A Systematic Literature Review on Emotional Text for Malay Corpus

Hafizuddin Muhd Adnan, Hamwira Yaacob, Normi Sham Awang Abu Bakar,
Kulliyyah of Information & Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia,
hafizuddin.adnan@live.iium.edu.my, hyaacob, nsham @iium.edu.my,

Abstract— The existence of corpus, especially online corpus, fabricates benefits not only to the linguistic area, but also benefits to others, for instance to academicians especially researchers. Research trends on corpus shows that study is done to upgrade the performance and functionality, for example toward the emotional corpus. These emotional corpora been used widely not only by researcher in linguistic but also by psychologist, the legal firm and also by stakeholder by referring to the emotional annotation in the corpus to understand about their client emotion and behavior. However, the study on the Malay emotional corpus is still left behind where currently there is no available study or paper been published regarding on Malay emotional corpus and also no development method yet been study on Malay Emotional Corpus. Therefore, this paper aims to focus on applying the systematic literature review (SLR) regarding the emotional corpus. The objective of this study is to understand and identify the available technique and methods that have been used in developing the emotional corpus based on text. The methodology used in this paper is by applying a systematic approach known as the Preferred Reporting Items for Systematic reviews and Meta- Analyses (PRISMA). The articles that have been reviewed through SLR are obtained from the Scopus online database. Articles are selected based on the time range between years 2011 until years 2020.These articles are reviewed and analyzed through four PRISMA approaches which is the Identification, Screening, Eligibility and Included approach. After going through the process, this paper managed to get 35 articles for the purpose of SLR for detail reviewed. Initiating SLR in this study, several components or areas are identified for the reviewing purpose. This paper review regarding the method applied, the domain or application, modality, name of the corpus involved, the label of corpus (language) and then what is the measurement based they used in the methodology in their study. At the end of the study, this paper significantly brought a summary in helping to understand the available method and technique used for emotion recognition based on text for the purpose of emotional Corpus development hence identifying the most applicable method used. This bring as a source based for the future works in developing Malay emotional Corpus.

Keywords— *emotional corpus, emotion recognition, text and word based, Methodologies.*

I. Introduction

A text corpus, in its most basic definition, is a collection of written or spoken material which is machine-readable. Text corpora have been generally regarded as a source of information by academics, broadly across various disciplines [2] There are several corpora available in Malay language. The existence of these corpus provides different service application to the users [3]. For instance, sealang.net corpus with their service on collocation analysis and concordance, glosbe.com corpus as a global dictionary, lexilogos.com corpus is a Malay – English dictionary and prpm.dbp.gov.my corpus for define and provide detailed meaning of word in their service and application. Currently, there is a corpus been developed known as Movic. Movic stand for Malay online virtual corpus. This corpus is a virtual corpus which is serve as a one stop centre for searching activity through corpus because it was integrated together with 4 different Malay corpora behind it [3]. It used a web crawler technology to provide result when user do a searching through it. The corpora that was integrated by this Movic are from sealang.net corpus, Malay Concordance Project (MCP) corpus, MyBaca corpus and Pusat Rujukan Persuratan Melayu, Dewan Bahasa dan Pustaka (DBP) corpus. Through Movic, it gives benefits in term of time saving and effort saving when user do a searching because it concepts as a one stop centre.

Concurrently, study on the emotional corpus induced from several approaches. For instance, from speech, emoticons, audio and also from text or from word. The basis of approaches is derived from social media, websites and online corpus. From the existing study on this, it urged this paper to explore and identify the methodologies been used and applied in the development of several Emotional corpus available [1]. The paper will apply SLR in order to analysed and identify the method been used for the emotional corpus development. The study or the content been analysed in SLR is identified as a primary study even though the SLR itself considered as a secondary study [27].

One of the approaches used as a basis for emotional corpus is from word or text. Text and word are mostly available when involving social media and online website surfing to do searching and comments. As a sequence, the paper focus to identify the applicable method used for emotional corpus been developed based on text and word approaches.

The aim of this paper is to identify and summarize the available method used in developing Emotional corpus through systematic literature review (SLR). The paper applies the SLRs because it is a process of aggregating knowledge toward any topic of focused research [27].

This paper is organized as follows: Section II brief the background of the study, then Section III elaborates the methodology used for the SLR. After that, the findings are discussed in Section IV, and lastly Section V highlights the conclusion and future work.

II. Background of Study

Emotion recognition based on text or word been studied through several methods [1,2,5,7,13,15] and the purpose of study is to have an emotion annotation for those text or word been studied [9]. As a result, from these studies, several methods have been applied in developing emotional corpus [16,19,21,24]. Some examples of the text and word-based emotion recognition method are given in Table 1:

Author	**Method**	**Domain**	**Modality**
Bandhakavi, A., Wiratunga, N., Massie, S., & Padmanabhan, D. (2017).	Machine Learning	Online Website	Blogs and News Portal
Kušen, E., Cascavilla, G., Figl, K., Conti, M., & Strembeck, M. (2017, August).	Natural Language Processing	Online Websites	Lexicons
Lin, X., & Han, C. (2018, November).	Sentiment Analysis	Online Websites	Corpus
Tan, Z., Zhang, Y., Zhang, C., Huang, R., Lei, P., & Duan, X. (2019, October).	Machine Learning	Online Websites	Corpus
Zhang, C., Xie, L., Aizezi, Y., & Gu, X. (2019).	Deep Learning (Machine Learning)	Social Media	Microblogs Twitter
Guan, X., Peng, Q., Li, X., & Zhu, Z. (2019, December).	Neural Network	Social Media	News portal
Zhao, J., Yang, X., Qiao, Q., & Chen, L. (2020, December).	Machine Learning and Sentiment Analysis	School Database	School Record System

The prior study on the emotion recognition method based on the word and text caters several methods. The domain and modality not only focus on corpus rather on social media, blogs and also news portal [25,28].

As the aim of this study is to identify the available methods for emotional corpus development, therefore this study will do a categorizing based on the available emotion recognition technique, the domain based and their modality. This categorizing will cover all the available methods based on text or word generally, not only focus on Corpus yet.

As a result, once the study completed, we will identify and choose what is the most applicable method for emotion recognition based on text or word that can be applied in developing a comprehensive Corpus with emotional labels in future.

III. Methodology

While we searching for Articles on the emotion recognition based on text or word, the results showed consisting of various disciplines and publisher. Thus, the systematic review was performed through Scopus online database, where it consisting of IEEE Xplore, ProQuest, Springer Link, Elsevier and ACM.

In this study, a systematic approach known as the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) [29] is used. This approach includes the following phases:

1) Identification: By finding the number of records from selected online databases (Scopus);

2) Screening: The process involves the exclusion of the articles which is duplicated and unrelated;

3) Eligibility: Through this process, the filtration will exclude based on the reasons, then finally most relevant articles to the topic were selected.

4) Included: this final process will choose the article with the exact content of the focused topic.

In this study we will explain in detail how the articles obtained for the purpose of study is extracted. The explanation will show from first approach until the last approach. The procedures involved are discussed in the following sub-sections. In addition, our article selection criteria and filtering process are also explained.

A. *Articles Extraction Procedure*

For the first approach (Identification), the articles were searched through Scopus online database using the default setting. The key word used for searching are text emotion, word emotion and emotional corpus.

The first run of search resulted in 233 articles. The articles were identified in Scopus consisting from IEEE Explore, ProQuest, Springer Link, Elsevier and ACM.

Then move to second approach which is the screening. The articles from the search result in the first approach will be filter out from duplicated articles and exclude based on filter setting for searching.

The next section demonstrates Screening approach which involving the filtering/excluding process. Filtering/Excluding Process In this process, each of article was manually reviewed in four stages, as shown in Fig. 1.

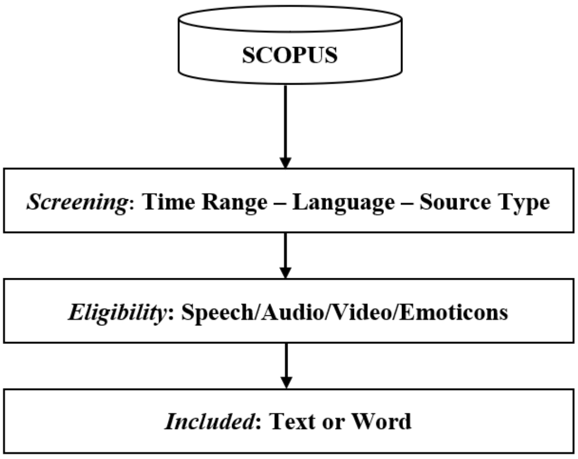
SCOPUS
Screening: Time Range – Language – Source Type
Eligibility: Speech/Audio/Video/Emoticons
Included: Text or Word

Fig. 1. Search strategy in Text or Word-based Emotion Recognition according to the PRISMA statement (Source: [4])

Through the first of review for filtering, initially, we managed to obtain 233 articles. However, there is no duplicated articles in our search result and the total articles remain as 233.

The next phase of approach is Screening by filtering the articles based on the time range of articles. We set out the time range: 2011 until 2020 for years of publication. The other criteria for filtration are due to the Language written and also the source type of the articles. Here, for the language, choose on the English and the source type is only for article from the journal and conference paper. From this process, 98 articles were filtered out and the remaining are 135 articles.

Then the third phase, the Eligibility approach. In this phase, the full text of articles was assessed for their eligibility accordingly to the focus topic or content. In this phase, 78 articles excluded because of the reason the main focus of study is based on speech, audio, video and emoticons. The remaining articles we have now are 57 articles to be reviewed.

The final phase is the included approach. This phase, we read in detail the whole articles and then we filter out by removing articles with the irrelevant topics. Any articles that do not discussing on text or word approach for emotion recognition were excluded which is 31 articles. Finally, the remaining articles available for the review are 26 articles.

The diagram in Fig. 1 shows the phases involves to perform the four approaches in order to obtain the relevant articles that meet the main focus for our study criteria.

IV. FINDINGS

This paper set the main aims of the literature search where to obtain the several categories of information for emotion recognition based on the method used, the domain, modality and the measurement used for the development of emotional corpus.

In general, the development of emotional corpus involves several methods for emotion recognition [17,21,22,25]. This paper identified that these several methods been used since the early study on corpus until this time being.

We identified that at the recent study on emotion recognition for corpus based on text or word been studied through the machine learning and sentiment analysis method. Some studies they used a combination method of sentiment analysis and machine learning method [26].

In category of machine learning and sentiment analysis method for emotional corpus development, the studies were initiated through several techniques. These two methods seem correlate each other as combining method.

Several study on emotion recognition done based on the text or word for the emotional corpus development. Xu and Wu [22,23] used the machine learning method together with sentiment analysis method for emotional annotation. They do a dynamic adjustment algorithm of sentiment word weight and classify it and distinguish the word from the emoticons [24] do some modification of on a feature based on CRF (Conditional random fields) for their model of emotional keywords annotation.

Then Rao [17] have a combination method, machine learning and sentiment analysis through Pearson's correlation as their measurement. The annotation is based on the Sem Eva and Sina from Chinese corpus. In addition, study by [22] also used Machine learning method for their emotional analysis in different perspectives including word-based approach. The study involved the Chinese corpus, Ren-CECps.

Furthermore, studies focus only on machine learning method also been initiated. For instance, study on hierarchical emotion classification and emotion component analysis on Chinese micro-blog posts corpus been initiated [23] through feature extractions and data mining algorithm in machine learning for Weibo corpus, also a Chinese corpus. For their hierarchical emotion classification, each test example from online blogposts is classified from the top level successively to the bottom level.

Another study in Indonesian language emotion annotation involve the English corpus called WordNet through machine learning using KNN; Naïve bayes; SVM-SMO based on the online web portal, a collection of historical stories and folklore in Indonesia [20]. This study used WordNet Corpus as their basis to understand the linguistic phrase that been used for emotion translation.

Then study on how to determine word-emotion associations from tweets by multi-label classification. In this study they used Multi label classification in machine learning for Word2Vec English corpus [5]. Their measurement is on the Intrinsic evaluation which is include Brown clusters, POS tags, and word2vec embeddings in their study.

In addition, there is also study based on Natural language processing [8] through the cosine similarity measure in several corpus which is, NRC, DepecheMood and EmoSenticNet corpus.

Consequently, study on emotion recognition for online corpus been initiated through these methods since 2011 until 2020. For instance, in machine learning and sentiment analysis [12,25,19]. Sentiment analysis method been used in study by [12,26]. Even though all these recent studies are based on the modality online corpus but most of studies are focus on the Chinese language.

V. CONCLUSION AND FUTURE WORK

Initially, the systematic literature review in this paper brought some new ideas in the area of emotional corpus. According to the discussions in the reviewed papers, there are several methodologies that we understand and identified which can be applied for the implementation of Emotional Corpus especially to the Malay online corpus which is currently no study or publication yet is done related to emotional Malay corpus.

The study leads a clear direction for the future works related to the emotional corpus development especially for Malay online corpus. Perhaps, the future works of this research area is towards the implementation of Emotional Malay corpus through most applicable method available which will benefit to researchers and others to use it as reference in their research or works.

REFERENCES

[1] Apandi, N., & Jamil, N. (, November). "An analysis of Malay language emotional speech corpus for emotion recognition system". In 2016 *IEEE Industrial Electronics and Applications Conference* (IEACon) (pp. 225-231). IEEE.(2016)

[2] Asyafie, M. A., Harun, M., Shapiai, M. I., & Khalid, P. I. (December). "Identification of phoneme and its distribution of Malay language derived from Friday sermon transcripts". In 2014 *IEEE Student Conference on Research and Development* (pp. 1-6). IEEE.(2014)

[3] Bakar, N. S. A. A. "The Development of an Integrated Corpus for Malay Language". In *Computational Science and Technology* (pp. 425-433). Springer, Singapore. (2020).

[4] Bakar, N. S. A. A., Yaacob, H., Handayani, D., & Abuzaraida, M. A. (July). "Malay Online Virtual Integrated Corpus (MOVIC): A Systematic Review". In 2018 *International Conference on Information and Communication Technology for the Muslim World* (ICT4M) (pp. 243-248). IEEE.(2018)

[5] Bravo-Marquez, F., Frank, E., Mohammad, S. M., & Pfahringer, B. (, October). "Determining word-emotion associations from tweets by multi-label classification". In 2016 *IEEE/WIC/ACM International Conference on Web Intelligence* (WI) (pp. 536-539). IEEE.(2016)

[6] Hijazi, M. H. A., Libin, L., Alfred, R., & Coenen, F. (, 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.(2016)

[7] Jimenez, I. A. C., García, L. C. C., Violante, M. G., Marcolin, F., & Vezzetti, E." Commonly Used External TAM Variables in e-Learning, Agriculture and Virtual Reality Applications". *Future Internet*, 13(1), 7. (2021).

[8] Kušen, E., Cascavilla, G., Figl, K., Conti, M., & Strembeck, M. (, August). "Identifying emotions in social media: comparison of word-emotion lexicons". In 2017 5th *International Conference on Future Internet of Things and Cloud Workshops* (FiCloudW) (pp. 132-137). IEEE.(2017).

[9] Lee, L. W., & Low, H. M. "The development and application of an online Malay language corpus-based lexical database". *Kajian Malaysia*, 32(1), 151. (2014).

[10] Lee, L., & Low, H. (2011). "Developing an online Malay language word corpus for primary schools". *International Journal of Education and Development Using ICT*, 7(3).

[11] Li, C. H., Yang, J. C., & Park, S. C. "Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet". *Expert Systems with Applications*, 39(1), 765-772.(2012).

[12] Lin, X., & Han, C. (, November). "Chinese text sentiment analysis based on improved convolutional neural networks". In 2018 *IEEE 9th International Conference on Software Engineering and Service Science* (ICSESS) (pp. 296-300). IEEE.(2018).

[13] Matalon, Y., Magdaci, O., Almozlino, A., & Yamin, D. "Using sentiment analysis to predict opinion inversion in Tweets of political communication". *Scientific reports*, 11(1), 1-9. (2021).

[14] Medhat, W., Hassan, A., & Korashy, H. "Sentiment analysis algorithms and applications: A survey". *Ain Shams engineering journal*, 5(4), 1093-1113. (2014).

[15] Nicholson, J., & Baldwin, T. (, June). "Web and corpus methods for Malay count classifier prediction". In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Companion Volume: Short Papers (pp. 69-72).(2009)

[16] Quan, C., & Ren, F. "A blog emotion corpus for emotional expression analysis in Chinese". *Computer Speech & Language*, 24(4), 726-749.(2010).

[17] Rao, Y., Quan, X., Wenyin, L., Li, Q., & Chen, M. "BuildingWord-Emotion Mapping Dictionary for Online News". In *SDAD@ ECML/PKDD* (pp. 28-39). (2012).

[18] Saloot, M. A., Idris, N., & Mahmud, R. "An architecture for Malay Tweet normalization". *Information Processing & Management,* 50(5), 621-633. (2016).

[19] Wang, P., Nakov, P., & Ng, H. T. "Source language adaptation approaches for resource-poor machine translation". *Computational Linguistics*, 42(2), 277-306. (2016).

[20] Winarsih, N. A. S., & Supriyanto, C. (, August). "Evaluation of classification methods for Indonesian text emotion detection". In 2016 *International seminar on application for technology of information and communication* (ISemantic) (pp. 130-133). IEEE.(2016).

[21] Wu, Y., Kita, K., & Matsumoto, K. "Three predictions are better than one: Sentence multi - emotion analysis from different perspectives". *IEEJ Transactions on Electrical and Electronic Engineering*, 9(6), 642-649. (2014).

[22] Wu, Y., Kita, K., Ren, F., Matsumoto, K., & Kang, X. (November). "Modification relations based emotional keywords annotation using

conditional random fields". In 2011 4th *International Conference on Intelligent Networks and Intelligent Systems* (pp. 81-84). IEEE.(2011)

[23] Xu, H., Yang, W., & Wang, J "Hierarchical emotion classification and emotion component analysis on Chinese micro-blog posts". *Expert systems with applications*, 42(22), 8745-8752. (2015).

[24] Xu, Y. Q., Zhu, Y. H., Wen-hua, W., & Gao, L. C. (March). "A dynamic adjustment algorithm research of sentiment word weight based on context". In 2011 3rd *International Conference on Computer Research and Development* (Vol. 3, pp. 19-22). IEEE.(2011)

[25] Zhang, C., Xie, L., Aizezi, Y., & Gu, X "User Multi-Modal Emotional Intelligence Analysis Method Based on Deep Learning in Social Network Big Data Environment". *IEEE Access*, 7, 181758-181766. (2019).

[26] Zhao, J., Yang, X., Qiao, Q., & Chen, L. (, December). "Sentiment Analysis of Course Evaluation Data Based on SVM Model". In 2020 *IEEE International Conference on Progress in Informatics and Computing* (PIC) (pp. 375-379). IEEE. (2020).

[27] B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, S. Linkman, "Systematic literature reviews in software engineering – A tertiary study," *Information and Software Technology*, Vol. 52, Issue 8, pp 792-805, (2010).

[28] Bandhakavi, A., Wiratunga, N., Massie, S., & Padmanabhan, D. "Lexicon generation for emotion detection from text". *IEEE intelligent systems*, 32(1), 102-108. (2017).

[29] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.," *PLoS Med*., vol. 6, no. 7, p. e1000097, July (2009).