IMPROVING ACCURACY IN SENTIMENT ANALYSIS FOR FORMAL AND INFORMAL MALAY LANGUAGE USING SEMANTIC INFORMATION

NURUL AIDA OSMAN[1], DUC NGHIA PHAM[2]

[1] Lecturer, Department of Computer & Information Sciences, Universiti Teknologi PETRONAS, Perak, Malaysia

[2] Principal Data Scientist, Advanced Intelligence Lab, MIMOS Berhad, W.P. Kuala Lumpur, Malaysia

E-mail: [1]nurulaida.osman@utp.edu.my, [2]nghia.pham@mimos.my

ABSTRACT

This paper presents a lexicon-based sentiment analysis model that is specifically designed for analyzing sentiments in formal and informal Malay language texts. The main challenge when dealing with Malay language is handling noisy texts and sarcasm. To overcome this hurdle, we propose a method to enhance the lexicon by incorporating semantic information and gloss information from Kamus Dewan and synonym chains from WordNet Bahasa to obtain sentiment terms. The goal is to utilize these semantic information and sentiment terms to enhance the accuracy of sentiment analysis in both formal and informal Malay language. The proposed model generates a semi-supervised Malay sentiment lexicon for both formal and informal Malay language and utilizes semantic information to further enhance its performance accuracy. We manually annotated two evaluation datasets, one in formal Malay and one in informal Malay, with sentiment values. We then conducted experiments on two models corresponding to formal Malay language and informal Malay language using these datasets. The results demonstrated that the proposed approach achieved an average accuracy of 90.0% and 88.4% for formal and informal Malay language, respectively. This confirmed that semantic information can effectively boost the performance accuracy of sentiment analysis model (in comparing with existing models) for Malay language.

Keywords: *Sentiment Analysis, Opinion Mining, Lexicon Based Sentiment, Malay Lexicon, Informal Malay Lexicon, Formal Malay Lexicon*

1. INTRODUCTION

Sentiment analysis (or opinion mining) is one popular research field studied over the years. It has been widely applied due to the huge increment and growth of internet content especially in social media. Lots of opinions can be obtained from product reviews, forums, blogs, and social media posts. It is quite challenging for a user to find those relevant contents, analyze the opinions expressed in those contents, then organize and summarize them into useful insights [1]. It is even more difficult for a computing system to perform these tasks. Given its great beneficial impact for practical applications, several research works (mainly based on machine learning) have been proposed by researchers from both academia and the industry to address this challenge [2].

Existing research in sentiment analysis primarily revolves around the problem of understanding opinions by categorizing text into positive or negative classes. The main challenge faced by sentiment analysis is the difficulty in handling informal language styles in communication. The use of slangs and sarcasms further creates additional complexities and challenges for researchers working on sentiment analysis. For analyzing sentiment of formal and informal Malay texts, the common challenges and limitations include the handling of Bahasa Pasar, which involves the use of "*Manglish*" (which mixes Malay and English terms in one sentence), short forms, and emoticons.

In this paper, we propose a new method for mining the opinion of Malay contents in both formal and informal writing styles. We extract Malay textual data from Twitter, blogs, and news and manually annotate them with sentiment values to form a gold standard dataset for performance evaluation. Though, compiling a sentiment lexicon manually is quite a tedious process regarding time and effort. Therefore, several research works have been proposed for automatically generating sentiment lexicon. These works can be classified into two main categories: (i) dictionary-based approaches that utilize lexical resources and digital dictionaries, and (ii) corpus-based approaches that compute occurrence statistics obtained from text corpora.

With that, we propose a least supervised model for generating Malay language sentiment lexicon model, which utilizes both WordNet Bahasa (a Malay variant of WordNet) and Kamus Dewan (a Malay language dictionary). The two sources are selected due to their reliable lexical words and widely used human-defined glossary information. The model first extracts subjective terms from both WordNet Bahasa and Kamus Dewan and then severalizes them into positive and negative classes. It then uses words from these two sentiment-oriented classes as a bootstrap to automatically computes a reliable and highly coverage sentiment lexicon for enhancing sentiment analysis tasks for Malay language.

2. RELATED WORKS

Sentiment analysis has attracted interests from both academic researchers and industry practitioners. This work aims to improve the accuracy of sentiment analysis for both formal and informal in Malay language. In Malaysia, people tend to use informal style to express their opinions on social media, which usually creates lots of noises for sentiment analysis [3]. These noisy texts are generally hard for software applications to process and obtain the information from. As mentioned by Saloot and Mahmud [4], noisy text can be considered as main challenges in using NLP applications. Therefore, prior research papers related to our work were investigated.

2.1 Sentiment Analysis

Hartmann *et al.* [5] highlighted on a fundamental of sentiment to human communication. The paper mentioned on how lexicons can generate sentiment score from individual words and expressions. For that, an experimental framework consisting of different types of research questions (including data characteristics and analytical resources) were conducted to enable informed method decisions depending on the application environment. To form reliable performance evaluation, few features been selected including content variables, the sentiment classes, and the sizes of text. The classifier needs to have the capability to handle noisy and unstructured text. Hence, a new sentiment analysis model known as SiEBERT has been proposed. This approach is used to obtain the highest possible accuracy.

Sharma *et al.* [6] has conducted a survey on sentiment analysis of Twitter that uses machine learning technique. One of the main challenges in processing the tweets is because of the tweet complex format. The size of tweet text is so short and contains set of issues with the use of slang and abbreviations. In this study, the methodologies and models used in Twitter sentiment analysis research were studied and analyzed. As a result, they found that Tweeter Analyzer has the capability to solve their current issues. Machine learning techniques were proved to be simpler and more efficient than Symbolic techniques for sentiment analysis of tweets.

AlBadani *et al.* [7] introduced a novel method based on machine learning for Twitter sentiment analysis. It combined the “universal language model fine-tuning” (ULMFiT) model with support vector machine (SVM) to increase the efficiency and accuracy of sentiment word detection specifically on user preferences for certain products based on their comments. The experiment results on three datasets showed that their proposed model outperforms others.

2.2 Sentiment Analysis for Malay Language

As our study focus on Malay language, we reviewed existing research on sentiment analysis for Malay language both for formal and informal languages. Malaysian millennials tend to use distorted words in their social media interactions. Nazman *et al.* [8] attempted to classify these morphological distortions. The study was based on tweets extracted from a sample of 50 randomly selected Twitter accounts from Malaysia. They found that distorted words were mainly modified based on their inflections to fit some sounds. Many millennials deliberately coined these distortions to appear trendy. Meanwhile, some merely followed the trend without knowing the actual word. However, this study did not conduct any experiments. They found out that the normalization task was challenging because it had similarities with spell checking but differed in that ill-formedness in text messages.

Another interesting work by Zabidi and Sheikh [9] implemented a model based on deep learning architecture for sentiment analysis of informal Malay tweets. The authors developed their Malay sentiment analysis model based on the Convolutional Neural Network (CNN). They first built a Word2Vec embedding model from scratch to transform text input data to numerical vectors before feeding these vectors to a CNN for training. Hyper-parameter tuning was also conducted. The final optimized model achieved an accuracy of up to 77.59% in their experimental evaluation.

Abu Bakar *et al.* [10] reviewed and discussed the state of the art of sentiment analysis for Malay language, identified its challenges, and outlined different potential directions for future works. The papers covered various approaches, datasets, performance measures, and pre-processing techniques used in the previous works on sentiment analysis of the Malay text. It found that noisy Malay texts exhibit several informal patterns and rules such as ambiguous and messy informal words. It also acknowledged that there was no available technique to handle Malay sarcasm at the time of the review was conducted.

Zabha *et al.* [11] proposed a lexicon-based sentiment analysis classifier for cross-lingual Malay twitter data. They developed two lexicons of languages that combined English and Malay languages in the systems. Experimental results showed that the model could predict sentiment with high recall rate. However, it achieved a low average accuracy due to frequent usage of slangs, shorten words and dialects in the test data.

Rodzman *et al.* [12] proposed a lexicon-based sentiment analysis on Malay documents in different specific domains. The objective was to develop different sentiment lexicons on specific domains such as song, politic and product to build the best lexicon-based sentiment analysis classifier for Malay documents in each specific domain. From the experiment, the result showed that lexicon-based classification outperformed Naïve Bayes sentiment analysis classification in overall three topics which were song, politic and product in average of 70% compared to 50% average for Naïve Bayes. However, they still could not handle the informal language issues due to the usage of slangs, shorten words and dialects.

Abu Bakar *et al.* [13] proposed an enhancement of Malay social media text normalization for lexicon-based sentiment analysis. They proposed an architecture of the Malay Text Normalizer which comprises four dictionaries which are dialect dictionary, trend dictionary, global Malay/English dictionary (“kite” can be referred to “saya/kita” or “layang-layang”. and noisy dictionary. The proposed Malay Text Normalizer was proven to improve performance of lexical based method. However, it still cannot handle the Malay sarcasms sentences.

3. OUR APPROACH: LEXICON-BASED SENTIMENT MODEL FOR MALAY LANGUAGE

This section describes a lexicon-based sentiment analysis model that is designed to work in scenarios where resources are limited, and human participation is minimized. The model requires only two input sentiment orientations, positive and negative words, and then generates a sentiment lexicon in three main steps:

- Step 1 – Generating synonym terms from a WordNet using predefined seed terms.
- Step 2 – Extracting the human-defined gloss information of entries in Kamus Dewan.
- Step 3 – Combining the results from Steps 1 and 2.

The main components and flow process for our approach for sentiment analysis in Malay language are depicted in Figure 1.

Step 1 involves extracting synonym words from WordNet using predefined seed words. The seed words are used to form synonym words by searching all the synsets related to the words within the predefined seed synsets. The lexicon is then expanded with newly added words as the synonym words expanded on every repetitive search. The goal is to protect the sentiment properties as the synonym word is retrieved. Synonym words consist of words meaning as well as semantic orientation.

Step 2 involves extracting the words glossary information in Kamus Dewan. In this step, all words that contain within the glossary of an entry word are considered to have a similar semantic orientation to the entry word itself. For example, if a word generated from Step 1 matches a word within the glossary of the entry in the Kamus Dewan, the entry itself will be added to the lexicon.

Step 3 combines the results from Steps 1 and 2. In this step, the coverage of the sentiment lexicon can be highly optimized by utilizing the two most prominent features of semantic relation in Kamus Dewan which are synonymy and hyponymy. As mentioned previously, the experiment utilized two widely used lexical resources for Malay language which are WordNet Bahasa [14] and Kamus Dewan (3rd Edition) by Dewan Bahasa dan Pustaka [15].

The main reason for selecting a dictionary-based approach is that a dictionary has a high coverage of words used in natural language and a high level of semantic relation equivalencies between an entry word and human-defined glossary words. Additionally, a dictionary preserves rich content of lexical or semantic relations such as synonymy and hyponymy. In this experiment, the model generated focused only on a set of words labelled as positive or negative as the final lexicon, while all the objective words were discarded. As the glossary information in WordNet Bahasa is limited, we added the Kamus Dewan (3rd Edition), an openly available Malay dictionary with reliable glossary information.

3.1 Generating List of Synonym Words from WordNet Bahasa

WordNet Bahasa (WNB) is the Global WordNet Association standardized independent WordNet version specifically for Malay language. It consists of 49,668 synsets, 145,696 senses and 64,431 unique words. In this work, seven predefined positive and negative words each was used as a seed word [16].

```
S(pos) = Synset{
        ('good.a.01'),
        ('nice.a.01'), ('excellent.a.01'), ('positive.a.01'),
        ('fortunate.a.01'),                 ('correct.a.01'),
        ('superior.a.01')
}

S(neg) = Synset{
        ('bad.a.01'),
        ('nasty.a.01'),
        ('poor.a.01'),                      ('negative.a.01'),
        ('unfortunate.a.01'),                  ('wrong.a.01'),
        ('inferior.a.01')
}
```

Algorithm 1 shows the pseudo-code for generating synonym chains from WNB as follows:

```
Input: Predefined 7 seed terms of positive and
negative terms

Output: Generated positive and negative terms of
WordNet Bahasa relation expansion (pos_terms and
neg_terms)

# Step 1:
FOR each positive term in positive seed set:
    Look up for other positive synonyms words in
    WNB and add to temporary positive seed set;
    Look up for other similar words in WNB and add
    to temporary positive seed set;
    Look up for antonyms words in WNB and add to
    temporary positive seed set;
    Add temporary positive seed set to positive terms
    (pos_terms);

# Step 2:
FOR negative term in negative seed set:
    Look up for other negative synonyms words in
    WNB and add to temporary negative seed set;
    Look up for other similar words in WNB and add
    to temporary negative seed set;
    Look up for antonyms words in WNB and add to
    temporary negative seed set;
    Add temporary negative seed set to negative
    terms (neg_terms);
```

Algorithm 1: Pseudo-code algorithm for generating synonym chains from WNB

Based on the Algorithm 1, the sets of expansion positive and negative words from WordNet have been aligned with WordNet Bahasa. There are total of 561 positive terms and 538 negative terms of WordNet Bahasa. The total output from this step will be used as the input for the following step which is mining Kamus Dewan for semantic information.

3.2 Extracting the Kamus Dewan for Semantic Information

Using the expanded seed words of positive and negative WordNet Bahasa from the previous step, the glossary information of entries in Kamus Dewan (KD) was extracted to enhance coverage of the lexicon. The detail process of this step is shown as follows (refer to Figure 2).

As depicted in Figure 2, the semantic relation is used in which the glossary of an entry is related to the entry term semantically. All words form WordNet Bahasa are compared with the glossary information from Kamus Dewan database. All matched glossaries are retrieved, and entry terms are added to the list of positive and negative Kamus Dewan. The results from the series of conducted experiments shows that using relations between entries and adjectives within their glossaries proven t to get the best results as compared to only nouns, verbs, and adverbs. Here, 561 positive words in WordNet Bahasa and the 538 negative words in WordNet Bahasa are used as the initial seed words, while the positive seed words (KD+) and negative seed words (KD-) involved in this current step of extracting Kamus Dewan are initially empty sets.

Algorithm 2 below shows the pseudo-code for extracting Kamus Dewan for semantic information.

```
Input: Kamus Dewan, list of positive and negative
expansion terms from WNB (pos_terms and
neg_terms)
```

```
Output: Generated positive and negative terms from
Kamus Dewan

# Step 1:
FOR each term in the WNB expansion list (output
from Algorithm 1):
    Look up for the similar terms in Kamus Dewan
    gloss information;
    FOR each similar terms found:
        Remove “bkn” to avoid mismatches;
Get all matches terms and add to gloss terms;
Normalize all the terms in gloss terms to lower case
terms;

# Step 2:
IF gloss terms consist of word "tidak", "tak", "bukan",
"jangan" or "belum":
    Call negation function to handle negation;

# Step 3:
FOR each positive term in pos_term matches with
gloss terms in Kamus Dewan:
    Add it to positive terms pos_terms;
FOR each negative terms in neg_terms that matches
with gloss terms in Kamus Dewan:
    Add it to negative terms neg_terms;
```

Algorithm 2: Pseudo-code algorithm for extracting Kamus Dewan for semantic information

In Algorithm 2, we used negations to avoid the issue of negation in glossaries, if a matched gloss term is led by a negation term, the corresponding entry is not retrieved, since negators used to reverse the sentiment orientation of the words. Negation in the context of Malay language is expressed using the negators such as ‘tidak’, ‘bukan’, ‘jangan’ and ‘belum’ [17, 18]. For example, one of the glossaries for the entry word ‘sakit’ is ‘tidak sihat’, the negation rule would ignore the word of ‘sihat’ being matched, and the entry ‘sakit’ being added to the negative class KD-.

4. EXPERIMENTS AND RESULTS

The experiments were conducted on publicly available twitter and news datasets containing formal and informal Malay sentences. We conducted two different sets of experiments which is the formal Malay language model and informal Malay language model. The experiments have been done in two different domains which are formal Malay news and informal Twitters comments. The offline evaluation with historical datasets was chosen. The baseline dataset consists of 1000 formal Malay language and 1000 informal Malay language. Accuracy is one of the most widely applicable and fundamental measures in which sentiment analysis models are evaluated [19]. It is computed as follows.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} * 100 \qquad (1)$$

where tp, tn, fp and fn represent true positives, true negatives, false positives, and false negatives respectively. This evaluation metric is used to measure the model’s overall performance specifically the accuracy of sentiment classification of terms. Table 1 depicts the results for accuracy for both informal and formal Malay language model. We performed the experiments for over six weeks to obtain the best accuracy results. The pruning process has been done by weekly to filter out the un-certainty terms and wrongly categorized terms. The pruning process is done manually to clean up the resultant sentiment lexicon.

Table 1: Accuracy results for both formal and informal Malay language model

Accuracy %						
Polarity	W1	W2	W3	W4	W5	W6
Formal Positive	69.8	69.39	72.39	88.8	88.6	88.8
Formal Negative	80.6	80.4	78.6	76.6	91.2	91.2
Informal Positive	43.8	45.6	78.4	81.8	82.19	84.2
Informal Positive	74.6	81	72.8	74	74	92.6

Graph has been plotted for the achieved results (refer to Figure 3). Results has been analyzed and discussed in the next section.

5. DISCUSSION AND CONCLUSION

Table 2 illustrates the accuracy model to be labelled as FormalTerms for formal positive and formal negative terms, and InformalTerms for informal positive and informal negative terms. As referred to Table 2, the model accuracy results are 88.8% for formal positive terms and 91.2% for formal negative terms for week 6 (w6). The overall accuracy for both formal and informal Malay model as follows (refer to table 2):

Table 2: Overall accuracy for both formal and informal Malay model

Accuracy Positive (%)	Accuracy Negative (%)	Accuracy Overall (%)
FormalTerms	88.8	90.0
InformalTerms	84.2	88.4

The model's overall accuracy for formal Malay terms classification is 90.0%. From here we could see that there is a 1.60% drop in accuracy which is from 90.0% to 88.4% when we moved from a formal Malay to an informal Malay classification problem. This shows that the informal Malay, which involves classifying terms as subjective or objective, and classifying subjective terms as positive or negative, is a more challenging problem. This is one of the main problems [20], in which the assumption of the lexical resource or dictionary used only contains subjective terms where classifying terms as positive and negative while ignoring he objective class.

Furthermore, dealing with informal Malay language containing lot of noisy text become a major problem. According to [10], so many garbage contents including noisy text expressed in social media platforms like Facebook and Twitter that lead to some issues in sentiment analysis process. To the best extend of knowledge, only few works on sentiment analysis focusing on Malay text and one of them is a work by [21]. This is a vital process because most NLP applications are normally trained and generated on formal and structured text [22].

From the total of 55,725 entries in Kamus Dewan (KD), there is only 6,155 terms (11%) were labelled as subjective (2,869 positive and 3,286 negative), while the remaining 89% were labelled as objective. Liu [23] explained that a manually pruning and cleaning process to build sentiment lexicon is a one-time effort. All the errors and irrelevant terms were discarded manually from the generated lexicon and produced a structured and clean version publicly available lexicon for future research on Malay sentiment analysis and natural language processing tasks.

The experimental results demonstrate that our proposed approach produces higher accuracy scores for both formal and informal Malay models and outperforms other existing works. Although we were unable to compare our approach against existing works on the same datasets, it is worthy to note that none of the existing works reported an accuracy more than 90% like ours. Since there are very few works in Malay sentiment analysis, the comparison can only be made against similar works in Malay sentiment models, such as [8]. There is another work in Malay model [11], but the dataset is too small, with less than 2000 tweets used. Although there are limitations in this comparison, it can serve as an indicator of the current standing of this project. Our proposed approach achieves a more than 90% accuracy, which is comparably better than the overall accuracy of only 81% by the existing deep learning approach by [8]. Our proposed approach has a significant advantage due to its data-driven characteristic, meaning that the accuracy improves with more data being fed into the model.

The weakness of our approach is that it does not distinguish between different word senses. This is because it is designed to analyze only typical and common formal and informal texts found on the social platforms and online webs. For instance, the word "suka", which means "like" in English, can express positive sentiment in "saya suka awak" or can be used as a "kata ganda" or double words in "saya hanya suka-suka sahaja mendendangkan lagu ini". Our proposed model produces a neutral score for "suka-suka". However, for phrases that contain "suka" in its idiom list with a positive score, our model will supersede the neutral score with the positive score for "suka" when applied in common positive texts, such as "dia suka", "saya suka", "mereka suka", and "kami suka". Hence, disambiguating between different word senses in Malay language is a promising avenue for future research to improve the accuracy of the results.

REFERENCES:

[1] Liu, B.: Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing, (1), 1-38. Retrieved from https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf (2010).

[2] Pang, B., Lee, L., & Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 79-86. Retrieved from http://portal.acm.org/citation.cfm?id=1118693.1118704 (2002).

[3] Chekima, K. and Alfred, R.: ''Sentiment analysis of Malay social media text,'' Com-put. Sci. Technol., vol. 488, pp. 205–219, Feb. (2018).

[4] Saloot, M. A., Idris, N. and Mahmud, R.: ‘‘An architecture for Malay Tweet normalization,’’ Inf. Process. Manage., vol. 50, no. 5, pp. 621–633, Sep. (2014).

[5] Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. More than a Feeling: Accuracy and Application of Sentiment Analysis (2022).

[6] Aiswarya, M. K: Sentiment Analysis of Twitter using Machine Learning. Journal of Research Proceedings, 1(2), 216-225 (2021).

[7] AlBadani, B., Shi, R., & Dong, J.: A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. Applied System Innovation, 5(1), 13 (2022).

[8] Nazman, N. N. N., Chuah, K. M., & Ting, S. H. Tryna b Kewl: Textual Analytics of Distorted Words among Malaysian Millennials on Twitter (2020).

[9] Ong, J. Y., Mun'im Ahmad Zabidi, M., Ramli, N., & Sheikh, U. U.: Sentiment analysis of informal Malay tweets with deep learning. IAES International Journal of Artificial Intelligence, 9(2), 212 (2020).

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

[11] Zabha, N. I., Ayop, Z., Anawar, S., Hamid, E., & Abidin, Z. Z. Developing cross-lingual sentiment analysis of malay Twitter data using lexicon-based approach. Int. J. Adv. Comput. Sci. Appl., 10(1), 346-351 (2019).

[12] Rodzman, S. B., Rashid, M. H., Ismail, N. K., Abd Rahman, N., Aljunid, S. A., & Abd Rahman, H.: Experiment with Lexicon Based Techniques on Domain-Specific Malay Document Sentiment Analysis. In 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE) (pp. 330-334). IEEE (April 2019).

[13] Bakar, M. F. R. A., Idris, N., & Shuib, L.: An Enhancement of Malay Social Media Text Normalization for Lexicon-Based Sentiment Analysis. In 2019 International Conference on Asian Language Processing (IALP) (pp. 211-215). IEEE (November 2019).

[14] Bond, F., Lim, L.T., Tang, E.K., Riza, H.: The combined wordnet bahasa. NUSA: Linguistic studies of languages in and around Indonesia 57, 83-100 (2014)

[15] Perkamusan, D.B.d.P.B.: Kamus dewan. Dewan Bahasa dan Pustaka, (1984)

[16][18] Turney, P. D., & Littman, M. L.: Measuring praise and criticism: Inference of se-mantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4), 315-346 (2003).

[17] Idris, A.A.: Modality in Malay. (1980)

[18] Kroeger, P.: External negation in Malay/Indonesian. Language 90(1), 137-184 (2014)

[19] Rezaeinia, S. M., Ghodsi, A., & Rahmani, R.: Improving the accuracy of pre-trained word embeddings for sentiment analysis. arXiv preprint arXiv:1711.08609 (2017).

[20] Esuli, A., Sebastiani, F.: Determining Term Subjectivity and Term Orientation for Opinion Mining. In: EACL, (2006)

[21] Handayani, D., Awang Abu Bakar, D. N. S.,Yaacob, H., and Abuzaraida,M. A.: ‘‘Sentiment analysis for Malay language: systematic literature review,’’ in Proc. Int. Conf. Inf. Commun. Technol. Muslim World (ICT4M), pp. 305–310 (Jul. 2018).

[22] Baldwin T. and Li, Y.: ‘‘An in-depth analysis of the effect of text normalization in social media,’’ in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., pp. 420–429 (2015).

[23] Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5(1), 1-167 (2012).

FIGURES:

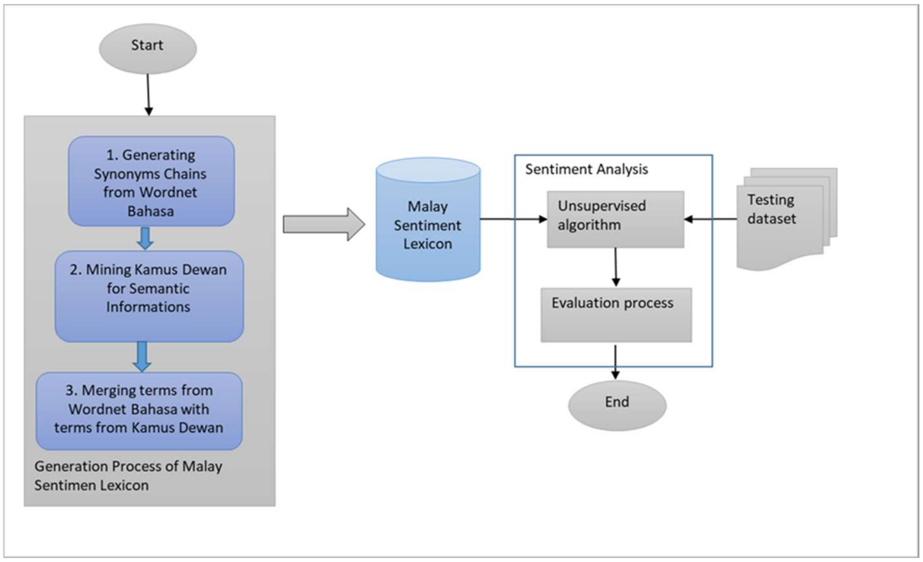
Start
1. Generating Synonyms Chains from Wordnet Bahasa
2. Mining Kamus Dewan for Semantic Informations
3. Merging terms from Wordnet Bahasa with terms from Kamus Dewan
Generation Process of Malay Sentimen Lexicon
Malay Sentiment Lexicon
Sentiment Analysis
Unsupervised algorithm
Evaluation process
Testing dataset
End

Figure 1: Main Components And Flow Process Of Proposed Approach

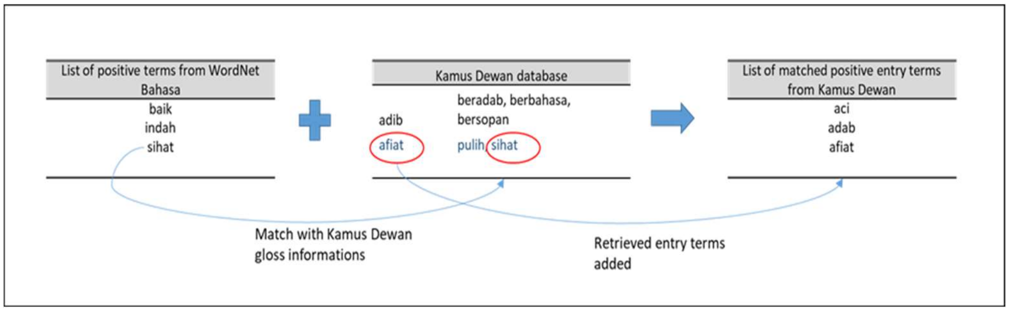
List of positive terms from WordNet Bahasa
baik
indah
sihat
Kamus Dewan database
adib
beradab, berbahasa, bersopan
afiat
pulih, sihat
List of matched positive entry terms from Kamus Dewan
aci
adab
afiat
Match with Kamus Dewan gloss informations
Retrieved entry terms added

Figure 2: Process Of Mining Kamus Dewan For Semantic Informations

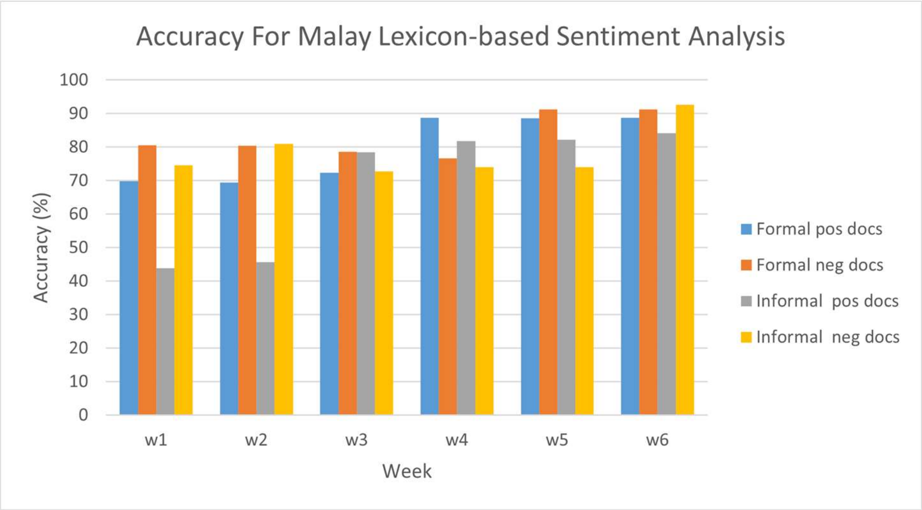

Figure 3: Graph For Accuracy For Formal Positive, Formal Negative And Informal Positive, Informal Negative