICON

M-COVIDLex Seed Words				
Malay Lexicon	***Malay Synonym***	***POS Tag***	***Polarity***	***Polarity Score***
lupa	*tidak ingat*	KK	Negative	-1
layak	*padan*	KA	Positive	+1
usaha	*daya upaya*	KN	Positive	+1

b) English lexicon synonyms: The expansion of the English lexicon synonyms is conducted using a lexical dictionary-based technique, namely WordNet [34]. However, this method only involves lexicons belonging to the FOR-NEG and FOR-POS POS tags. The reason for this is that only these two POS tags have the English lexicon in the M-COVIDLex seed word list. Algorithm 2 below presents how the English synonym expansion method is conducted using the WordNet application.

Algorithm 2: M-COVIDLex English Synonym Expansion

Input: seed_word *M-COVIDLex*, lexicon *L*, wordNet *W*, english_synonym S_i
Output: expansion *M-COVIDLex* = (*L*: *P*);
L: lexicon, P: lexicon_polarity
Start
  for all *L* in *M-COVIDLex* do
    if exist *L* in *W* then
      add S_i in *M-COVIDLex*
    end if
  end for
End

D. Sentiment Analysis

Sentiment analysis is the process of deciding sentiment polarity score, where the technique that will be implemented in this paper is based on (i) grammatical rules of four Malay POS tags: KNF (negation), KP (intensifier), KB (auxiliary), and KH (conjunction), and (ii) word frequency calculations. In this paper, the order of importance of these four Malay POS tags are presumed as follows: KH > K B > KNF > KP (see Algorithm 3). This order is established by recognizing the significance of each POS tag within grammatical rules for determining sentiment polarity scores. For instance, KP POS tag is of the highest importance because of its aptitude to increase or decrease the strength of the polarity of post sentiment, hence found at the end of the order. Fig 8 shows how the sentiment analysis phase happens, and Table V presents the criteria of sentiment polarity used in this paper.

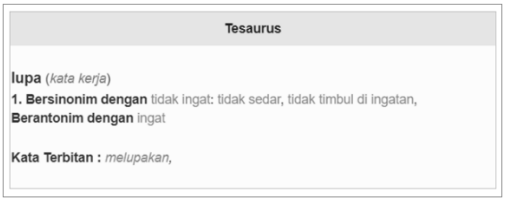

Fig. 7. Lupa thesaurus on the PRPM website.

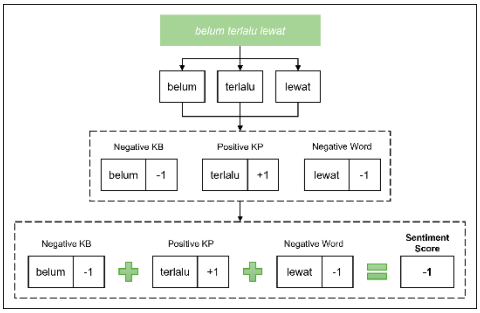

Fig. 8. Demonstration of how M-COVIDLex sentiment analysis phase analyzes the sentiment of a text.

Algorithm 3: M-COVIDLex Sentiment Analysis

Input: Post *S*, Lexicon *L*
Output: Post *S*, Lexicon *L*, Polarity *P*
Start
  if there exists *L* == positive in *S*, then
    score *L* == 1;
  else exists *L* == negative in *S*, then
    score *L* == -1
  end if
  if there exists *L* == conjunction in *S*, then
    classify *S* using conjunction rules;
  else if exists *L* == auxiliary in *S*, then
    classify *S* using auxiliary rules;
  else if exists *L* == negation in *S*, then
    classify *S* using negation rules;
  else if exists *L* == intensifier in *S*, then
    classify *S* using intensifier rules;
  else
    classify *S* using word count
  end if
End

TABLE V. SENTIMENT POLARITY CRITERIA USED IN THIS PAPER

Sentiment Score Criteria	**Sentiment Polarity**
Sentiment score value > 0	Positive
Sentiment score value = 0	Neutral
Sentiment score value < 0	Negative

E. Sentiment Classification

Sentiment classification is the process of classifying posts according to their corresponding polarity, where in this paper, the post is classified into either positive (sentiment polarity score equal to 1), negative (sentiment polarity score equal to -1), or neutral (sentiment polarity score equal to 0) based on the results

gained from sentiment analysis phase. The sentiment classification technique used in this paper is a simple classification otherwise known as lexicon-based classification [89,98]. Algorithm 4 enlightens how this sentiment classification process is executed, and Fig 9 reveals the implementation results of this phase.

Out[1]:

	tweets	KH	KB	<KNF>	KP	SC	LexClass
0	aaron kwok years old withdraw kumpulan wang si...	1	1	1	1	4	1
1	abah bantuan sara hidup bantuan prihatin rakya...	1	1	1	1	4	1
2	abah guru kaunseling sekolah rendah pagi abah ...	1	1	-1	1	2	1
3	abah kau tak keluarkan duit kumpulan wang simp...	1	1	-1	1	2	1
4	abah kawan yang tinggal program perumahan raky...	1	1	1	1	4	1
...	...	...	...	...	...	...	...
16595	zarina anjou right actually haha celebrities w...	1	1	-1	1	2	1
16596	zon hijau kuarantin dan swab test tak ensure f...	1	0	-1	1	1	1
16597	zon kebenaran polis hm	1	1	1	1	4	1
16598	zoom sekat nak belajar online december sudah b...	-1	1	-1	-1	-2	-1
16599	zulkifly bekas menteri kesihatan lepas yang bo...	0	-1	1	-1	-1	-1

16600 rows × 7 columns

Fig. 9. M-COVIDLex sentiment classification results.

Algorithm 4: M-COVIDLex Sentiment Classification

Input: Post *S*, Lexicon *L*, Positive_Word P_W, Negative_Word N_W, Conjunction_Rules R_{KH}, Auxiliary_Rules R_{KB}, Negation_Rules R_{KNF}, Intensifier_Rules R_{KP}

Output: Post *S*, Conjunction_Rules R_{KH}, Auxiliary_Rules R_{KB}, Negation_Rules R_{KNF}, Intensifier_Rules R_{KP}, Lexicon_Frequency K_L, DataFrame *RD*, Classification *SC*, Polarity_Score *P*, Sentiment *LexClass*

Start

count total P_W in *S*
count total N_W in *S*
for every *L* in *S*
  count frequency *L* == R_{KH}
    add frequency *L* in R_{KH} [*RD*]
      print *P* frequency *L* in R_{KH} [*RD*]
  count frequency *L* == R_{KB}
    add frequency *L* in R_{KB} [*RD*]
      print *P* frequency *L* in R_{KB} [*RD*]
  count frequency *L* == R_{KNF}
    add frequency *L* in R_{KNF} [*RD*]
      print *P* frequency *L* in R_{KNF} [*RD*]
  count frequency *L* == R_{KP}
    add frequency *L* in R_{KP} [*RD*]
      print *P* frequency *L* in R_{KP} [*RD*]
end for
for every *P* in [*RD*]
  count *P* in R_{KH} [*RD*] + R_{KB} [*RD*] + R_{KNF} [*RD*] + R_{KP} [*RD*]
    add *P* in *SC*
end for
if *P* in *SC* == 0;
  print *P* in *LexClass* == 0
else if P in *SC* > 0;
  print *P* in *LexClass* == 1
else if *P* in *SC* < 0
  print *P* in *LexClass* == -1
end if

End

IV. Experiment and Results

The evaluation of M-COVIDLex will involve an analysis of sentiments from two distinct social media text corpora: a domain-specific corpus (the one presented in this paper) and a general domain corpus as referenced in [64].

A. M-COVIDLex

The third phase has successfully resulted in the development of M-COVIDLex, a domain-specific mixed code sentiment lexicon. The lexicon in M-COVIDLex is restricted to two sentiment polarities: positive sentiment, assigned a score of +1, and negative sentiment, assigned a score of -1. Neutral sentiment lexicons with a score of 0 are excluded from the seed words of M-COVIDLex, as they lack sentiment value. M-COVIDLex contains 6,698 lexicons associated with positive sentiment and 3,813 lexicons associated with negative sentiment.

B. Performance Evaluation

The performance of M-COVIDLex is initially assessed on a domain-specific social media text corpus, concentrating on the effects of government initiatives aimed at addressing public health emergencies on the daily routines of Malaysians. The experimental dataset corpus was derived from the execution of the second phase. This study utilized a subset of 800 posts from a total of 16,600 normalized posts. The selection of posts was conducted randomly, based on sentiment polarity, resulting in 400 posts with positive sentiment and 400 posts with negative sentiment. The chosen experimental datasets are accompanied by confusion matrix tables to derive their values. Fig 10 illustrates a portion of the data annotation results, while Table VI displays the corresponding values derived from these results.

teks	polariti	penilaian
aaron kwok years old withdraw kumpulan wang simpanan pekerja already	1	True Positive
abah bantuan sara hidup bantuan prihatin rakyat citra ilestari isinar lepas ringgit malaysia ewallet hear shit one time going lose ringgit malaysia bloody ewallets given promote cashless transactions bloody financial aid	1	False Positive
abah guru kaunseling sekolah rendah pagi abah pergi sekolah jam pagi sebentar langit masih bergelap untuk menyambut anak anak muridnya kalau sahajalah murid darjah yang menangis tak mahu sekolah abah pujuk masuk	1	True Positive
abah kau tak keluarkan duit kumpulan wang simpanan pekerja akaun bukan nak dapat okay sekali sudah okay nak harap bantuan prihatin nasional yang dapat biarlah yang perlukan duit mengeluaran duit diwaktu terdesak	1	True Positive
abah kawan yang tinggal program perumahan rakyat sudah golongan bottom forty cakap beli beras yang harganya kampung ringgit malaysia boleh makan minggu tetapi nak bayar ansuran atau sewa rumah kereta dan data internet untuk pengajaran dan pembelajaran rumah	1	False Positive

Fig. 10. M-COVIDLex data annotation results.

TABLE VI. Confusion Matrix Table Values for M-COVIDLex Performance Evaluation

Confusion Matrix	Experimental Dataset
True Positive (TP)	274 posts
False Positive (FP)	126 posts
True Negative (TN)	305 posts
False Negative (FN)	95 posts

The values presented in Table VI serve to evaluate the effectiveness and quality of M-COVIDLex in sentiment analysis

of the experimental dataset. This paper employs the following evaluation metrics to assess the performance of the proposed sentiment lexicon: (i) error rate, (ii) accuracy, (iii) sensitivity, (iv) specificity, (v) precision, and (vi) F1-score. Fig 11 illustrates the performance outcomes of M-COVIDLex, utilising the confusion matrix alongside six evaluation metrics.

Error rate. This evaluation measure was selected to assess the effectiveness of M-RuleScore in analysing the sentiment of the dataset, indicating that a lower error rate value corresponds to improved performance of the proposed M-RuleScore.

$$Error\ rate = \frac{(FP + FN)}{(TP + FP + FN + TN)} \quad (1)$$

Accuracy. This evaluation measure was selected to assess the proportion of posts accurately predicted from the entire dataset, with a higher accuracy value indicating superior performance of the proposed M-RuleScore.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (2)$$

Sensitivity. This evaluation measure was selected to assess the classification error of the dataset, indicating that a higher sensitivity value correlates with improved performance of the proposed M-RuleScore.

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (3)$$

Specificity. This evaluation measure was selected to assess the suitability of M-RuleScore for analysing the dataset, indicating that a higher specificity value reflects improved performance of M-RuleScore in analysing negative sentiments.

$$Specificity = \frac{TN}{(TN + FP)} \quad (4)$$

Precision. This evaluation measure was selected to assess the effectiveness and robustness of the proposed M-RuleScore in analysing the sentiment of the dataset.

$$Precision = \frac{TP}{(TP + FP)} \quad (5)$$

F1-score. This evaluation measure was selected to assess the mean values of sensitivity and accuracy for the proposed M-RuleScore in analysing the dataset's sentiment.

$$F1 - score = \frac{(2 \times precision \times sensitivity)}{(precision + sensitivity)} \quad (6)$$

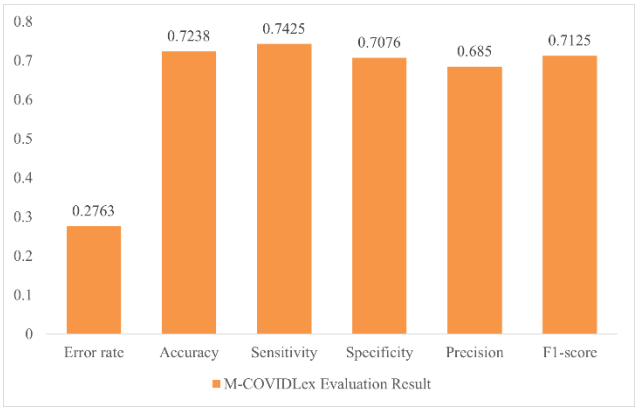

Fig. 11. M-COVIDLex performance evaluation result.

The performance evaluation results of M-COVIDLex presented in Fig 11 yield several conclusions.

- M-COVIDLex achieved an accuracy of 72.38% in the analysis of the experimental dataset. The analysis indicates that, from a total of 800 posts by Malaysians, M-COVIDLex successfully evaluated 274 posts expressing positive sentiments and 305 posts expressing negative sentiments regarding the government's response to the COVID-19 crisis.
- M-COVIDLex recorded an error rate of 27.63% in the analysis of the experimental dataset. Of the 800 posts analysed concerning the government's efforts in addressing the COVID-19 crisis and their impact on the daily lives of Malaysians, 126 posts were incorrectly classified as expressing positive sentiments, while 95 posts were misclassified as expressing negative sentiments.
- M-COVIDLex achieved a precision of 0.6850 in the analysis of the experimental dataset. The precision value indicates that, among the 400 posts predicted to express positive sentiments in the experimental dataset, only 68.50% accurately reflected positive sentiments and aligned with the government's efforts in addressing the COVID-19 crisis.
- M-COVIDLex demonstrated a sensitivity of 0.7425 in the analysis of the experimental dataset. The sensitivity value indicates that, among 369 genuinely positive posts in the experimental dataset, only 74.25% expressed positive sentiments regarding the government's efforts to address the COVID-19 crisis.
- M-COVIDLex demonstrated a specificity of 0.7076 when analysing the experimental dataset. The specificity value indicates that, among the 400 posts predicted to express negative sentiments in the experimental dataset, 70.76% accurately reflect negative sentiments opposing the government's efforts in addressing the COVID-19 crisis.

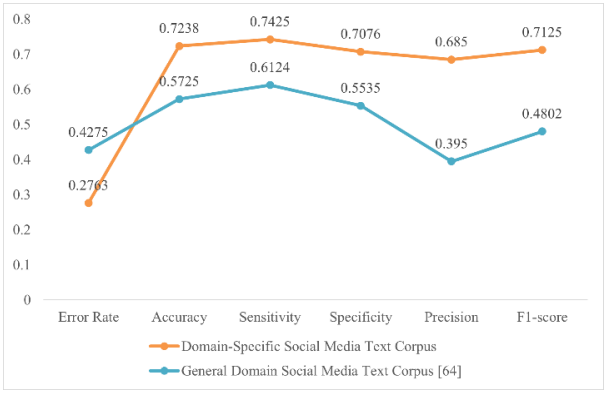

Fig. 12. Comparison of evaluation metrics between domain-specific and general domain social media text corpus.

To ensure comparability in the number of posts used for model comparison with other datasets and for evaluating the performance of M-COVIDLex, the analysis was limited to 800 posts. The performance of M-COVIDLex may be impacted by

this limitation. Evaluating performance across all 16,600 posts is likely to produce more favourable outcomes for M-COVIDLex. M-COVIDLex effectively assessed Malaysian sentiment concerning the government's response to the COVID-19 crisis and its effects on daily life. Despite M-COVIDLex achieving an accuracy of 72.38 percent and an error rate of 27.63 percent, which are lower than many leading models in sentiment analysis for comparable domains, the performance evaluation indicates that a significant majority of Malaysians approve of and support the government's crisis management efforts.

C. Model Comparison

This section evaluates the performance of M-COVIDLex on the general domain social media text corpus as referenced in [64]. The corpus selected by [64] was generated from Malay social media text, focussing exclusively on the contextual use of Malay social media terminology without emphasising any specific domain. The purpose of this assessment is to evaluate the suitability and effectiveness of M-COVIDLex as a sentiment lexicon for analysing sentiment within this domain. This study utilised a dataset comprising 800 posts selected from a total corpus of 1,791 posts, with 400 posts representing positive sentiment and 400 posts representing negative sentiment. The selection of these posts was conducted randomly, determined by the polarity of their sentiment. Table VII presents a comparison of the confusion matrix values for domain-specific social media text versus general domain social media text. Fig 12 presents a comparison of the evaluation metrics for the social media texts analysed.

TABLE VII. COMPARISON OF CONFUSION MATRIX VALUES BETWEEN DOMAIN-SPECIFIC AND GENERAL DOMAIN SOCIAL MEDIA TEXT CORPUS

Performance Evaluation	Domain-Specific Social Media Text Corpus	General Domain Social Media Text Corpus [64]
Corpus size	800 posts	
Positive sentiment posts	400 posts	
Negative sentiment posts	400 posts	
True Positive (TP)	274 posts	158 posts
False Positive (FP)	126 posts	242 posts
True Negative (TN)	305 posts	306 posts
False Negative (FN)	95 posts	94 posts

The comparison of M-COVIDLex performance evaluation results between the domain-specific social media text corpus dataset and the general domain social media text corpus dataset [64] in Fig 12 allows for several conclusions to be drawn.

- M-COVIDLex demonstrated superior error rates and accuracy on the domain-specific social media text corpus compared to the general domain social media text corpus. The observed result aligns with expectations, as M-COVIDLex was developed utilising lexicons that contain sentiment values within a domain-specific social media text corpus.
- The quantity of posts exhibiting positive sentiment in both datasets is identical, totalling 400 posts. Nevertheless, the sensitivity outcomes of M-COVIDLex on the domain-specific social media text corpus dataset surpass those of the general domain social media text corpus dataset. M-COVIDLex accurately predicted 274 out of 400 posts with positive sentiment in the domain-specific social media text corpus, compared to 158 in the general domain corpus. Additionally, the classification error for M-COVIDLex in the domain-specific dataset is marginally lower than that in the general dataset. The sensitivity results indicate that M-COVIDLex effectively identified true positive posts and significantly minimised false negative posts within the domain-specific social media text corpus dataset.
- The quantity of posts exhibiting negative sentiment in both datasets is identical, totalling 400 posts. Nevertheless, the specificity outcomes of M-COVIDLex are superior in the domain-specific social media text corpus compared to the general domain social media text corpus. Although the number of negative posts indicating negative sentiment in the domain-specific social media text corpus dataset is slightly higher than in the general domain dataset, the number of negative posts misclassified as positive in the domain-specific dataset is significantly lower than in the general dataset. The findings demonstrate that M-COVIDLex effectively identifies true negative posts and minimises false positive posts within the domain-specific social media text corpus dataset.
- The precision results of M-COVIDLex are superior on the domain-specific social media text corpus dataset compared to the general domain social media text corpus dataset, exhibiting a difference of 29 percent. The findings indicate that the sentiment analysis model demonstrates improved accuracy in predicting true positive sentiment posts within the domain-specific social media text corpus dataset, thereby minimising the classification error associated with false positive posts. This result was anticipated, as the sentiment analysis is dependent on the lexicon list in M-COVIDLex.
- The F1-score results of M-COVIDLex are superior on the domain-specific social media text corpus dataset compared to the general domain social media text corpus dataset. The results demonstrate that M-COVIDLex effectively balances the analysis of sentiment within the domain-specific social media text corpus dataset.

The sentiment analysis results derived from the general domain social media text corpus dataset are more significant than those from the domain-specific social media text corpus dataset. This result demonstrates that M-COVIDLex, developed from a domain-specific social media text corpus, is exclusively appropriate for sentiment analysis within the same domain corpus. The application of this method on a general domain-specific social media text corpus is not advisable, as the resulting analysis yields significantly lower outcomes and fails to offer any valuable insights. Despite M-COVIDLex demonstrating superior performance on a domain-specific social media text corpus, its application to analyse sentiments in other specific domains, such as banking, entertainment, and food, yields a low probability of accurate sentiment analysis results for those domains. The primary purpose of constructing M-COVIDLex is to address the public health emergency related to COVID-19 in Malaysia. Given that most domains possess distinct sub-languages, the implementation of M-COVIDLex on other public

health emergency domain-specific corpora, such as Influenza, is anticipated to yield superior sentiment analysis performance compared to general domain corpora. The lexicon in M-COVIDLex includes general terms related to public health emergencies, including vaccines, protection, monitoring, and prevention. The enumeration of these general terms demonstrates that M-COVIDLex's contribution extends beyond COVID-19 to encompass other public health emergencies.

V. Discussion

The experiments conducted in the previous section highlight the significance of analysing sentiment within a domain-specific social media text corpus dataset, utilizing a sentiment lexicon derived from the same corpus. The performance evaluation of M-COVIDLex on the social media text corpus dataset indicates that, despite the inclusion of 10,511 sentiment lexicons, the accuracy results remained below 80 percent. The limitations on the expansion of Malay synonyms may stem from the selection of only level one synonym words as seed words. Additionally, given that M-COVIDLex is derived from a domain-specific social media text corpus, it is logical that the accuracy results of this dataset surpass those of a general domain social media text corpus dataset. This analysis confirms the assertion from [55] that sentiment lexicons developed from domain-specific corpora are capable of providing high classification accuracy and thorough insights exclusively for that domain. The efficiency of the proposed method is significantly affected by the quality of the developed lexical dictionaries for all seven Malay POS tags [8]. The process necessitates the engagement of human resources to manually annotate the data [94], a task that is both time-consuming and costly, as indicated by multiple studies [12,29,63,99]. This proposed method may establish a baseline for future research on sentiment analysis of multilingual, code-mixed, or code-switching social media texts. Future initiatives in safety sectors during public health emergencies require the creation of an application that can analyse code-mixed sentiment on social media. A tool of this nature would facilitate prompt notifications to government entities and pertinent organisations concerning individuals who breach movement control orders, rather than relying exclusively on physical enforcement strategies such as roadblocks. The performance evaluation results of the proposed method will assist industrial researchers and relevant agencies in comprehending the Malaysian sentiment regarding government-structured relief aids during the health crisis. The performance result can be enhanced by (i) utilising a larger corpus, ideally the entire dataset of 16,600 X posts, rather than a subset, and (ii) broadening the lexical dictionaries to include a greater variety of words, including antonyms and third-level synonyms.

VI. Conclusion

This paper outlines a method for constructing a domain-specific mixed code sentiment lexicon, referred to as M-COVIDLex, through the integration of corpus-based and dictionary-based techniques, utilising seven Malay POS tags: KA, KK, KAD, KN, FOR-NEG, FOR-POS, and NEG. This mixed code sentiment lexicon addresses the absence of a sentiment lexicon tailored for the public health emergency context, specifically regarding COVID-19 in Malaysia. The M-COVIDLex construction method consists of five fundamental phases: (i) Acquiring the sentiment of Malaysians from social media platform X regarding the impact of government efforts in addressing the COVID-19 crisis on their daily lives; (ii) Processing the acquired data with an enhanced Malay Normaliser and an improved set of 46 Malay POS tags; (iii) Constructing a mixed code sentiment lexicon through seed word selection, annotation, and synonym expansion; (iv) Analysing the sentiments of the acquired data using grammatical rules of four Malay POS tags: KNF, KP, KB, and KH, along with word frequency calculations; and (v) Classifying the sentiments using a lexicon-based classification technique. The evaluation of sentiment classification is conducted through a confusion matrix and six metrics: error rate, accuracy, sensitivity, specificity, precision, and F1-score. The performance evaluation of the proposed M-COVIDLex on the domain-specific social media text corpus dataset exceeds its performance on the general domain social media text corpus dataset.

Acknowledgment

The first author has rendered M-COVIDLex available for public use and future research endeavours. The list is available for download at the following link: https://doi.org/10.6084/m9.figshare.26826250.v1.

References

[1] Rajkumar Buyya, Calheiros, R. N., & Amir Vahid Dastjerdi. (2016). Big data : principles and paradigms. Elsevier/Morgan Kaufmann.

[2] Bakar, M. F. R. A., Idris, N., Shuib, L., & Khamis, N. (2020). Sentiment Analysis of Noisy Malay Text: State of Art, Challenges and Future Work. IEEE Access, 8, 24687-24696.

[3] Zunic, A., Corcoran, P., & Spasic, I. (2020). Sentiment analysis in health and well-being: systematic review. JMIR medical informatics, 8(1), e16023.

[4] Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social network analysis and mining, 11(1), 81.

[5] Enjop, V., Adnan, R., Jamil, N., Ahmad, S., Zainol, Z., & Ahmad, S. A. (2022). Does Google Translate Affect Lexicon-Based Sentiment Analysis of Malay Social Media Text?. Malaysian Journal of Computing, 7(2), 1236-1249.

[6] Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. International Journal of Research in Marketing, 40(1), 75-87.

[7] Cambridge Dictionary. (2019, December 4). SLANG | meaning in the Cambridge English Dictionary. Cambridge.org. https://dictionary.cambridge.org/dictionary/english/slang

[8] Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. Procedia Computer Science, 161, 707-714.

[9] Darwich, M., Mohd, S. A., Omar, N., & Osman, N. A. (2019). Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. J. Digit. Inf. Manag., 17(5), 296.

[10] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134.

[11] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.

[12] Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. Applied Sciences, 13(7), 4550.

[13] Yusuf, A., Sarlan, A., Danyaro, K. U., Rahman, A. S. B., & Abdullahi, M. (2024). Sentiment Analysis in Low-Resource Settings: A

Comprehensive Review of Approaches, Languages, and Data Sources. IEEE Access.

[14] Batanović, V., Cvetanović, M., & Nikolić, B. (2020). A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. PLoS One, 15(11), e0242050.

[15] Meetei, L. S., Singh, T. D., Borgohain, S. K., & Bandyopadhyay, S. (2021). Low resource language specific pre-processing and features for sentiment analysis task. Language Resources and Evaluation, 55(4), 947-969.

[16] Kumari, D., Ekbal, A., Haque, R., Bhattacharyya, P., & Way, A. (2021). Reinforced nmt for sentiment and content preservation in low-resource scenario. Transactions on Asian and Low-Resource Language Information Processing, 20(4), 1-27.

[17] Marreddy, M., Oota, S. R., Vakada, L. S., Chinni, V. C., & Mamidi, R. (2022). Am I a resource-poor language? Data sets, embeddings, models and analysis for four different NLP tasks in telugu language. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(1), 1-34.

[18] Ekbal, A., Bhattacharyya, P., Saha, T., Kumar, A., & Srivastava, S. (2022, June). HindiMD: A multi-domain corpora for low-resource sentiment analysis. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 7061-7070).

[19] Zabha, N. I., Ayop, Z., Anawar, S., Hamid, E., & Abidin, Z. Z. (2019). Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach. International Journal of Advanced Computer Science and Applications, 10(1).

[20] Mahadzir, N. H. (2021). Sentiment Analysis of Code-Mixed Text: A Review. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(3), 2469-2478.

[21] Konate, A., & Du, R. (2018). Sentiment analysis of code-mixed Bambara-French social media text using deep learning techniques. Wuhan University Journal of Natural Sciences, 23(3), 237-243.

[22] Srinivasan, R., & Subalalitha, C. N. (2023). Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distributed and Parallel Databases, 41(1), 37-52.

[23] Hidayatullah, A. F., Apong, R. A., Lai, D. T., & Qazi, A. (2023). Corpus creation and language identification for code-mixed Indonesian-Javanese-English Tweets. PeerJ Computer Science, 9, e1312.

[24] Laumann, F. (2022, June 10). Low-resource language: what does it mean? NeuralSpace. https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5

[25] Nasharuddin, N. A., Abdullah, M. T., Azman, A., & Kadir, R. A. (2017). English and Malay cross-lingual sentiment lexicon acquisition and analysis. In Information Science and Applications 2017: ICISA 2017 8 (pp. 467-475). Springer Singapore.

[26] Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264.

[27] Deng, D., Jing, L., Yu, J., & Sun, S. (2019). Sparse self-attention LSTM for sentiment lexicon construction. IEEE/ACM transactions on audio, speech, and language processing, 27(11), 1777-1790.

[28] Alsolamy, A. A., Siddiqui, M. A., & Khan, I. H. (2019). A corpus based approach to build arabic sentiment lexicon. International Journal of Information Engineering and Electronic Business, 11(6), 16-23.

[29] Machová, K., Mikula, M., Gao, X., & Mach, M. (2020). Lexicon-based sentiment analysis using the particle swarm optimization. Electronics, 9(8), 1317.

[30] Wang, Y., Yin, F., Liu, J., & Tosato, M. (2020). Automatic construction of domain sentiment lexicon for semantic disambiguation. Multimedia Tools and Applications, 79, 22355-22373.

[31] Du, M., Li, X., & Luo, L. (2021). A Training-Optimization-Based Method for Constructing Domain-Specific Sentiment Lexicon. Complexity, 2021(1), 6152494.

[32] Chathuranga, P. D. T., Lorensuhewa, S. A. S., & Kalyani, M. A. L. (2019, September). Sinhala sentiment analysis using corpus based sentiment lexicon. In 2019 19th international conference on advances in ICT for emerging regions (ICTer) (Vol. 250, pp. 1-7). IEEE.

[33] Tho, C., Heryadi, Y., Lukas, L., & Wibowo, A. (2021, April). Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach. In Journal of Physics: Conference Series (Vol. 1869, No. 1, p. 012084). IOP Publishing.

[34] Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

[35] Shayaa, S., Jaafar, N. I., Bahri, S., Sulaiman, A., Wai, P. S., Chung, Y. W., ... & Al-Garadi, M. A. (2018). Sentiment analysis of big data: methods, applications, and open challenges. Ieee Access, 6, 37807-37827.

[36] Labille, K., Gauch, S., & Alfarhood, S. (2017, August). Creating domain-specific sentiment lexicons via text mining. In Proc. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM) (pp. 1-8).

[37] Sazzed, S. (2020, August). Development of sentiment lexicon in bengali utilizing corpus and cross-lingual resources. In 2020 IEEE 21st International conference on information reuse and integration for data science (IRI) (pp. 237-244). IEEE.

[38] Piryani, R., Piryani, B., Singh, V. K., & Pinto, D. (2020). Sentiment analysis in Nepali: exploring machine learning and lexicon-based approaches. Journal of Intelligent & Fuzzy Systems, 39(2), 2201-2212.

[39] Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. Information fusion, 36, 10-25.

[40] Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. Asian Journal of Computer Science and Technology, 8(S2), 1-6.

[41] Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. IEEE access, 8, 23522-23530.

[42] Ofek, N., Caragea, C., Rokach, L., Biyani, P., Mitra, P., Yen, J., ... & Greer, G. (2013, May). Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon. In 2013 international conference on social intelligence and technology (pp. 109-113). IEEE.

[43] Han, H., Zhang, J., Yang, J., Shen, Y., & Zhang, Y. (2018). Generate domain-specific sentiment lexicon for review sentiment analysis. Multimedia Tools and Applications, 77, 21265-21280.

[44] Liu, Y., Jiang, C., & Zhao, H. (2019). Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. Decision Support Systems, 123, 113079.

[45] Chekima, K., Alfred, R., & Chin, K. O. (2017). Rule-based model for Malay text sentiment analysis. In Computational Science and Technology: 4th ICCST 2017, Kuala Lumpur, Malaysia, 29–30 November, 2017 (pp. 172-185). Springer Singapore.

[46] Shamsudin, N. F., Basiron, H., & Sa'aya, Z. (2016). Lexical based sentiment analysis-verb, adverb & negation. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 8(2), 161-166.

[47] Mahadzir, N. H., Omar, M. F., Nawi, M. N. M., Salameh, A. A., Hussin, K. C., & Sohail, A. (2022). Melex: The construction of malay-english sentiment lexicon. Computers, Materials and Continua.

[48] Suhaimi, S. H., Bakar, N. A. A., & Azmi, N. F. M. (2021). Proposing Malay Sarcasm Detection on Social Media Services: A Machine Learning Approach. Open International Journal of Informatics, 9(Special Issue 2), 1-10.

[49] Tan, Y. F., Lam, H. S., Azlan, A., & Soo, W. K. (2016, April). Sentiment Analysis for Telco Popularity on Twitter Big Data Using a Novel Malaysian Dictionary. In ICADIWT (pp. 112-125).

[50] Anbananthen, K. S. M., Selvaraju, S., & Krishnan, J. K. (2017). The generation of malay lexicon. Am. J. Applied Sci, 14, 503-510.

[51] Selvaraju, S., & Anbananthen, K.S. (2019). Opinion Extraction on Online Malay Text. American Journal of Applied Sciences, 16, 134-142.

[52] Li, W., Zhu, L., Guo, K., Shi, Y., & Zheng, Y. (2018). Build a tourism-specific sentiment lexicon via word2vec. Annals of Data Science, 5, 1-7.

[53] Muhammad, S. H., Brazdil, P., & Jorge, A. (2020). Incremental approach for automatic generation of domain-specific sentiment lexicon. In Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42 (pp. 619-623). Springer International Publishing.

[54] Kreutz, T., & Daelemans, W. (2018). Enhancing general sentiment lexicons for domain-specific use. In Proceedings of the 27th International Conference on Computational Linguistic, Santa Fe, New Mexico, USA, August 20-26, 2018 (pp. 1056-1064).

[55] Feng, J., Gong, C., Li, X., & Lau, R. Y. (2018). Automatic approach of sentiment lexicon generation for mobile shopping reviews. Wireless Communications and Mobile Computing, 2018.

[56] Almatarneh, S., & Gamallo, P. (2018). Automatic construction of domain-specific sentiment lexicons for polarity classification. In Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection-15th International Conference, PAAMS 2017 15 (pp. 175-182). Springer International Publishing.

[57] Bergsma, T., van Stegeren, J., & Theune, M. (2020, May). Creating a sentiment lexicon with game-specific words for analyzing NPC dialogue in the elder scrolls V: Skyrim. In Workshop on Games and Natural Language Processing (pp. 1-9).

[58] Shaukat, K., Hameed, I. A., Luo, S., Javed, I., Iqbal, F., Faisal, A., ... & Adeem, G. (2020). Domain Specific Lexicon Generation through Sentiment Analysis. International Journal of Emerging Technologies in Learning (iJET), 15(9), 190-204.

[59] Singh, V., Singh, G., Rastogi, P., & Deswal, D. (2018, December). Sentiment analysis using lexicon based approach. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 13-18). IEEE.

[60] Alexander, N. S., & Omar, N. (2017). Generating a Malay Sentiment Lexicon Based on WordNet. Asia-Pacific Journal of Information Technology and Multimedia, 6(1).

[61] bin Rodzman, S. B., Rashid, M. H., Ismail, N. K., Abd Rahman, N., Aljunid, S. A., & Abd Rahman, H. (2019, April). Experiment with lexicon based techniques on domain-specific Malay document sentiment analysis. In 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE) (pp. 330-334). IEEE.

[62] Sukawai, E. Z. U. A. N. A., & Omar, N. A. Z. L. I. A. (2020). Corpus Development for Malay Sentiment Analysis Using Semi Supervised Approach. Asia-Pacific Journal of Information Technology and Multimedia, 9(01), 94-109.

[63] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

[64] Ariffin, S. N. A. N., & Tiun, S. (2022). Improved POS Tagging Model for Malay Twitter Data based on Machine Learning Algorithm. International Journal of Advanced Computer Science and Applications, 13(7).

[65] Safiah, N., Onn, F. M., Musa, H. H., & Mahmood, A. H. (2010). Tatabahasa Dewan Edisi Ketiga. Kuala Lumpur: Dewan Bahasa dan Pustaka.

[66] Ariffin, S. N. A. N., & Tiun, S. (2020). Rule-based text normalization for Malay social media texts. International Journal of Advanced Computer Science and Applications, 11(10).

[67] Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodriguez-Garcia, M. A., & Valencia-Garcia, R. (2017). Sentiment analysis on tweets about diabetes: An aspect-level approach. Computational and mathematical methods in medicine, 2017(1), 5140631.

[68] Ikoro, V., Sharmina, M., Malik, K., & Batista-Navarro, R. (2018, October). Analyzing sentiments expressed on Twitter by UK energy company consumers. In 2018 Fifth international conference on social networks analysis, management and security (SNAMS) (pp. 95-98). IEEE.

[69] Jiang, K., & Li, Y. (2020, December). Mining customer requirement from online reviews based on multi-aspected sentiment analysis and Kano model. In 2020 16th Dahe Fortune China Forum and Chinese High-educational Management Annual Academic Conference (DFHMC) (pp. 150-156). IEEE.

[70] Yuan, H., Tang, Y., Xu, W., & Lau, R. Y. K. (2021). Exploring the influence of multimodal social media data on stock performance: an empirical perspective and analysis. Internet Research, 31(3), 871-891.

[71] Zhi, S., Li, X., Zhang, J., Fan, X., Du, L., & Li, Z. (2017, August). Aspects opinion mining based on word embedding and dependency parsing. In Proceedings of the International Conference on Advances in Image Processing (pp. 210-215).

[72] Chen, Y., & Ji, W. (2021). Public demand urgency for equitable infrastructure restoration planning. International Journal of Disaster Risk Reduction, 64, 102510.

[73] Aqlan, A. A. Q., Manjula, B., & Naik, R. L. (2019). A Study of Sentiment Analysis: Concepts, Techniques, and Challenges. In Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 147-162). Springer, Singapore.

[74] Sanagar, S., & Gupta, D. (2020). Unsupervised genre-based multidomain sentiment lexicon learning using corpus-generated polarity seed words. IEEE Access, 8, 118050-118071.

[75] Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. Behavioral Science, 7(4), 484.

[76] Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).

[77] Wilson, T., Wiebe, J., & Hoffmann. P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of HLT/EMNLP.

[78] Sebastiani, F., & Esuli, A. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th international conference on language resources and evaluation (pp. 417-422). European Language Resources Association (ELRA) Genoa, Italy.

[79] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American society for information science and technology, 61(12), 2544-2558.

[80] Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010, November). Senticnet: A publicly available semantic resource for opinion mining. In 2010 AAAI fall symposium series.

[81] Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.

[82] Mohammad, S. M., & Turney, P. D. (2013). Nrc emotion lexicon. National Research Council, Canada, 2, 234.

[83] Hutto, C., & Gilbert, E. (2014, May). 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).

[84] Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

[85] Loria, S. (2018). textblob Documentation. Release 0.15, 2(8), 269.

[86] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature.

[87] Handayani, D., Bakar, N. S. A. A., Yaacob, H., & Abuzaraida, M. A. (2018, July). Sentiment analysis for Malay language: systematic literature review. In 2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M) (pp. 305-310). IEEE.

[88] Bakar, N. S. A. A., Rahmat, R. A., & Othman, U. F. (2019). Polarity classification tool for sentiment analysis in Malay language. IAES International Journal of Artificial Intelligence, 8(3), 259.

[89] Hijazi, M. H. A., Libin, L., Alfred, R., & Coenen, F. (2016, October). Bias aware lexicon-based Sentiment Analysis of Malay dialect on social media data: A study on the Sabah Language. In 2016 2nd International Conference on Science in Information Technology (ICSITech) (pp. 356-361). IEEE.

[90] Talbot, J., Charron, V., & Konkle, A. T.34(2021). Feeling the void: lack of support for isolation and sleep difficulties in pregnant women during the COVID-19 pandemic revealed by Twitter data analysis. International Journal of Environmental Research and Public Health, 18(2), 393.

[91] Google Trends. (n.d.). Google's Year in Search. Retrieved June 22, 2021, from https://trends.google.com/trends/yis/2020/MY/

[92] Shakeel, S., Ahmed Hassali, M. A. & Abbas Naqvi, A. 2020. Health and economic impact of covid-19: Mapping the consequences of a pandemic in malaysia. Malaysian Journal of Medical Sciences 27(2): 159–164. doi:10.21315/mjms2020.27.2.16

[93] Abdullah, N. A. S., & Rusli, N. I. A. (2021). Multilingual Sentiment Analysis: A Systematic Literature Review. Pertanika Journal of Science & Technology, 29(1).

[94] Sadia, A., Khan, F., & Bashir, F. (2018, February). An overview of lexicon-based approach for sentiment analysis. In 2018 3rd International Electrical Engineering Conference (IEEC 2018) (pp. 1-6).

[95] Ariffin, S. N. A. N., & Tiun, S. (2018). Part-of-Speech Tagger for Malay Social Media Texts. GEMA Online® Journal of Language Studies, 18(4).

[96] Pustejovsky, J., & Stubbs, A. (2012). Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. " O'Reilly Media, Inc.".

[97] Shamsudin, N. F., Basiron, H., Saaya, Z., Rahman, A. F. N. A., Zakaria, M. H., & Hassim, N. (2015). Sentiment classification of unstructured data using lexical based techniques. Jurnal Teknologi, 77(18).

[98] Iqbal, M., Karim, A., & Kamiran, F. (2015, April). Bias-aware lexicon-based sentiment analysis. In Proceedings of the 30th Annual ACM Symposium on Applied Computing (pp. 845-850).

[99] Hota, H. S., Sharma, D. K., & Verma, N. (2021). Lexicon-based sentiment analysis using Twitter data: a case of COVID-19 outbreak in India and abroad. In Data science for COVID-19 (pp. 275-295). Academic Press.