Sentiment Analysis for Malay Language:Systematic Literature Review

Dini Handayani,
School of Computing and IT
Taylor's University
Selangor, Malaysia
dinioktarina.dwihandayani@taylors.edu.my

Normi Sham Awang Abu Bakar,
Hamwira Yaacob
Department of Computer Science
International Islamic University
Malaysia,
Kuala Lumpur, Malaysia
{nsham,hyaacob}@iium.edu.my

Mustafa Ali Abuzaraida
Computer Science Department
Faculty of Information Technology
Misurata University, Libya
abuzaraida@it.misuratau.edu.ly

***Abstract*—Recent research and developments in Sentiment Analysis (SA) have simplified sentiment detection and classification from textual content. The related domains for these studies are diverse and comprise fields such as tourism, costumer review, finance, software engineering, speech conversation, social media content, news and so on. SA research and developments field have been done on various languages such as Chinese and English language. However, SA research on other languages such as Malay language is still scarce. Thus, there is a need for constructing SA research specifically for Malay language. To understand trends and to support practitioners and researchers with comprehension information with regard to SA for Malay language, this study exhibit to review published articles on SA for Malay language. From five online databases including ACM, Emerald insight, IEEE Xplore, Science Direct, and Scopus, 2433 scientific articles were obtained. Moreover, through the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Statement, 10 articles have been chosen for the review process. Those articles have been reviewed depend on a few categories consisting of the aim of the study, SA classification techniques, as well as the domain and source of content. As a result, the conducted systematic literature review shed some light about the starting point to research in term of SA for Malay language.**

Keywords—sentiment analysis, opinion mining, sentiment mining, Malay, classification

I. INTRODUCTION

Sentiment Analysis (SA) is a study that is used to recognize as well as interpret opinion, sentiments, emotions, attitude, feeling and subjective information for any issue, product, movie or an event [1]–[4]. SA has been a very emerging research field in the past two decades. SA is currently comprehensively connected in various domain namely tourism [5], Costumer review [6]–[8], finance [9]–[11], software engineering [12], speech conversation [1], social media content [2], [3], [13]–[19], movie review [4], news [20] and so on.

The main task of SA is to distinguish whether a text has positive, negative or neutral sentiments. Positive sentiments are those which consist of good words or appraisals [2]. Whilst, the negative sentiments are those which consist of bad words or criticism [2]. Furthermore, neutral sentiments are those that are neither positive nor negative [2]. Previous SA studies have been done to classify whether a text having positive or negative sentiments [3], [4], [6], [8], [10], [11], [13], [14], [17], [21]. However, some other works are attempted to distinguish a text between positive, negative or neutral sentiments [1], [9], [12], [15], [16], [20], [22].

At present, SA is conducted at three levels: word, sentence, and document. The most important part is to know the sentiment information about the smallest part of the content itself, which is a word known as a lexicon. Moreover, manual interpretation of a sentiment lexicon is tedious in terms of time and effort. On the basis of the problem, many studies have emerged with the aim of automating lexicon. This includes classification at the 'term' level, and the classification result can help in the estimation of bigger pieces of content.

The research on SA is categorized into two approaches: corpus-based machine learning (ML) approach and lexicon-based method. The commonly used techniques for corpus based ML methods are Naive Bayesian (NB) [1], [3], [6], [8], [9], [14], [17], Support Vector Machines (SVM) [1], [3], [6], [8]–[11], [14], [15], [17], K-means algorithm (KNN) [2], [11], [19], Decision three [4], [8], [10], Adaboost [4], bagging [4], stacking [4], Logistic Regression [9], [10], [14], [17], Boosted Tree [10], Vector Space Model (VSM) [18], skip gram [21], Continuous Bag of Words [21], Recurrent Neural Network (RNN) [21], Stochastic Gradient Descent (SGD) [22], and Random Forest Classifier [3], [5], [10], [22]. While for Lexicon based method the commonly used lexicon data set are NRC [7], [8], VADER [1], [9], WordNet [15], SentiWordNet [9], [16], [17], Public Sentiment Lexicon [15], TextBlob [2], [9], [17], AFINN Dictionary [2], NTUSD [11] and SentiStrength [7]. Some researcher combined both approaches on their previous work called NB, SVM, and VADER [1], NB, SVM, Decision Three and NRC [8], NB, Logistic Regression and VADER, SentiWordNet, TextBlob [9], NB, SVM and SentiWordNet, TextBlob [17], SVM, K-means algorithm and NTUSD [11], SVM and WordNet, Public Sentiment Lexicon [15], KNN and TextBlob, AFINN Dictionary [2].

SA research has been conducted in Indonesian [20],

978-1-5386-7525-0/18/$31.00 ©2018 IEEE
DOI 10.1109/ICT4M.2018.00063

Persian [7] and Arabic language [18], [19]. Nevertheless, earlier work in SA field have been conducted on widespread languages such as Chinese [5], [6], [11], [14] and English language [1]–[4], [8]–[10], [12], [13], [15]–[17], [21], [22]. With reason are many available resources in Chinese and English language. On the other hand, SA research on other languages such as Malay language is still scarce. Thus, there is a need for constructing SA research specifically for Malay language.

During the literature review process, the articles on SA for Malay language cut across various disciplines. Consequently, conducting a comparative investigation of articles is challenging since distinctive journals and different scientific domains have diverse aims, content and investigation techniques.

The main goal of this study is to compile and assess the recent scientific literature review on SA for Malay language via literature reviews. Generally, the credibility of a systematic review relies upon on what was done, what was found, and the clarity of the details [23]. Moreover, the reviewing quality of systematic reviews varies, limiting readers' ability to evaluate the qualities and shortcomings of those articles.

In Section II, the methodology of this study is conferred. Discussion of this study is expounded in Section III. Conclusion and future work are described in Section IV.

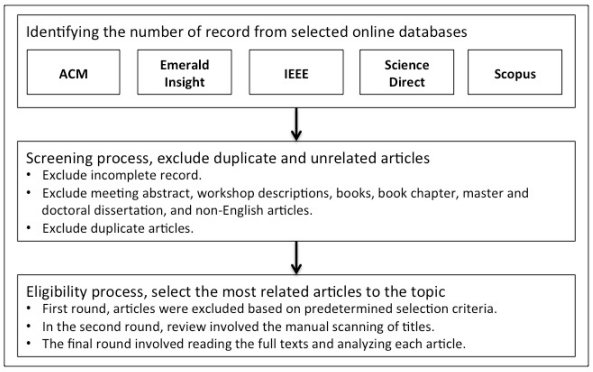

Fig 1. Search strategy in Sentiment Analysis for Malay Language according to the PRISMA statement (Source: [23])

II. METHODOLOGY

In this review, the recognition of acceptable studies was executed according to Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) [23], which include a few steps including (as shown in fig 1):

A. Identifying the number of record from selected online databases.

Utilizing five online databases including ACM, Emerald insight, IEEE Xplore, Science Direct, and Scopus carried out the systematic review. In this study, the presently search string was examined "Sentiment Analysis for Malay Language". This search string resulted on a total of 2433 articles, where 2158 articles were identified from ACM. 62 articles were identified from Emerald insight. Meanwhile, from IEEE Xplore six articles were identified. Moreover, 191 articles were identified from Science Direct. Finally, 16 articles were identified from Scopus.

B. Screening process, exclude duplicate and unrelated articles.

Four criteria were used in this screening process, to choose and acknowledge the articles for further review. Articles were eliminated if they did not meet the following criteria:

- Exclude incomplete record (e.g. no authors names, no articles title, no articles years etc.). The remaining articles after this step are 2403.
- Exclude meeting abstract, workshop descriptions, books, book chapter, master and doctoral dissertation, and non-English articles. After this step, the numbers of articles are 2359.
- 167 articles were excluded, due to duplicate articles.
- Articles must be to some extent current. In this respect, we chose articles that were released between 2008 and 2018. This 10 years period could be considered to correspond to the main research period of sentiment analysis. 436 articles were excluded after this step.

After the screening process, as a result 1754 articles reminded for the next eligibility process. Which is choose the most relevant articles to the topic.

C. Eligibility process, select the most related articles to the topic.

In this step, each article was manually checked on in three rounds. This was done to exclude non-SA based works.

- During the first round of review, articles were eliminated based on predetermined selection criteria. The articles that did not address SA were excluded. After applying our selection criteria, 618 articles remained.
- In the second round, the review included the manual filtering of titles. This review round eliminates articles in which the topic of the discussion was centered on subjects other than SA for Malay Language. At the end of this round, 11 articles were left.
- The last round included perusing the full texts and investigate each article. Finally, articles that do not focus on SA for Malay Language were eliminated from the analysis. In this round, only 10 articles remain for review process.

III. DISCUSSION

Accentuate on SA for Malay Language, reviews on 10 chosen articles were accomplished depend on a few categories including the aim of the study, SA classification technique, and domain and source of content as shown in Table I.

A. Category 1: The Aim of the study

Previous studies on SA for Malay Language have different aims. The aim of previous work done by Isa, Puteh, & Kamarudin, 2013, is to propose preprocessing methods for stemming Malay content, to be specific Reverse Porter Algorithm (RPA) and Backward Forward Algorithm (BFA) and Artificial Immune Network (AIN) for extracting sentiment meaning from Malay newspaper articles [24].

On the other hand, previous work done by A Alsaffar & Omar, 2014, has three main goals [25]. The first goal is to assess the most precise feature selection technique for Malay sentiment classification. The second goal is to dispose of which classification strategy performs best. The third goal is to explore how feature selection strategy contributes to the advancement of the classification result of three ML classifiers on Malay SA.

Moreover, the main goal of an earlier study done by Ahmed Alsaffar & Omar, 2015, was to find the best set of features that improve Malay SA classifications [26]. Additionally, this earlier study proposes a new technique for automated SA in Malay movie reviews.

The main aim of the study done by Tan, Lam, Azlan, & Soo, 2016, was to explore the techniques for analyzing the tweets to figure trends, patterns and business decision [27]. This previous study carry out SA classification on a set of tweets related to five local mobile telecommunication service in both English and Malay.

Furthermore, the objective of the work done by Hasbullah, Maynard, Chik, Mohd, & Noor, 2016, was to propose automated content analysis tools to help Malaysian legal firm and the Malaysian official leaders recognize public sentiment through social media [28]. This previous work explores and applies Semantic Role Labeling (SRL) to classify the content dataset.

On the other hand, the objective of the work done by Darwich, Noah, & Omar, 2017, was to propose a minimally supervised sentiment lexicon model for Malay [29]. The expected result from the proposed approach is to classify the sentiment word to become positive or negative sentiment.

Moreover, Al-Moslmi, Omar, Albared, & Alshabi, 2017, attempted to propose an enhanced ensemble of ML classification methods for Malay SA [30]. The expected result of this previous study is the improvement of the performance Malay sentiment based classification.

Additionally, Nasharuddin, Abdullah, Azman, & Kadir, 2017, has objective to introduce a cross-lingual sentiment lexicon acquisition method for Malay and English languages [31]. This previous work develops an automatic method for the cross-lingual sentiment lexicon construction. The proposed method is expected as a solution to solve a problem in Malay SA.

A study done by Chekima & Alfred, 2018, has objective is to proposed a framework to handle the most trivial challenges posed by Malay social media content (casual text) called RojakLex [32]. RojakLex contains 4 different lexicons, namely: MySentiDic (Malay Lexicon), English Lexicon (translated version of MySentiDic), Emotion lexicon (9 online emoticons) and Neologism lexicon (neologism words).

Chekima, Rayner Alfred, & Chin, 2018, in their work has a goal to come out with new feature sets that convert the conventional sentiment feature extraction technique [33]. The proposed method has taken contextual valence shifter into attention from a specific perspective as compared to the previous research concerning on Malay language. They proposed a rule-based model to address explicit Valence shifter in Malay SA, which includes negation, intensifies, diminished, and contrast.

The second reviewed component is the SA classification technique. Various methodologies were employed to establish SA for Malay Language. Most of these studies were using corpus-based ML approach and lexicon-based method.

B. Category 2: SA Classification Technique

Isa et al., 2013, use Artificial Immune System (AIS) to set up SA algorithm [24]. This study also proposed pre-processing methods for stemming Malay content, particular RPA, BFA and AIN. The result from stemming the Malay content utilizing new RPA shows enhancement in processing time compares to BFA. However, the SA accuracy utilizing AIN with both stemming methods show nearly comparative result.

Furthermore, A Alsaffar & Omar, 2014, utilizes 6 feature selection methods for SA research, specifically: information gain (IG), principal components analysis (PCA), relief-F, gini index, uncertainty, and chi-square [25]. While for the classification techniques, ML approach techniques used in this previous study, namely SVM, NB and KNN. The main outcome of this study is to find out the right feature selection technique as well as classification methods for developing the Malay SA. The best result was achieved by using SVM with 87% accuracy.

Similarly with A Alsaffar & Omar, 2014, in the following year, Ahmed Alsaffar & Omar, 2015 utilize KNN for the classification technique [26]. However, this previous study combines the lexicon approach as well as ML approach. The performance of the classifier in this previous study was obtained at 86.43%.

In contrast, Tan et al., 2016, uses a lexicon-based approach to classify the Malay sentiment text [27]. This previous work

proposed new sentiment corpus to overcome the lack of resources in Malay SA namely SentiLexM. The accuracy for confidence level 72.9% was 78.5%, while for confidence level 95% the accuracy was obtained 84.1%.

On the other hand, Hasbullah et al., 2016, proposed the rule-based method to automate analyze social media content such as Facebook and Twitter [28]. The proposed method was developed based on an existing rule-based system.

The previous study done by Darwich et al., 2017, makes use the lexicon-based approach to generate SA model [29]. In this previous study, WordNet Bahasa and Kamus Dewan were utilized as a corpus to create a minimally supervised sentiment lexicon model. The proposed model achieves 73.7% accuracy.

Furthermore, the earlier study done by Al-Moslmi et al., 2017, use three classification approaches which are similar with previous study done by Ahmed Alsaffar & Omar, in 2014 and 2015, namely: NB, SVM and KNN [30]. Additionally, this previous work also utilize five ensemble classification algorithms, namely: bagging, stacking, voting, adaboost and metacost. The experimental result shows that the use of SVM, NB, and KNN with Metacost algorithm achieved the best result with an F-measure of 85.81%.

Moreover, Nasharuddin et al., 2017, utilize word score summation method to classify the sentiment of the news articles [31]. Some part of speech (POS) tags are being tested using the Word Score Summation method to obtain the sentiment of the news articles. The proposed techniques carry out 50% of accuracy and work excellent for verbs and negations in both English and Malay news articles.

On the other hand, Chekima & Alfred, 2018 employ RojakLex as a lexicon-based approach to classify the informal Malay sentiment text [32]. Human-centric accession was utilized to translate neologism terms. The proposed system achieved 79,28% which shows improvement from the baseline.

Finally, Chekima et al., 2018 proposed a rule-based model to handle explicit Valence shifter in Malay SA [33]. The proposed system shows the improvement in Malay SA compares to the existing techniques after considering valence shifter. A total if 16 features have been recognize to look at the result of Valence shifter. During the features contraction, Malay lexicon namely MySentiDic were utilize as a baseline.

C. Category 3:Domain and Source of Content

The last reviewed component is the content of evaluation. The various content was used to develop SA for Malay Language. Most of the content comes from website, online forum, blog, social media, etc.

Isa et al., 2013 used the formal Malay language from Malaysian newspaper, to be specific Berita Harian (BH) [24]. The corpus contains 1080 of sentence, with an average of 40 words per sentence, with a total approximation of 4000 words. SA research using newspaper content can be used to prevent negative effect when it associates with serious issues [24].

Next, previous study done by Ahmed Alsaffar & Omar, 2014 and 2015, Al-Moslmi et al., 2017, was using opinion corpus for Malay (OCM) which is composed of an assortment of reviews web pages in Malay language [25], [26], [30]. OCM consist of 2,000 reviews; 1,000 are considered as positive sentiment while the other 1,000 considered as a negative sentiment. Additionally, for the sentiment lexicon, 2,478 Malay sentiment lexicons were utilized [26]. The English WordNet was used to gather more exact sentiment words that are utilized in English. These words were deciphered into the Malay language for appropriation as sentiment words in the lexicon. Similarly, Darwich et al., 2017 use WordNet Language and Board Dictionary, to create a minimally supervised lexicon sentiment model [29].

Furthermore, Tan et al., 2016 use SentiLexM which contains 26,004 English and Malay sentiment words [27]. A total of 51,295 tweets are collected. The new corpus with derived from existing English corpus specifically AFINN.

Similarly, Hasbullah et al., 2016, utilize social networks such as Twitter and Facebook [28]. The proposed idea in this previous study is to help Malaysian legal firms and the Malaysian official leaders to understand public sentiment through social media.

In the previous work done by Nasharuddin et al., 2017, a total of 883 of an unstructured English and Malay news articles were chosen as the test documents [31]. The news articles consist of a diverse essence, which includes politics, business, economy, executive reports, and sports.

Similar to Tan et al., 2016 and Hasbullah et al., 2016, Chekima & Alfred, 2018 in their previous study also utilize Twitter and Facebook as a source of data collection [32]. A total of 8,026 comments were composed for evaluation purposes.

Moreover, a total of 14,780 documents defined in previous work by Chekima et al., 2018 [33]. The data were collected from a newspaper, Malaysian lifestyle media forum (car, mesra, and lazada), and social media, such as Twitter and Facebook.

The elaborated systematic review expects to draw consideration to different aims, proposed techniques to create the SA for Malay Language, and different content for assessment. Generally, there is a reasonable demand for research and development about SA for Malay Language.

IV. CONCLUSION AND FUTURE WORK

The conducted systematic literature review shed some light about the starting point to research in term of SA for Malay language. The reviewed component offers the clear thought with respect to the various pursuits, proposed classification procedure and different content used in each

study of SA for Malay language. The systematic literature review also assures that there is strong reason for more research and some study are shifting within the similar direction.

In this study, we will expand considerably the latter work; by composing an SA for Malay language corpus to be used as references for evaluation in future studies on sentiment analysis for Malay language. Although there are varieties of sentiment analysis corpus, at first glance they seem less availability for sentiment analysis corpus in Malay language.

ACKNOWLEDGMENT

This research is fully funded from International Islamic University's RIGS grant, ID: RIGS-17-156-0731.

REFERENCES

[1] S. Maghilnan and M. R. Kumar, "Sentiment Analysis on Speaker Specific Speech Data," *2017 Int. Conf. Intell. Comput. Control*, 2018.

[2] S. Ahuja and G. Dubey, "Clustering and Sentiment Analysis on Twitter Data," *2017 2nd Int. Conf. Telecommun. Networks (TEL-NET 2017)*, vol. 2, no. 1, pp. 178–184, 2017.

[3] J. Vora and A. M. Chacko, "Sentiment Analysis of Tweets to Identify the Correlated Factors That Influence an Issue of Interest," *2017 2nd Int. Conf. Telecommun. Networks (TEL-NET 2017)*, 2017.

[4] S. Kumar and R. Singh, "Comparative analysis of ensemble classifiers for sentiment analysis and opinion mining," *2017 3rd Int. Conf. Adv. Comput. Autom.*, pp. 1–5, 2017.

[5] Y. H. Gu, S. J. Yoo, Z. Jiang, Y. J. Lee, Z. Piao, H. Yin, and S. Jeon, "Sentiment analysis and visualization of Chinese tourism blogs and reviews," *2018 Int. Conf. Electron. Information, Commun.*, 2018.

[6] Y. Fang, H. Tan, and J. Zhang, "Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness," *IEEE Access*, vol. 6, pp. 20625–20631, 2018.

[7] M. E. Basiri and A. Kabiri, "Translation is Not Enough: Comparing Lexicon-based Methods for Sentiment Analysis in Persian," *2017 Int. Symp. Comput. Sci. Softw. Eng. Conf.*, no. Ml, pp. 36–41, 2017.

[8] Z. Singla, S. Randhawa, and S. Jain, "Sentiment Analysis of Customer Product Reviews Using Machine Learning," *2017 Int. Conf. Intell. Comput. Control*, 2017.

[9] S. Sohangir, N. Petty, and D. Wang, "Financial Sentiment Lexicon Analysis," *2018 12th IEEE Int. Conf. Semant. Comput.*, pp. 0–3, 2018.

[10] P. Chakraborty, U. S. Pria, M. R. A. H. Rony, and M. A. Majumdar, "Predicting Stock Movement using Sentiment Analysis of Twitter Feed," *2017 6th Int. Conf. Informatics, Electron. Vis. 2017 7th Int. Symp. Comput. Med. Heal. Technol.*, pp. 1–6, 2017.

[11] W. Zongyue and Q. Sujuan, "A Sentiment Analysis Method of Chinese Specialized Field Short Commentary," *2017 3rd IEEE Int. Conf. Comput. Commun.*, 2017.

[12] M. R. Islam and M. F. Zibran, "A Comparison of Software Engineering Domain Specific Sentiment Analysis Tools," *SANER 2018*, no. 978, pp. 487–491, 2018.

[13] D. K. Tayal and S. K. Yadav, "Analysis of Sentiments & Polarity Computation of Opinions," *2017 2nd Int. Conf. Telecommun. Networks (TEL-NET 2017)*, pp. 2–7, 2017.

[14] Y. Li and B. Shen, "Research on Sentiment Analysis of Microblogging Based on LSA and TF-IDF," *2017 3rd IEEE Int. Conf. Comput. Commun.*, 2017.

[15] K. Lavanya and C. Deisy, "Twitter Sentiment Analysis Using Multi-Class SVM," *2017 Int. Conf. Intell. Comput. Control*, 2017.

[16] T. S. K. Chaitanya, V. S. Harika, and B. Prabadevi, "A sentiment analysis approach by identifying the subject object relationship," *2017 2nd Int. Conf. Commun. Electron. Syst.*, no. Icces, pp. 62–68, 2017.

[17] N. Bhan and M. D'silva, "Sarcasmometer using Sentiment Analysis and Topic Modeling," *2017 Int. Conf. Adv. Comput. Commun. Control*, pp. 1–7, 2017.

[18] F. S. Al-anzi and D. AbuZeina, "A Micro-Word based Approach for Arabic Sentiment Analysis," *2017 IEEE/ACS 14th Int. Conf. Comput. Syst. Appl.*, pp. 910–914, 2017.

[19] D. Abuaiadah, D. Rajendran, and M. Jarrar, "Clustering Arabic Tweets for Sentiment Analysis," *2017 IEEE/ACS 14th Int. Conf. Comput. Syst. Appl.*, pp. 449–456, 2017.

[20] Y. Liang, B. Fu, and Z. Li, "Sentiment Tendency Analysis of THAAD event in Indonesian News," *2017 14th Web Inf. Syst. Appl. Conf.*, pp. 211–214, 2017.

[21] Y. Sharma, G. Agrawal, P. Jain, and T. Kumar, "Vector Representation of Words for Sentiment Analysis Using GloVe," *2017 Int. Conf. Intell. Commun. Comput. Tech.*, pp. 279–284, 2017.

[22] A. A. Aziz, A. Starkey, and M. C. Bannerman, "Evaluating Cross Domain Sentiment Analysis using Supervised Machine Learning Techniques," *Intell. Syst. Conf. 2017*, no. September, pp. 689–696, 2017.

[23] 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, Jul. 2009.

[24] N. Isa, M. Puteh, and R. M. H. R. M. H. R. Kamarudin, "Sentiment classification of malay newspaper using immune network (SCIN)," *Proc. World Congr. Eng.*, vol. 3, pp. 1543–1548, 2013.

[25] A. Alsaffar and N. Omar, "Study on feature selection and machine learning algorithms for Malay sentiment classification," *Conf. Proc. - 6th Int. Conf. Inf. Technol. Multimed. UNITEN Cultiv. Creat. Enabling Technol. Through Internet Things, ICIMU 2014*, pp. 270–275, 2014.

[26] A. Alsaffar and N. Omar, "Integrating a Lexicon based approach and K nearest neighbour for Malay sentiment analysis," *J. Comput. Sci.*, vol. 11, no. 4, pp. 639–644, 2015.

[27] Y. F. Tan, H. S. Lam, A. Azlan, and W. K. Soo, "Sentiment analysis for telco popularity on twitter big data using a novel Malaysian dictionary," *Front. Artif. Intell. Appl.*, vol. 282, pp. 112–125, 2016.

[28] S. S. Hasbullah, D. Maynard, R. Z. W. Chik, F. Mohd, and M. Noor, "Automated Content Analysis: A Sentiment Analysis on Malaysian Government Social Media," in *IMCOM '16: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, 2016, p. 30:1--30:6.

[29] M. Darwich, S. A. M. Noah, and N. Omar, "Minimally-supervised sentiment lexicon induction model: A case study of malay sentiment analysis," *Multi-disciplinary Trends Artif. Intell.*, vol. 10607, no. November, pp. 225–237, 2017.

[30] T. Al-Moslmi, N. Omar, M. Albared, and A. Alshabi, "Enhanced Malay Sentiment Analysis with an Ensemble Classification Machine Learning Approach," *Journal of Engineering and Applied Sciences*, vol. 12, no. 20. pp. 5226–5232, 2017.

[31] N. A. Nasharuddin, M. T. Abdullah, A. Azman, and R. A. Kadir, "English and Malay cross-lingual sentiment lexicon acquisition and analysis," vol. 424, p. 4154, 2017.

[32] K. Chekima and R. Alfred, "Sentiment Analysis of Malay Social Media Text," in *Computational Science and Technology*, 2018, pp. 205–219.

[33] K. Chekima, Rayner Alfred, and K. O. Chin, "Rule-Based Model for Malay Text Sentiment Analysis," *Int. Conf. Comput. Sci. Technol. ICCST 2017 Comput. Sci. Technol.*, pp. 172–185, 2018.

TABLE I. TABLE ON SENTIMENT ANALYSIS FOR MALAY LANGUAGE

No	Reference (Title, Author, Year)	Aim	SA Classification Technique	Domain and Source of Content
1	Sentiment classification of Malay newspaper using immune network (SCIN), Norulhidayah Isa, Mazidah Puteh, and Raja Mohamad Hafiz Raja Kamarudin, 2013, [24]	To propose preprocessing techniques for stemming Malay text called RPA and BFA and AIN for extracting sentiment from Malay newspaper articles	ML approached AIS	1080 of sentence, with an average of 40 words per sentence, with total estimation of 4000 words
2	Study on Feature Selection and Machine Learning Algorithms For Malay Sentiment Classification, Ahmed Alsaffar, Nazlia Omar, 2014, [25]	• To evaluate the most accurate feature selection method of Malay sentiment classification. • To determine which learning method performs best. • To investigates how feature selection methods contribute to the improvement of the classification performance of three machine-learning classifiers on Malay SA.	ML approached Feature selection: IG, PCA, relief-F, gini index, uncertainty, and chi-square. Classification: SVM, NB and KNN.	OCM Consist of 2,000 reviews; 1,000 are considers as positive sentiment; 1,000 considered as negative sentiment
3	Integrating a Lexicon Based Approach and K Nearest Neighbor for Malay Sentiment Analysis, Ahmed Alsaffar and Nazlia Omar Center, 2015, [26]	• To identify an optimized set of features that enhances the Malay sentiment analysis and classifications. • To propose, implement and evaluate a new approach for automatic sentiment analysis of Malay movie reviews	Hybrid technique KNN WordNet	OCM 2,478 Malay sentiment lexicon from WordNet translate to Malay
4	Sentiment Analysis for Telco Popularity on Twitter Big Data Using a Novel Malaysian Dictionary, Yi-Fei Tan, Hai-Shuan Lam, Asyraf Azlan, and Wooi King Soo, 2016, [27]	• To explores the techniques for analyzing the tweets to determine trends, patterns and business decision • To perform SA on a set of tweets related to five local mobile telecommunication service in both English and Malay	Lexicon approach SentiLexM	SentiLexM 26,004 English and Malay sentiment words; The corpus consist of 51,295 words from twitter
5	Automated Content Analysis: A Sentiment Analysis on Malaysian Government Social Media, Siti Salwa Hasbullah, Diana Maynard, Rita Zaharah Wan Chik, Farahwahida Mohd, Myzan Noor, 2016, [28]	• To propose automated content analysis tools to help Malaysian legal firm and the Malaysian official leaders understand public sentiment through social media • To explore and apply SRL to classify the content data set	Rule-based approach	Comment on official Malaysian government leaders' social media such as Twitter and Facebook
6	Minimally-Supervised Sentiment Lexicon Induction Model: A Case Study of Malay Sentiment Analysis, Mohammad Darwich, Shahrul Azman Mohd Noah, and Nazlia Omar, 2017, [29]	To propose a minimally supervised sentiment lexicon model for Malay	Lexicon based approach	WordNet Bahasa and a Malay language dictionary named Kamus Dewan
7	Enhanced Malay Sentiment Analysis with an Ensemble Classification Machine Learning Approach, Tareq Al-Moslmi, Nazlia Omar, Mohammed Albared, Adel Alshabi, 2017, [30]	To propose an enhanced ensemble of ML classification methods for Malay SA	ML approach NB, SVM and KNN Five ensemble classification algorithms	OCM
8	English and Malay Cross-lingual Sentiment Lexicon Acquisition and Analysis, Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah, Azreen Azman, Rabiah Abdul Kadir, 2017, [31]	• To introduce a cross-lingual sentiment lexicon acquisition method for Malay and English languages • To develop an automatic method for the cross-lingual sentiment lexicon construction which in turn will help the Malay language SA in the future	Lexicon based approach Word Score Summation technique	883 of an unstructured English and Malay news articles
9	Sentiment Analysis of Malay Social Media Text, Khalifa Chekima, Rayner Alfred, 2017, [32]	To propose a framework to handle a few of the most common challenges posed by Malay social media content (casual text) called RojakLex	Lexicon based approach RojakLex	8,026 Twitter and Facebook comments
10	Rule-Based Model for Malay Text Sentiment Analysis, Khalifa Chekima, Rayner Alfred, Kim On Chin, 2017, [33]	• To construct new feature sets that transform the traditional sentiment feature extraction technique • To propose a rule-based model to handle explicit Valence shifter in Malay SA	Rule-based approach	14,780 document were successfully annotated from newspaper, Twitter and Facebook