Journal of Theoretical and Applied Information Technology
30th June 2023. Vol.101. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BILINGUAL SENTIMENT ANALYSIS ON MALAYSIAN SOCIAL MEDIA USING VADER AND NORMALISATION HEURISTICS

JAMES MOUNTSTEPHENS[1], MATHIESON TAN ZUI QUEN[2], LAI PO HUNG[3]

[1,2,3]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia

Email: james@ums.edu.my[1*] (corresponding author), bond071396@gmail.com[2], laipohung@ums.edu.my[3]

ABSTRACT

This research addresses a number of important issues involved in performing Sentiment Analysis (SA) on Malaysian Social Media (SM), including an analysis of bilingual or mixed language, choice of sentiment lexicon, normalisation heuristics, and the use of public datasets. This work is the first to quantify the level of language mixing in informal Malaysian text. Analysis of the 2M tweet Malaya dataset revealed a significant level of English sentiment content in Malaysian social media (13.5%), demonstrating the neccessity of a bilingual approach to Malaysian Sentiment Analysis. Significant patterns in noisy Malaysian SM text were identified and heuristics for normalising them were devised. The popular and effective English lexicon-based SA system VADER (*Valence Aware Dictionary and sEntiment Reasoner*) was translated to Malay using automatic and manual methods, with the combination of English and Malay VADER yielding a bilingual SA system. A subset of the Malaya dataset was both corrected and extended from two to three classes in order to properly test the bilingual SA system. Bilingual VADER with normalisation heuristics was able to achieve an impressive level of performance on a three-class problem (accuracy=0.71, mean F1=0.72), as compared to Malay VADER alone and several popular machine learning-based algorithms.

Keywords: *Sentiment Analysis, Malaysian Social Media, Mixed language, VADER, Normalisation*

1. INTRODUCTION

Sentiment Analysis is "the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral." [1]. This task is an important and active research topic with significant applications but also considerable challenges [2][3]. Given input text, Sentiment Analysis systems output the text's valence as either positive, neutral, or negative, making it a 3-class problem. We note that a large number of Sentiment Analysis systems have attempted to reduce the problem to only 2 classes (positive and negative) which greatly simplifies the task but is of questionable correctness [4]. Non-neutral valence comes largely from the positive and negative words present in the text but neutral valence is a result of their absence, making it a much more open-ended and difficult problem.

Ideally, Sentiment Analysis systems should be able to effectively handle a wide range of text content from a variety of sources [5]. However, the broader the content, the more challenging the task, and some text sources are more difficult than others. Unlike formal articles, Social Media text is especially noisy and challenging: informal language, typos, poor grammar, and many other issues exist in Social Media, making the process of text normalisation before performing Sentiment Analysis important [6][7].

In practice, Sentiment Analysis systems are usually either Lexicon-based or are based on Machine Learning, each with advantages and disadvantages [8][9][10]. Lexicon (or knowledge)-based systems use a specific list of words with sentiment ratings in conjunction with a set of rules which allow them to process whole sentences and to handle negation and intensity. Machine Learning-based Sentiment Analysis treats the task as a classification problem and trains a classifier

using text data which has been labelled with valence [11].

Machine Learning-based SA has received considerable attention recently (eg. [12][13]) but the single best approach is still not settled. ML-based systems require large datasets which might not always be available, whereas LB systems do not. And while ML systems have achieved very impressive results on restricted domains, they often fail to generalise to broader domains once trained [14][15]. In contrast, LB systems usually start with a very broad lexicon and can easily update their coverage by adding to it. Plus, unlike the "black box" of a ML system, the lexicon and rules can be understood by people and can be used to better understand a problem domain and increase our knowledge. The preceding concepts and issues apply to all SA systems but SA on SM in a Malaysian context has several added and potentially serious complications [16][17].

The main research gap that this work addresses is the bilingual problem of Malay social media (i.e. the mixture of Malay and English languages) which has been identified as a difficulty in earlier work [6][14][15] but has not yet been quantified, analysed, or explicitly addressed. So far, most non-Malaysian SA research (including existing systems and datasets) have addressed problem domains that use a single language (primarily English). However, it is known informally and is acknowledged by prominent Malaysian scholars that the colloquial language of Malaysia, *Bahasa Rojak*, is a mixed language – primarily of Malay and English [18]. At an everyday level, this degree of language mixing is significant enough to warrant government campaigns to preserve pure Malay [19]. And within the specific domain of Malaysian SA it has been acknowledged that Malaysian Social Media text contains both Malay and English content [6][14][15].

However, despite these acknowledgements, the following important research questions about English content have not been addressed. What is the actual proportion of English content in informal Malaysian text? For SA, is it significant enough to be worth addressing explicitly? If English content carries important sentiment, how does this relate to Malay sentiment content in the same text? Does it replace, match, or contradict Malay sentiment? How can the mixture of languages be dealt with effectively for SA, or even exploited to improve performance?

The answers to these questions will affect performance for all forms of SA on Malaysian SM text. Most obviously, a lexicon-based approach would have to deal also with an English lexicon rather than a purely Malay one. But even ML based approaches that do not require an explicit lexicon are still affected implicitly by the increased complexity of a bilingual search space. Adequate handling of English content would give the *potential* for performance improvements proportional to the level of English sentiment content. For example, if it were found that x% of Malaysian tweets expressed sentiment using only English words then a pure Malay lexicon would automatically fail at least x% of the time. It would also mean that *up to x*% improvement would exist if English were handled properly. Addressing these questions seriously for the first time, and an initial investigation in exploiting this potential are the main purposes of this paper. As will be shown later, significant English sentiment content is found in Malaysian social media text and it does affect SA performance.

Appropriate for this analysis, we will concentrate on a lexicon-based approach here, although an ML system will also be tested for benchmarking purposes. Some work using lexicon-based SA on Malaysian SM has been conducted so far, with a degree of success. Alfred and Chekima [14][15] have attempted to construct a Malay sentiment lexicon by translating existing English sentiment lexicons, including Sentiwordnet [20]. However, the prominent English lexicon-based system VADER [21][22] has not yet been explored for its potential translation to Malay. VADER currently has over 3400 citations and has a number of advantages which may outperform other lexicon-based systems in this domain. Like SentiWordnet, VADER covers a broad domain. However, unlike Sentiwordnet, it is a full SA system that comprises a set of rules for applying the sentiment lexicon. And unlike SentiWordnet, it is explicitly designed for SM short text and contains a comprehensive list of emoticons and short forms. VADER's sentiment ratings are from human raters and can be considered trustworthy. Its code and lexicon are publicly-available in English, making it possible to analyse and extend. And unlike the work of Alfred and Chekima [15], VADER handles 3 class output which is a much more difficult problem.

VADER's advantages have already motivated attempts to translate it to Swedish [23], German [24], Bengali [25], and Assamese [26] with promising results. However, despite the potential shown by this earlier work, VADER has never yet been translated to the Malay language.

Here, we explore its translation to Malay using both machine translation and manual translation. The latter is a considerable task. VADER will provide the lexicon and the basic rules for processing tweets used in this work. Various methods of combining English and Malay lexicons for conducting bilingual Sentiment Analysis will be explored.

Appropriate Malaysian SM data is crucial for the English analysis and the bilingual SA system development and testing just described. Ideally, the dataset would be large, broad domain, reliably labelled with 3 sentiment classes, and publically-available. Such datasets exist for English (eg. [8]) but not yet for Malay.

Other researchers [14][15][16][17] have noted that it is a serious shortcoming of current SA work in Malaysia that each researcher uses their own proprietary dataset for testing: some are 2 class, some are narrow domain, and none are public. This makes it impossible to objectively assess current performance of Malaysian SA; high performance results quoted previously may not be as they seem.

The best currently-available candidate we have found is from the Malaya Natural-Language-Toolkit library for Bahasa Malaysia [27] which is large (>2M tweets), broad domain, and publicly-available. Unfortunately, this dataset has only 2 class labels (positive and negative) but due to its other advantages we will use it here and conduct time-consuming manual correction work to extend a subset of it to 3 classes.

Lastly, analysis of this data will also be used to develop normalisation heuristics specifically for Malaysian SM. Patterns of repetition and negation will be identified and implemented before the SM data is processed by the bilingual VADER SA system.

It must be pointed out that this work is only at an exploratory level and is not intended to completely solve the bilingual sentiment analysis problem (ie to produce a system that exceeds the performance of all others). The research scope here was to study the relative changes in performance for Malay and English lexicons individually, and then in combination. The results of a bilingual approach over a standard machine learning approach on the same data would be encouraging. This also implies that we will not *directly* compare our system's performance to existing Malaysian SA work. The main reason for this was stated earlier: in contrast to much English-language SA research, Malaysian SA researchers make neither their unique test data nor their code public, precluding a straightforward comparison. This notwithstanding, we will make an estimate of comparative performance at the end of this paper.

1.1. Problem Statement

To summarise the research problems addressed by this work:

1. The bilingual problem of Malay social media (i.e. the mixture of Malay and English) has been identified as a difficulty in earlier work [6][14][15] but has not yet been analysed and quantified, or explicitly addressed in a SA system.
2. Despite the potential shown by earlier translation work [23][24][25[26], VADER has not yet been translated to the Malay language.
3. No Malaysian SA work has used a large, broad domain, 3 class social media dataset *which is publicly-available* [14][15][16][17] making comparison between work difficult.

1.2. Research Questions

To summarise the research questions of this work:

1. What is the actual proportion of English content in informal Malaysian text? For SA, is it significant enough to be worth addressing explicitly?
2. How does English sentiment content relate to Malay sentiment content in the same text? Does it replace, match, or contradict Malay sentiment?
3. How can the mixture of languages be dealt with effectively for SA, or even exploited to improve performance?

2. BACKGROUND

2.1. Valence Aware Dictionary and sEntiment Reasoner (VADER)

VADER's [21] main component is its lexicon of 7,517 sentiment words and emoticons which were derived from established sentiment word banks and rated for valence by 10 pre-screened English speakers. In addition to its lexicon, which is common to most knowledge-based SA systems, VADER also takes into account a number of features that are particularly important in social media text. These include: exclamation and interrogation marks, which can modify sentiment intensity; capitalisation, which will increase the intensity of sentiment-laden words; negators, which switch the polarity of sentiment

words; boosters, which increase intensity; and other language-dependent rules. Given a sentence, VADER will attempt to find each word in the lexicon, extract its sentiment level, and then potentially modify this level based on the presence of exclamation marks and capitalisation. If the word is not found in the lexicon, it may still be identifed as a booster or a negator, which will affect the intensity and valence of other sentiment words in the sentence. The final outputs of the algorithm are individual ratings for positive, negative, neutral sentiments for the given sentence, and a single compound rating which is their weighted sum. When tested on a variety of text types, VADER was able to achieve an F1 score of 0.96 on social media, which outperformed popular Machine Learning algorithms (Naïve Bayes and SVM) and also human raters when compared to prevalidated ground truth.

2.2 Existing VADER Translations

VADER's high performance has led to recent attempts to adapt it to other languages. Swedish VADER used automatic translation of the VADER lexicon to Swedish [23]. German [24], Bengali [25], and Assamese [22] VADER repeated the entire VADER methodology of lexicon construction using human subjects. None of these attempts have yielded results comparable to English VADER but the work is encouraging and provides proof of principle that VADER could be adapted to the Malay language.

3. METHOD

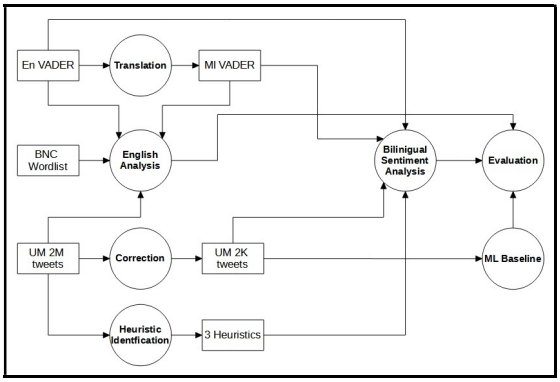

Figure 1: Bilingual Sentiment Analysis System Overview

In order to address the research problems and questions listed in 1.1 and 1.2, our overall method, as shown in figure 1, will be as follows:

1. **Translate English VADER to Malay** (research problem 2, question 3)
2. **Acquire and correct publicly-available Malaysian social media dataset** (research problem 3, question 3)
3. **Analysis of English Sentiment Content in Malaysian Social Media** (research problem 1, questions 2 and 3)
4. **Identification of Abnormal Text Patterns and Normalisation Heuristics** (research question 3)
5. **Machine Learning Performance Baseline** (research question 3)
6. **Bilingual Sentiment Analysis** (research problem 2, question 3)

During the initial preparation stage, English VADER is translated to Malay and a representative sample of the Malaya SM dataset is corrected and extended to 3 class labels. Tnen using the translated VADER lexicon, analysis of the English content of Malaysian SM will be conducted which will identify the potential for performance improvement for a bilingual system. The SM data will be analysed for patterns of abnormal text that can be corrected by normalisation heuristics. A performance baseline from Machine Learning SA algorithms will be obtained from training and testing on the corrected Malaya data. English and Malay VADER will be combined into a bilingual SA system and the normalisation heuristics will be tested for effectiveness. Lastly, the performance of the various combinations will be asessed and compared to the ML baseline. Considerable custom Python code was developed to integrate the various components and processes.

3.1 VADER Translation to Malay

The original English VADER lexicon consists of 7,517 terms which must be translated to Malay (minus the 450 which are emoticons and common short forms). The nltk implementation of VADER was used in this research [28]. In addition to the main lexicon, VADER's lists of English boosters and negators must also be translated. The process is not trivial because Malay and English differ in a number of significant ways. A clear difference is their relative vocabulary size: there are many different words in English that have the same translation in Malay.

For objectivity, speed, and efficiency, the translation process was intended to be done by machine. However, the desire for objectivity had to be balanced by the current imperfect capabilities of

machine translation. It was decided that automatic translation would be tried first and if the results were not considered satisfactory upon inspection then manual translation would be carried out.

For automatic translation, the Google translate API (using the freely-available PyTrans wrapper [29]) was used. For manual translation, a native Malay speaker with university-level English proficiency was available and 3 other native Malay speakers were available to validate the translation. Although this would be time-consuming, it was expected that the results would be more natural and accurate than Google Translate.

3.2 Malaya Social Media Dataset correction

Malaya is a Natural-Language-Toolkit library for Bahasa Malaysia, powered by Deep Learning (Tensorflow) [27]. It consists of a comprehensive set of Python modules for all aspects of NLP and ML in Malay is the only publicly available library of its kind for that language. Malaya has currently been downloaded 372k times and has an active community on Discord. The Malaya Twitter dataset consists of 2M general purpose tweets extracted from Twitter in Malaysia. Each is labelled with positive and negative valence and there are 1M positive, 1M negative tweets. The 2M tweets comprise a total of 27,546,111 words. Although this data is large and broad domain, there are no neutral labels. Also, inspection reveals that a very small fraction of the labels seem incorrect.

In order to ensure the most reliable testing for our bilingual SA system, the valence labels for a subset of the Malaya dataset were manually inspected and corrected. Correction means to change any positive to negative (and vice versa) and also to identify any neutral that have been labelled as positive or negative. A random sample of 2000 tweets was extracted for this correction work and the distribution was validated against the full dataset to demonstrate that the sample was representative of the population.

3.3 Analysis of English Sentiment Content in Malaysian Social Media

The purpose of this analysis was to estimate both the overall and sentiment-specific English content of Malaysian social media and to determine how it relates to Malay sentiment. To achieve statstical significance, the full Malaya dataset of 2M positive and negative tweets will be analysed.

For objectivity, it was intended to perform the analysis with as little manual intervention as possible. However, automatically identifying English words is not always straightforward. Refering to a dictionary is not always correct since there are certain Malay words that exist in the English dictionary with different meanings but are virtually unused in the English language. For example, the Malay word "yang" does exist in the English dictionary but cannot be considered truly English. Therefore, we used word frequency in use as a criterion for identifying English words. The British National Corpus (BNC) is a large general-purpose English corpus which freely provides a list of 6,985 words with frequency >= 10 per million in spoken and written English [30]. This should contain "normal" English words rather than those that only exist in a dictionary.

The BNC list contains both neutral and sentiment-laden words so it is not useful to estimate specifically sentiment content. However, the English and translated Malay VADER lexicons only contain sentiment words will be used here. The level of English sentiment in Malaysian SM will provide the potential for performance increase by successful bilingual processing.

3.4 Identification of Abnormal Text Patterns and Normalisation Heuristics

Inspection of tweets in the Malaya dataset was carried out in order to identify commonly-occurring patterns in Malaysian SM. Estimates of the frequency of these abnormal patterns were also made.

3.5 Machine Learning Performance Baseline

As stated earlier, VADER is a lexicon-based SA system which has certain advantages over ML systems. However, ML SA systems have demonstrated very high performance in certain domains and a currently very popular. To give context to the performance our system achieved, three popular ML algorithms: SVM, Naive Bayes, and kNN classifier were trained and tested on the same Malaya data as our bilingual lexicon based system. The freely-available Python SkLearn library was used for preprocessing, classifier, and metric functions [31]. The 2000 corrected tweets were divided randomly into 60% training and 40% testing sets of sizes 1200 and 800 respectively. Preprocessing was carried out as follows: stopword removal, numerical and special character removal,

conversion to lower case, TfIDF vectorisation, chi-squared feature reduction to 1000 features.

3.6 Bilingual Sentiment Analysis

The purpose of bilingual sentiment analysis is to combine English and Malay VADER in order to cover more sentiment content in Malaysian SM. This combination can be done in a number of ways. For the purposes of this current research, we explored two methods. In Winner-Takes-All (WTA) mode, each tweet is processed by both Malay and English VADER and the strongest predicted valence is taken as the winner. In English Override mode, all tweets are processed by Malay VADER and only if English sentiment words are detected in the tweet will English VADER be used. These two bilingual modes were also combined with the normalisation heuristics identified earlier. The results of bilingual SA will be compared to the performance of Malay VADER alone in order to demonstrate the importance of English sentiment content in Malaysian social media.

4. RESULTS

4.1 VADER Translation to Malay

The Google Translate API was used to translate the English VADER lexicon. Inspection revealed serious problems with the results. Some representative examples of problematic translations are as follows. Some translations were simply wrong, as seen below.

Table 1: Examples of Google mistranslations and manual corrections

Valence	English	Google	Manual
2.1	admire	pengembara	meminati
2.6	adore	mesra	memuja rasa hormat dan sayang yang dalam
-1.8	agony	dengan tidak mempedulikan	penderitaan
1.5	agree	apatis	bersetuju
-2.7	anger	penjamin	kemarahan

But the biggest problem was that Google conflated many English terms to a single Malay term which lost important distinctions. As expected, many different English words had the same Malay translation. In total, only 4,719 different Malay words were found as equivalents to the 7,517 English words.

Table 2: Examples of Google conflations and manual corrections

Valence	English	Google	Manual
-1.2	avoid	berkabung	mengelak
-1.7	avoidance	berkabung	elakkan
-1.1	avoidances	berkabung	elakkanelakkan
-1.4	avoided	berkabung	dihindari
-2	irritating	orang yang kalah	menjengkelkan
-2	irritatingly	orang yang kalah	secara menjengkelkan

Because of these problems, manual translation was considered necessary. This manual translation was able to achieve subtler distinctions than Google and yielded 5,195 different terms for the 7,517 English words. Examples of corrected mistranlations and conflations are also shown above.

4.2 Malaya Social Media Dataset correction

A random sample of 2000 tweets was extracted from the 2M tweet Malaya dataset for this correction work. As can be seen for the top 20 terms below, we can be confident that this sample is representative of the full dataset because its lexical distribution is very close to the original.

Journal of Theoretical and Applied Information Technology
30th June 2023. Vol.101. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: Top 20 term frequencies in 2M Malaya dataset vs 2k random sample

Full Dataset	Random Sample
('aku', 357230), ('yang', 271219), ('di', 256077), ('dan', 216130), ('yg', 202570), ('nak', 166359), ('ni', 165600), ('ada', 158764), ('tak', 154941), ('dia', 125649), ('ke', 115283), ('ini', 109117), ('dah', 103527), ('tapi', 96415), ('tu', 95971), ('dengan', 95484), ('buat', 95121), ('untuk', 91950), ('orang', 91871), ('dari', 89465),	('aku', 322), ('di', 287), ('yang', 228), ('yg', 197), ('dan', 184), ('nak', 157), ('ada', 146), ('tak', 145), ('ke', 144), ('ni', 134), ('dia', 131), ('ini', 112), ('dah', 112), ('-', 107), ('dari', 106), ('tu', 100), ('buat', 97), ('tapi', 96), ('.', 92), ('dengan', 90),

In total, 237 (12%) tweet labels were corrected. 160 (8%) became neutral, 40 (2%) negatives became positive, and 37 (2%) positives became negative. This indicates that overall, the original Malaya labels are accurate enough to not invalidate the English analysis performed next, but with correction should give more a more accurate evaluation of Sentiment Analysis performance during testing.

4.3 Analysis of English Sentiment Content in Malaysian Social Media

Using the BNC word frequency list to identify genuine English words in the 2M tweet Malaya dataset, 2,518,340 instances of 6,282 different English words were found. The 100 most frequent English words and their frequencies within the dataset are shown below.

[('la', 72280), ('ya', 61484), ('in', 53901), ('i', 47429), ('and', 30382), ('at', 29245), ('the', 27774), ('2', 26819), ('air', 23083), ('1', 21892), ('main', 21315), ('a', 20912), ('best', 20129), ('to', 19762), ('so', 19544), ('3', 18053), ('time', 17281), ('you', 16746), ('my', 16295), ('eh', 16207), ('dr', 15417), ('for', 14759), ('d', 14190), ('happy', 13804), ('me', 13495), ('like', 13176), ('of', 12965), ('is', 12655), ('4', 11361), ('m', 11166), ('5', 11064), ('video', 10756), ('ah', 10622), ('s', 10242), ('free', 9925), ('love', 9705), ('p', 9383), ('no', 9144), ('on', 9006), ('10', 8831), ('or', 8674), ('good', 8553), ('ye', 8342), ('be', 8169), ('shah', 8151), ('this', 8075), ('but', 7847), ('media', 7755), ('okay', 7446), ('ok', 7330), ('member', 7314), ('ready', 7054), ('with', 6866), ('follow', 6670), ('then', 6546), ('last', 6381), ('do', 6376), ('up', 6352), ('it', 6153), ('km', 6150), ('stress', 5906), ('thanks', 5821), ('all', 5809), ('first', 5801), ('motor', 5780), ('order', 5719), ('by', 5707), ('area', 5686), ('via', 5674), ('7', 5563), ('as', 5528), ('full', 5488), ('6', 5488), ('still', 5448), ('just', 5436), ('super', 5305), ('program', 5248), ('ph', 5240), ('try', 5177), ('one', 5146), ('birthday', 5142), ('oh', 5063), ('die', 5025), ('start', 5011), ('hp', 5010), ('open', 4969), ('your', 4920), ('ha', 4822), ('bank', 4817), ('from', 4809), ('day', 4679), ('please', 4484), ('call', 4409), ('baby', 4403), ('game', 4376), ('hotel', 4375), ('family', 4329), ('support', 4226), ('april', 4217), ('12', 4141)]

Figure 2: Top 100 English words in 2M tweet Malaya dataset

Inspection reveals that the BNC frequency method is not foolproof. Although the words {la, ya, air} are on the BNC list, it is reasonable to assume that in this context they have specifically Malaysian meanings (ie "air" really means the Malay word for "water"). So, to be conservative we removed these words from the analysis to leave 2,361,493 instances of English words. This means that at a per-word level, 8.6% of all words in Malaysian social media are English, which is clearly non-negligible. It can be seen that although most of these English words are neutral or functional words, there is still evidence of English sentiment content even at this high level. The sentiment-laden words {best, happy, like, ah, free, love, good, okay, ok, ready, stress, thanks, first, super, birthday, oh, die, ha, please, baby, support} all have significant frequencies.

However, it is essential to note that for SA on SM the correct unit of analysis is tweets and not words since even one English word per tweet could be enough to determine its sentiment. It was found that 1,054,114 = 53% of 2M tweets contain at least one English word, with many containing up to 6. This can be considered a significant level of overall English content in Malaysian social media. The next step is to determine how much of this English content is sentiment-laden and therefore relevant to Sentiment Analysis.

Given two possible languages (Malay and English) and two possible states (present or absent), there are four possibilities for sentiment content in the 2M Malaysian tweets. It was found that 131,823 (6.59%) tweets contained both English and Malay sentiment words, 137,099 (6.85%) contained only

English sentiment, 1,052,010 (52.60%) contained only Malay sentiment words, and 679,068 (33.95%) contained neither Malay nor English sentiment.

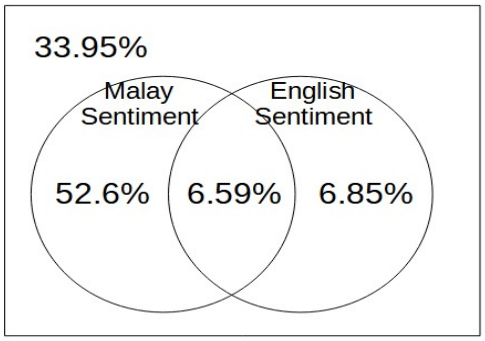

Figure 3: Venn diagram of English and Malay sentiment in 2M tweet Malaya dataset

Therefore, a total of 13.44% tweets contained English sentiment, which can be considered to be a significant level which requires addressing explicitly in Malaysian Sentiment Analysis. The 6.85% tweets that contain only English sentiment would fail automatically using only a Malay lexicon and therefore constitute the potential for improvement with a bilingual approach if English is handled correctly.

The 6.59% tweets that contain both English and Malay sentiment offer the potential to confuse SA processing if the valence in both languages conflict. Therefore, this class of tweets was analysed to determine the level of agreement in bilingual sentiment. The confusion matrix below summarises this analysis.

Table 4: Malay vs English valence within same tweet

	English neg	English pos
Malay neg	12%	12%
Malay pos	19%	57%

Although bilingual sentiment agrees in 69% (12%+57%) of tweets, 31% (12%+19%) tweets have conflicting valence. This amounts to around 2% of all tweets, which is small but non-negligible. In these cases, should Malay or English be chosen in conflicts? Exploring this issue is the purpose of the "Winner-takes-all" and "English Override" modes of bilingual SA in section 4.6.

Finally, the 34% of tweets without any detected sentiment words may either actually have sentiment words that were missed for reasons such as poor spelling and lexicon limitations, or they may genuinely be neutrals despite labelling by Malaya as positive or negative. We suspect it is mainly the former case since our manual dataset correction in section 4.2 yielded only 8% neutral tweets.

4.4 Identification of Abnormal Text Patterns and Normalisation Heuristics

Malaya tweet data was inspected to identify prominent patterns of abnormal text unique to Malaysian SM. Two patterns were identified and corresponding heuristics were developed to normalise them.

4.4.1 Joint Words Heuristic

Malay contains many multi-word units such as "peminat-peminat" and "sikap-jalang" which are separated by hyphens. However, in social media these crucial hyphens are often omitted. We estimate this to be the case in 4.4% of the tweets analysed. This is important because words taken individually may have very different meanings and valence to that when combined. For example, the multi-word unit “acuh-tak-acuh” means "indifferent" which is either neutral or mildly negative, but the individual word "acuh" has a positive valence. Since VADER processes sentences token by token and its lexicon does not explictly handle multiple tokens, VADER would process “acuh tak acuh” as three separate words with valences {positive, negative, positive} instead of a single unit with neutral or mildly negative valence.

The Joint Words heuristic attempts to replace missing hyphens to increase the chances of VADER recognising multi-word units. The translated Malay VADER lexicon contains multi-word units of between 2 to 4 individual words separated by hyphens. Each multiword unit is inserted into a search tree of depth 4, with each level corresponding to its position in the multi-word unit. In other words, all possible first words in a multi-word unit are at level 1, all possible second words are at level 2, and so on. When processing a tweet consisting of N tokens, each possible set of 4 contiguous tokens is input to the search tree and if a match of n tokens can be found, then those n individual tokens are replaced by a single hyphenated token.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.4.2 Repetitive Letter Heuristic

It is common in Malay social media for letters in a word to be repeated for emphasis. For example,

- mai pekena air coconut shake vanilla di pasar ramadhan keramat. **sedappppppppp**..... dato' keramat lrt station
- torres menang ucl dengan europa kot, **powerrrrrrr**

This repetition often occurs at the end of the word but can also occur internally, such as in "goooooood". We estimate this to occur in 5% of tweets. Clearly, these abnormal word forms would not be found in any lexicon and so would be classed as neutral words even though they actually convey emphasis and sentiment. The Repetitive Letter heuristic attempts to detect and normalise this situation. If a word is found to have repeated letters, the repeated section is reduced by one character until it is empty or the shortened word is found in either the Malay or English VADER lexicons. Note that we are only concerned with normalising sentiment-laden words (ie those in VADER) since any others would be neutral anyway.

4.4.3 Normalisation by Pattern Matching

The preceding heuristics were targetted to the specific characteristics of Malaysian social media text but it was also considered worthwhile to try a more general approach to normalisation. The extended version of the Gestalt Pattern Matching algorithm by Ratcliff and Obershelp found in the Python difflib library [32] was used to find closest matches between unrecognised words and the Malay and English VADER lexicons. Matches with confidence >85% were accepted and used to replace the unrecognised word.

4.5 Machine Learning Performance Baseline

Using the SkLearn library and the corrected 2000 tweets split 60/40 for training and testing, the following results were obtained for the three polarity values negative, neutral, positive {-, 0, +}.

Table 5: SA performance by popular Machine Learning algorithms

SVM				
Confusion Matrix		-	0	+
	-	269	0	100
	0	27	4	53
	+	85	0	262
Accuracy	0.67			
F1 score	0.64			
Naive Bayes				
Confusion Matrix		-	0	+
	-	304	0	65
	0	36	4	48
	+	125	0	222
Accuracy	0.66			
F1 score	0.62			
kNN				
Confusion Matrix		-	0	+
	-	36	0	333
	0	0	0	84
	+	8	0	339
Accuracy	0.47			
F1 score	0.35			

For a 3-class problem, the overall results are moderately good compared to chance (ie 33%). However, it is clear that all three classifiers could not identify neutral tweets and that their output became essentially two class (positive and negative). This may partially be the result of the uneven number of examples in each class. But, since neutral polarity is the result of what is missing from the tweet (ie sentiment-laden words) rather than what is present, it is likely due to the inherent difficulty of learning this problem from examples [4]. It may also be why so many other researchers restrict themselves to two classes. Overall, the SVM classifier gives the best performance of the three algorithms, with accuracy and F1 of 0.67 and 0.64, respectively. These results provide one important baseline to assess the performance of the Bilingual VADER system which will be described presently.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.6 Bilingual Sentiment Analysis

Another important baseline is to consider the performance of VADER using only either English or Malay lexicon alone (table 6).

Table 6: SA performance by English and Malay VADER alone

English VADER				
Confusion Matrix		-	0	+
	-	86	798	54
	0	5	146	9
	+	33	656	213
Accuracy	0.22			
F1 score	0.25			
Malay VADER				
Confusion Matrix		-	0	+
	-	600	165	172
	0	25	96	40
	+	62	247	593
Accuracy	0.64			
F1 score	0.68			

This will allow the proper evaluation of a bilingual approach. We know from the analysis performed earlier that neither Malay nor English alone cover the full scope of sentiment-laden words in Malaysian tweets. And since lexicon-based approaches treat words not found in the lexicon as neutral, this was expected to affect performance in proportion to lexicon coverage. The results for VADER in each language are shown in table 6 above.

As might be expected, English VADER was unable to recognise the predominantly Malay content of tweets and therefore classified most of them as neutral overall. However, 15% (86+213) of the non-neutral tweets were still correctly classified based only on the English sentiment, which tallies well with the estimate made earlier that 13.5% of Malaysian tweets contain English sentiment words. The overall 22% accuracy for English VADER is a combination of this genuine performance and the performance "by accident" on the neutral tweets in the dataset.

In contrast, Malay VADER achieved 64% accuracy overall which is a respectable level for a 3-class problem and is also comparable to the ML baseline. Malay VADER was able to genuinely distinguish positive and negative tweets from neutral and to do so equally well for either polarity. Of the 64% accuracy achieved, 4% was from neutral tweets and 60% (600+593) was on tweets with positive and negative sentiment content. This 60% very closely matches the level of Malay sentiment in the dataset estimated earlier (ie 59%). Importantly for the main research question here, this score of 64% accuracy provides a baseline to measure the improvements (if any) of the bilingual approach.

4.6.2 Bilingual VADER

As stated in 3.6, the English and Malay lexicons were combined in two ways: Winner-takes-all and English override and with results shown in table 7. In Winner-Takes-All (WTA) mode, each tweet was processed by both Malay and English VADER and the strongest predicted valence was taken as the winner. In English Override mode, all tweets were processed by Malay VADER and only if English sentiment words are detected in the tweet will English VADER override the Malay prediction.

Table 7: SA performance by English and Malay VADER combined

English Override				
Confusion Matrix		-	0	+
	-	451	297	190
	0	15	101	44
	+	43	222	637
Accuracy	0.59			
F1 score	0.64			
WTA				
Confusion Matrix		-	0	+
	-	614	136	187
	0	26	88	47
	+	62	156	684
Accuracy	0.69			
F1 score	0.72			

English override mode attempted to explore the question noted earlier of which language should win when a conflict between English and Malay valence exists. Compared to Malay VADER, it improved performance on positive tweets but decreased the performance on negative tweets yielding a poorer overall level of performance than Malay VADER. The reasons for this are still under analysis.

Like English override, the Bilingual system in WTA mode improved accuracy on positive tweets but did not decrease accuracy on negative tweets, therefore outperforming both the ML baseline and Malay VADER alone. In the English content analysis performed earlier, it was estimated that the potential for improvement for Bilingual VADER over Malay VADER was 6.85%. The improvement of 5% for WTA found here matches that estimate very well and can be considered a successful implementation of bilingual SA in a Malaysian context.

4.6.3 Bilingual VADER with Normalisation Heuristics

Since Bilingual VADER using WTA performed the best, it was then combined with the three normalisation heuristics described in 4.4 to produce the results shown below.

Table 8: SA performance by Bilingual VADER with normalisation

Bridging				
Confusion Matrix		-	0	+
	-	614	134	189
	0	26	88	47
	+	60	156	686
Accuracy	0.69			
F1 Weighted avg	0.72			
Repetitive				
Confusion Matrix		-	0	+
	-	612	135	190
	0	26	88	47
	+	60	152	690
Accuracy	0.69			
F1 Weighted avg	0.72			
Pattern Matching				
Confusion Matrix		-	0	+
	-	629	103	205
	0	34	71	56
	+	75	109	718
Accuracy	0.71			
F1 Weighted avg	0.72			

The repetitive heuristic was able to successfully normalise a number of noisy tokens in the tweets. Some examples are shown below.

Table 9: Examples of successful normalisation by Repetitive Heuristic

Original	Normalised
Puihhh	puih
Heeeeey	hey
hhhhheeeeeyyyy	hey
sedapppppppp.....	sedap.
takdeee.	takde.

Disappointingly, the bridging and repetitive heuristics yielded no discernable improvement on overall SA performance. However, the Pattern Matching heuristic was able to increase overall accuracy by 2%, giving the best performance overall among all the methods tried here.

5. CONCLUSION

Our analysis of English content in Malaysian informal text confirms existing observations about language mixing but for the first time quantifies them. We were able to show a significant level of English content in Malaysian social media and also that this content was sentiment-laden and relevant to Sentiment Analysis in a Malaysian context. Our attempts to handle this mixed language involved considerable work on translating English VADER to Malay and combining the two systems. We will conclude by returning to the three research questions listed in 1.2.

1. A total of 13.44% tweets were found to contain English sentiment, which can be considered to

be a significant level that requires addressing explicitly in Malaysian Sentiment Analysis. Our analysis estimated that up to 6.85% improvement in accuracy was possible for a bilingual system as compared to a purely Malay system. (sec 4.3)

2. Bilingual sentiment agrees in 69% of tweets, whereas 31% tweets have conflicting valence. (sec 4.3)
3. Bilingual performance was found to be superior to that of the pure Malay lexicon and the improvement (5%) was found to be comparable to the potential estimated by the English analysis conducted earlier. Normalisation somewhat improves performance (2%) although not as much as was expected. It was also shown that this bilingual system can outperform several popular machine learning based methods on the same data. (sec 4.4-4.6)

As stated in the introduction, achieving the best performance over all SA systems was not the scope of this exploratory work. It was also noted that directly comparing our performance with existing Malaysian SA work is very difficult due to the use of private datasets and algorithms. However, some comparison to current systems may still be worthwhile. For example, the 2023 non-Malaysian context SA work described in [33,34,35,36] can achieve > 80% accuracy using various machine learning approaches which exceeds our performance. However, the 2023 Malaysian SA work described in [37] achieves an F1 score of 0.72 on only two class sentiment data using a machine learning approach. This is the same F1 performance we have achieved on more difficult 3 class data using a bilingual lexicon approach. In absolute terms, our system's performance must be improved but the results obtained here are encouraging for the future continuation of this approach.

5.1 Future Work

The work presented here is only an initial exploration of bilingual SA in a Malaysian context. Future work will include correcting and testing on a far larger sample of the Malaya dataset, exploring other modes of combining English and Malay lexicons, refining the normalisation heuristics that were found to be ineffective earlier, and comparing VADER lexicon coverage to SentiWordNet. We also intend to compare directly to other Malaysian SA systems using the same dataset used here.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support for this work by Universiti Malaysia Sabah under Grant No. DN20090.

REFERENCES

[1] Oxford English Dictionary. (2023). "Sentiment analysis, n. 1." OED Online. Oxford University Press, Accessed March 2023.

[2] Wankhade M, Rao ACS, & Kulkarni C. (2022). "A survey on sentiment analysis methods, applications, and challenges." *Artif Intell Rev*. https://doi.org/10.1007/s10462-022-10144-1

[3] Qazi A, Raj RG, Hardaker G, Standing C, (2017) "A systematic literature review on opinion types and sentiment analysis techniques: Tasks and challenges", *Internet Research*, Vol. 27 Issue: 3, pp.608-630,

[4] Koppel M & Schler J. (2006). "The Importance of Neutral Examples for Learning Sentiment". *Computational Intelligence*. 22. 100-109. 10.1111/j.1467-8640.2006.00276.x.

[5] Ligthart A, Catal C, Tekinerdogan B. (2022). "Systematic reviews in sentiment analysis: a tertiary study". *Artif Intell Rev* 54:4997–5053

[6] Ahmad S, Asghar MZ, Alotaibi FM, Awan I. (2019). "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques". *Hum Centric Comput Inf Sci* 9(1):1–23

[7] Samsudin N, Puteh, M, Hamdan AR, Ahmad Nazri MZ. (2013). "Normalization of noisy texts in Malaysian online reviews". *Journal of Information and Communication Technology*. 12. 147-159. 1

[8] Cortis K, Davis, B. (2021). "Over a decade of social opinion mining: a systematic review". *Artificial Intelligence Review*. 54. 10.1007/s10462-021-10030-2.

[9] Medhat W, Hassan A, Korashy H. (2014). "Sentiment analysis algorithms and applications: A survey". *Ain Shams Engineering Journal*, Vol 5, Issue 4, pp. 1093--1113 (2014)

[10] Liu, B. *Sentiment Analysis and Opinion Mining*. Claypool Publishers. (2012)

[11] Khairnar J, Kinikar M. (2013). "Machine Learning Algorithms for Opinion Mining and Sentiment Classification," *Int. J. Sci. Res. Publ*., vol. 3, no. 6, pp. 1–6.

[12] Wankhade M, Annavarapu CSR, Verma MK. (2021). "CBVoSD: context based vectors over

sentiment domain ensemble model for review classification". *J Supercomput* 1–37

[13] Kumar KN, Uma V (2021). "Intelligent sentinet-based lexicon for context-aware sentiment analysis: optimized neural network for sentiment classification on social media". *J Supercomput* 77:12801–12825

[14] Chekima, K, Alfred, R. (2018). "Non-English Sentiment Dictionary Construction". *Advanced Science Letters*. 24. 1416-1420. 10.1166/asl.2018.10761.

[15] Chekima, K, Alfred, R. (2018). "Sentiment Analysis of Malay Social Media Text". In: Alfred, R., Iida, H., Ag. Ibrahim, A., Lim, Y. (eds) Computational Science and Technology. ICCST 2017. Lecture Notes in Electrical Engineering, vol 488. Springer, Singapore. https://doi.org/10.1007/978-981-10-8276-4_20

[16] Bakar, M, 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.

[17] Handayani D, Awang Abu Bakar NS, Yaacob H, Abuzaraida MA. (2018). "Sentiment Analysis for Malay Language: Systematic Literature Review," 2018 *International Conference on Information and Communication Technology for the Muslim World (ICT4M),* 2018, pp. 305-310, doi: 10.1109/ICT4M.2018.00063.

[18] Omar AH. (2004). "Encyclopedia of Malaysia Vol 9*:* Languages and Literature". Editions Didier Millet ISBN 981-3018-52-6.

[19] Wikipedia (2023). https://en.wikipedia.org/wiki/Bahasa_Rojak. Accessed March 2023.

[20] Baccianella S, Esuli A, Sebastiani F. (2010). "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

[21] Hutto C, Gilbert E. (2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216-225.

[22] Borg A, Boldt M. (2020). "Using VADER sentiment and SVM for predicting customer response sentiment". *Expert Systems with Applications*. 162.

[23] Gustafsson M, (2019). "Sentiment Analysis for Tweets in Swedish – Using a Sentiment Lexicon with Syntactic Rules". Bachelor's thesis. http://www.diva-portal.org/smash/get/diva2:1391359/FULLTEXT01.pdf. Accessed March 2023.

[24] Tymann K, Lutz M, Palsbroker P, Gips C. (2019). "GerVADER - A German adaptation of the VADER sentiment analysis tool for social media texts", *Conference at Humboldt-University zu Berlin*, Vol: 2454, No: 14.

[25] Amin A, Hossain I, Akther A, Alam KM. (2019) "Bengali VADER: A Sentiment Analysis Approach Using Modified VADER," *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1-6, doi: 10.1109/ECACE.2019.8679144.

[26] C. Dev, A. Ganguly and H. Borkakoty, "Assamese VADER: A Sentiment Analysis Approach Using Modified VADER," 2021 *International Conference on Intelligent Technologies (CONIT), 2021,* pp. 1-5, doi: 10.1109/CONIT51480.2021.9498455.

[27] Zolkepli, H (2018). "Malaya, Natural-Language-Toolkit library for Bahasa Malaysia, powered by Deep Learning Tensorflow". *GitHub repository*. https://github.com/huseinzol05/malaya. Accessed March 2023.

[28] https://www.nltk.org/_modules/nltk/sentiment/vader.html. Accessed March 2023.

[29] https://pypi.org/project/pytrans/. Accessed March 2023.

[30] https://ucrel.lancs.ac.uk/bncfreq/flists.html. Accessed March 2023.

[31] https://scikit-learn.org/stable/. Accessed March 2023.

[32] https://docs.python.org/3/library/difflib.html. Accessed March 2023.

[33] Zou H., Wang Z, Zou, Haochen, Wang, Zitao. (2023). "A semi-supervised short text sentiment classification method based on improved Bert model from unlabelled data". *Journal of Big Data*, 10 (1), 35.

[34] Qaqish, E., Aranki, A. & Etaiwi, W. (2023) "Sentiment analysis and emotion detection of post-COVID educational Tweets: Jordan case". *Soc. Netw. Anal. Min*. 13, 39.

[35] AminiMotlagh, M., Shahhoseini, H. & Fatehi, N. (2023). "A reliable sentiment analysis for classification of tweets in social networks". *Soc. Netw. Anal. Min*. 13, 7.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[36] Qi, Y., Shabrina, Z. (2023) "Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach". *Soc. Netw. Anal. Min*. 13, 31

[37] Kong, J.T.H.; Juwono, F.H.; Ngu, I.Y.; Nugraha, I.G.D.; Maraden, Y.; Wong, W.K. (2023) "A Mixed Malay–English Language COVID-19 Twitter Dataset: A Sentiment Analysis". In *Big Data Cogn. Comput*. 2023, 7, 61.