Subjectivity Analysis of an Enhanced Feature Set for Code-Switching Text

Emaliana Kasmuri, Halizah Basiron*
Fakulti Teknologi Maklumat Dan Komunikasi, Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—The phenomenon of code-switching has posed a new challenge to the linguistic computing area. Conventionally, the computer will process monolingual text or multilingual text. However, code-switching is different from this kind of text. Two or more languages are used to construct a piece of code-switching text, particularly a code-switching sentence. It is challenging for the computer to process a piece of code-switching text with languages that exist simultaneously. The challenge is more intense for the computer in subjectivity analysis, where the computer should distinguish subjective from objective code-switching text. This paper proposed three feature sets for subjectivity analysis on Malay-English code-switching text: Embedded Code-Switching Feature Sets, Unified Code-Switching Feature Sets, and Stylistic Feature Sets. These feature sets were enhanced from the monolingual feature set of subjectivity analysis. Experiments were conducted using the data harvested from Malay-English blogs. These data were labelled as either subjective or objective. Two machine learning classifiers – the Support Vector Machine (SVM) and Naive-Bayes, were used to evaluate the classification performance of the proposed feature sets. The experiments were carried out on individual feature sets and the combination of them. The results show the classification performance from combining the unified and stylistic feature sets surpassed other proposed feature sets at 59% accuracy. Therefore, it is concluded that the combination of unified and stylistic feature sets is necessary for the subjectivity analysis of Malay-English code-switching text.**

Keywords*—*Subjectivity analysis; code-switching; enhanced feature sets; Malay-English text

I. Introduction

It is common for a person to master multiple languages. Mastering multiple languages benefitted an individual in various ways, including vocabulary enrichment, communication improvement, and knowledge improvement through information exchange and sharing. It is also a common situation to see a multi-lingual person communicate, either in spoken or written, using a mix of languages. The use of mixed languages is known as code-switching [1].

A code-switching text is a piece of text that is constructed using words from at least two languages. For example, 'I really like matte satin silk nak buat collection tapi I am still a student tak ada income if dapat giveaway ni mesti best'. In the example, the sentence is constructed using English (underlined) and Malay (italic) words. Code-switching may occur within a text where the words from the second language interleave in between the words from the first language, or the first part of the text was constructed using the first language while the remaining part was constructed with the second language. For example, 'Aspirasi saya adalah banyak peluang dari segi pendidikan, ekonomi, sukan dan sebagainya untuk masyarakat khas for them to excel in those field'.

The code-switching phenomenon is not new in the linguistic research domain. It was acknowledged in the 1980s [2]. However, with the advancement of computer technology, the use of code-switching in open platforms such as blogs and social media has become more apparent than before. This situation has posed a new challenge to language computational areas in text analysis and machine comprehension of this new kind of text. In general, most of the research effort has been channelled towards computational and knowledge augmentation on monolingual and multilingual text. The presence of two languages simultaneously in a code-switching text was not considered in the current subjectivity analysis study.

A set of research questions are drawn to address the code-switching issues in subjectivity analysis. The research questions are as follows:

1) What are the most effective feature sets for subjectivity analysis in Malay-English code-switching text?

2) How do different machine learning classifiers perform in subjectivity classification when using the proposed feature sets?

3) What is the impact of combining various enhanced feature sets on the accuracy of subjectivity analysis in code-switching text?

Three objectives were drawn to align with the research questions. The first objective is to develop enhanced feature sets that effectively analyse subjectivity in Malay-English code-switching text. The second objective is, to compare the performance of machine learning classifiers using the proposed feature sets for subjectivity analysis. Finally, the last objective is to assess the accuracy improvement achieved by combining different feature sets for subjectivity classification in code-switched text.

Therefore, this article proposes a method that enhances feature sets to analyse subjectivity in Malay-English code-switching sentences to achieve the objectives. The enhanced feature sets used the subjectivity feature from both languages to represent the subjectivity of the code-switching feature sets. The enhanced feature sets consist of embedded, unified, and stylistic feature sets. Two machine learning classifiers, Naïve-Bayes (NB) and Support Vector Machine (SVM), were used to evaluate the feasibility of the feature sets classifying the Malay-English code-switching sentences into subjective and objective classes.

*Corresponding Author, halizah@utem.edu.my

This research contributes to the growing field of computational linguistics by addressing the challenge of subjectivity analysis in code-switching text. The findings from this study can be applied to improve text processing systems used in social media, blogs, and other platforms where code-switching is prevalent. The proposed enhanced features set that contains embedded, unified, and stylistic feature sets provide a foundation for future research and development in computational methods for analysing subjectivity in code-switching text.

The article's structure is as follows: In Section II, we review related works and discuss research motivation. Section III presents a description of our proposed solution and the features used. Section IV describes the dataset used in the experiment. Section V gives the details of our experiments and discusses the results. Sections VI and VII present the discussion and limitations, respectively. Lastly, Section VIII concludes and summarizes our work in this article.

II. Related Work

Subjectivity analysis is a linguistic computational task that determines the existence of a private state such as opinion, emotion and stance in a piece of textual document. Subjectivity analysis is a precursor to tasks such as sentiment analysis, and emotion classification.

Ting et al. investigated subjectivity classification using a window-based self-attention approach [2]. The approach improves the sentence encoding by leveraging context within variable-sized windows, instead of the entire sentence. Different sizes of windows (the size of surrounding words) were used to determine the importance of each word using trigrams or five-grams. The max-pooling method was used to extract the most relevant feature. As a result, multiple windows were captured from various phases where meaningful combinations of words were evident. This method simplifies attention computation, working directly on word embeddings and relying on n-gram features to provide context-aware sentence representations. The author used monolingual Cornell's movie data set to classify the review into subjective and objective classes. The proposed method performed at 94.60% accuracy.

Belisario et al. studied subjectivity analysis for Portuguese book reviews [3]. The authors used 350 reviews that were equally divided into subjective and objective sentences. The features of these sentences were extracted and represented using several methods including Sentilex-PT and WordnetAffectBR for the lexicon-based method, global centrality graph-based method using Eigenvector Centrality, Katz Index and PageRank and Naïve-Bayes, SVM and Neural Network for machine learning based method. The result reveals the Neural Network outperformed other methods at 83.20% accuracy.

Kasnesis et al. compare three transformer-based architectures to classify the the 10,000 Cornell movie review sentences into subjective and objective classes [3]. The authors used Bidirectional Encoder Representation (BERT), Robust Optimised BERT Pretraining Approach (RoBERTa) and Efficient Learning Encoder Classifies Token Replacements Accurately (ELECTRA) to evaluate the effectiveness of the subjectivity classification. ELECTRA achieved the highest accuracy score, 98.30% among the three.

Al Hamoud et al. analyse subjectivity expressed from an online English political and ideological debate forum [4]. Various controversial topics were debated in the forum including abortion, creationism, gay rights, the existence of God, gun rights and health care. The authors have created a dataset of 53, 453 sentences from the forum. These sentences were labelled as either subjective or objective. The features of the dataset were represented using one-hot encoding and GloVe pretraining embedding vectors. These features were experimented with using six deep learning models to find the most effective model that could classify the dataset into subjective/objective classes. The models are Long Short Term Memory Networks (LSTM), Gated Recurring Units (GRU), bidirectional GRU, bidirectional LSTM, LSTM with attention and bidirectional LSTM with attention. The results of the experiment show that LSTM with attention performed the best out of all deep learning models – 97.39% accuracy.

The studies that were presented above were working on monolingual textual documents – that is English and Portuguese. The excellent performance result in studies has shown the proposed classification solutions have generated effective classification models that has outperformed the other models in the respective experiment. Even though the studies can produce effective subjectivity classification models, their feasibility on code-switching text is still unknown. Therefore, this article attempts to fill the gap using Malay-English code-switching text for subjectivity analysis.

The issue of code-switching in linguistic computation has been addressed as early as the 1980s [5]. However, the issue received proper attention from the linguistic computation area due to the availability of the data sets and the maturity of computer language processing tools. With the advance of social media, the research of linguistic computation using code-switching text has started to receive attention. Among the research on code-switching computational are text generation [6], [7], [8] and sentiment analysis [9], [10], [11].

One of the hurdles in code-switching linguistic computation is data scarcity. The advancement of social media has accelerated the phenomenon of code-switching. Therefore, data become more available in this research area. However, the amount of data is insufficient for various tasks in linguistic computation. Code-switching data is artificially generated to support these tasks. Hu et al. generate code-switching training text to improve code-switching automatic speech recognition (ASR) [8]. The text injection method known as PaLM 2 and prompt tuning were used to generate the training data. The experiment to generate the data was carried out using the Large Language Model (LLM). The result shows that the proposed method achieves a 3.60% Word Error Rate (WER) for Mandarin-English. The research concludes leveraging LLMs for text generation and text injection benefits code-switching ASR tasks.

Three methods were compared to select the best method that generates code-switching data for Egyptian Arabic-English Hamed has been compared [7]. The first method is lexical replacement, which replaces a number of random Arabic words

with English words using code-switching point assignment and point prediction. Code-Mixing Index (CMI) was used to measure the performance of this method. CMI indicate the level of mixing from both languages. Both methods achieve 28.00% and 25.00% CMI. The second method used linguistic theories - Equivalence Constraint (EC) Theory and Matrix Language Frame (MLF) Theory. Both achieved 30.00% and 25.00% CMI. The final method used back translation. The method performed at 18.00% CMI.

Research by Chi et al. also addressed code-switching data scarcity [6]. Monolingual BERT models (Mandarin and English) were used to train a Transformer encoder-decoder model to translate between any pair of languages. The translated sentences are forced to code-switch to any degree. Disjoint union of vocabulary from two languages were used as parallel text. Grid beam search is the method to force the degree of the code-switching text. The proposed method performed at 30.58% WER.

A deep convolutional neural network (CNN) was used to determine the sentiment expressed in a comment that contains a mixture of English and Hindi languages [9]. A dataset that contains 17,155 comments was created, and the comments were extracted from a governmental website. Each comment was manually labelled as either positive or negative. The study used Word2Vec to represent the English words. The Hindi words embedding were developed using Word2Vec and trained using Hindi Wikipedia and other sources. The solution used on the dataset has yielded 67.00% of accuracy.

A study on a mixture of English-Spanish that leverages existing English and Spanish text analytical tools has been carried out using a supervised learning algorithm [10]. In this study, 3,062 tweets that were labelled as positive and negative were used. Four atomic features, which are word, lemma, psychometric properties and part-of-speech (POS) tags were used as basic features. The best accuracy obtained is 59.34% using combination lemmas and psychometric properties as features.

A recent study of sentiment analysis using a mixture of English-Hindi and English-Bengali tweets has demonstrated using deep learning [11]. The study classifies the dataset into positive, negative and neutral. Three deep learning algorithms, which are BiDirectional Long Short-Term Memory (BiLSTM) Convolutional Neural Network (CNN), a Double BiDirectional Long Short-Term Memory (D-BiLSTM) and an Attention-based model, were used in the experiment. Pre-trained word embeddings, which are Global Vector (GloVe) and Bidirectional Encoder Representations from Transformers (BERT), were used in the study. The study has revealed that accuracy performance between 42.00% and 68.00% was achieved using a 10-fold cross-validation classification. The result from the experiment shows the best performance for the English-Hindi dataset was achieved by the deep learning attention model using GloVe with 68.00% accuracy, whereas, for the English-Bengali dataset, the best performance was achieved by the deep learning attention model with 67.00% accuracy.

All the research efforts described in this section focus on code-switching sentiment classification and text generation. The attempt for subjectivity classification on code-switching text, especially for subjectivity analysis on Malay-English code-switching text, was found to be limited at the time this research was conducted. Therefore, this article is attempts to find out a feasibile solution for subjectivity analysis on Malay-English code-switching text.

III. Enhanced Feature Sets of Subjectivity Analysis for Code-Switching Text

The syntactical and semantical features were used to analyse subjectivity in a monolingual text such as English [12]. These features are described in Table I. These features were enhanced for a Malay-English code-switching text.

TABLE I. Initial Feature Set for Subjectivity Analysis

Syntactical Features	Semantical Features
1. Sentence that located in beginning of a paragraph. 2. The co-occurrence of words and punctuation.	1. Pronoun 2. Adjective 3. Cardinal number 4. Modal (other than will) 5. Adverb (other than not)

The enhanced feature sets are Embedded Code-Switching Feature Sets, Unified Code-Switching Feature Sets and Stylistic Feature Sets. The Embedded Code-Switching Feature Set include the Malay text analytical components in the code-switching feature set representation. The Unified Code-Switching Feature Set fused two text analytical components into a unified feature set. The stylistic feature sets used non-vocabulary features to represent the subjectivity clues in the code-switching text. Fig. 1 illustrates the proposed feature sets.

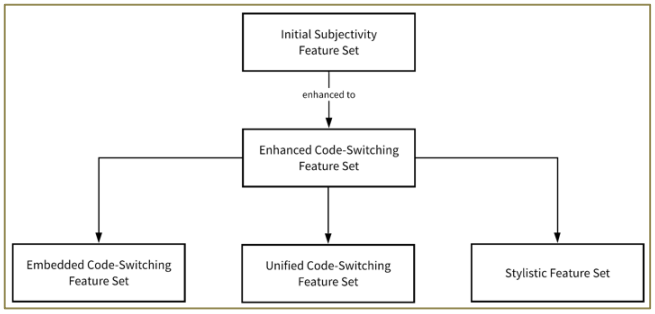

Fig. 1. Proposed feature set for code-switching subjectivity analysis.

A. Embedded Code-Switching Feature Set

The subjectivity analysis feature set for a monolingual text is derived from a monolingual part-of-speech (POS) tag. A code-switching sentence is constructed using at least two different languages. Therefore, two POS taggers are used to produce the feature set for the code-switching sentence. The union of the POS tags from two languages produces the embedded code-switching feature set. The union process is shown in Fig. 2.

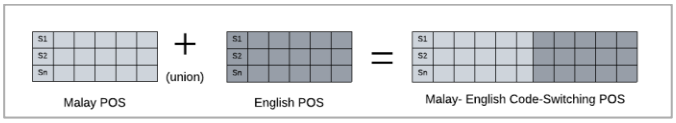

Fig. 2. The union process to construct embedded code-switching POS tags.

The Hawkin-UKM Malay and Penn Treebank POS tags are used to produce the embedded code-switching feature sets. The POS taggers consist of several tags to represent a single POS. For example, the Hawkin Malay POS tagger used ADJ, ADJS and ADJT to represent an adjective word, while the Penn Treebank POS used JJ, JJR and JJS. These POS tags are combined and grouped. Table II shows the grouping of the POS tags that represent the subjectivity features for the Malay-English code-switching text.

TABLE II. MALAY-ENGLISH CODE-SWITCHING POS TAGS

Part of Speech	Malay POS Label	English POS Label
Adjective	ADJ, ADJS, ADJT	JJ, JJR, JJS
Adverb	ADV	RB, RBR, RBS, WRB
Noun	KN, KNG, KNK, KNT, KPB	NN, NNP, NNPS, NNS
Verb	KK, KKIA, KKIK, KKIW, KKT	VB, VBD, VBG, VBN, VBP, VBZ
Pronoun	KGDD, KGDP, KGDT, KGNT, @KG	PRP, PRP$, WP
Conjunction	KH	CC, IN
Cardinal number	KBIL	CD
Modal	MD	MD

Modal POS was absent in the initial Hawkin-UKM Malay POS tagger. Modal is included in the Malay POS for a balanced and uniform feature. The modal POS tag is defined as MD. A list of English modals retrieved from MyEnglishPages.com[1] is translated into Malay using Cambridge Online English-Malay Dictionary[2], and Kamus Dwibahasa Bahasa Inggeris – Bahasa Melayu [13], [14]. Thirteen English modals were translated into Malay. The translation process produces 30 Malay words with equal meaning to the English modal.

The process to generate the Malay-English code-switching feature set is shown in Fig. 3. The process consists of five phases. The process begins with the sentence tokenization process in Phase 1. The code-switching sentences retrieved from the repository are tokenized. Tokenization is a process that breaks the words into the sentence into individual pieces. The individual piece is called a token. The space between the words is used as a boundary to mark the beginning and end of a word. This is known as a delimiter.

The process continues with POS tagging in Phases 2 and 3. The tokens are tagged using the Hawkin Malay and Penn Treebank POS taggers. Phases 2 and 3 produced two sets of tagged tokens. In Phase 4, the English POS-tagged tokens are embedded into the Malay POS-tagged tokens. The phase standardized the POS tags using the group described in Table III. In Phase 5, the presence of the feature is coded as 1, while others are coded as zero, indicating the absence of the feature. The features are extracted and converted into a matrix. This matrix is known as an embedded code-switching feature set.

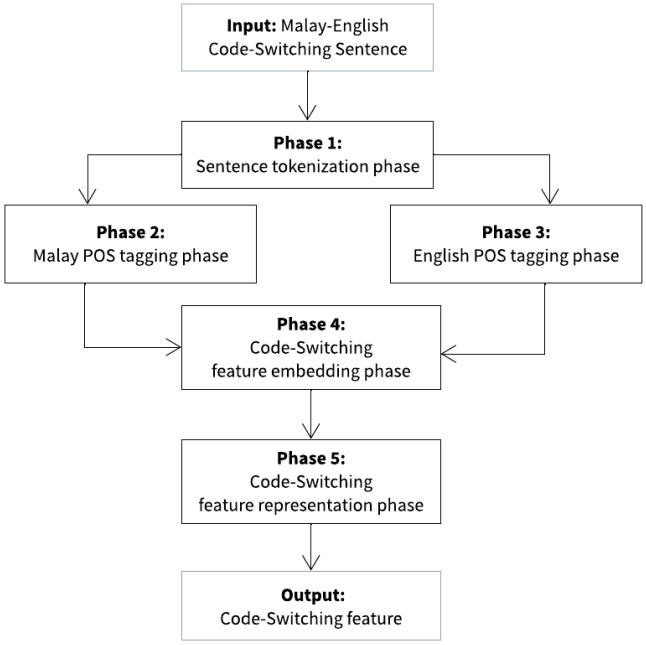

Fig. 3. The process to generate embedded code-switching feature set.

TABLE III. EMBEDDED CODE-SWITCHING SUBJECTIVITY FEATURE SET

Feature	Description
ADJ_MS	Presence of adjective for Malay words
ADV_MS	Presence of adverb for Malay words
NN_MS	Presence of noun for Malay words
VB_MS	Presence of verb for Malay words
PR_MS	Presence of pronoun for Malay words
CC_MS	Presence of conjunction for Malay words
CD_MS	Presence of cardinal number for Malay words
MD_MS	Presence of modal for Malay words

B. Unified Malay-English POS Feature Set

The second feature set is the unified Malay-English POS feature set. The process to produce the unified code-switching feature set is shown in Fig. 4. The process consists of six phases. The process starts with Phase 1, which tokenized the sentences. The tokens are POS-tagged using the Hawkin-Malay and Penn Treebank POS tags in Phase 2 and Phase 3. The phases produce two sets of tokens. However, there were discrepancies in token lengths identified in Phase 4. The length of the token is standardized by removing the additional token.

In Phase 5, the Malay and English POS Tags are unified using a rule-based algorithm. The algorithm was adopted from [15], as shown in Algorithm 1. The algorithm requires three (3) sentences: the sentence with language label tags, the sentence with English POS tags and the sentence with Malay POS tags. Each word in the sentence with language label tags will be mapped to either a sentence with English POS tags or a sentence with Malay POS tags. The mapping of each word will be according to the language label.

[1] https://www.myenglishpages.com/site_php_files/grammar-lesson-modals.php

[2] https://dictionary.cambridge.org/dictionary/english-malaysian/

Algorithm 1: Unified_cs_pos_tag

Input:

sentence_cs: Malay-English code-switching sentence with language tags

sentence_en_pos: Malay-English code-switching sentence with English POS tags

sentence_my_pos: Malay-English code-switching sentence with Malay POS tags

sentence_unified_POS = " "

For each word in sentence_cs

Get the language of the word

If the language labelled for the word is either as English or Shared

Get the matching word from sentence_en_pos

Concatenate word and POS Tag into sentence_unified_POS

Else if the word is labelled as Malay

Get the matching word from sentence_my_pos

Concatenate word and POS Tag into sentence_unified_POS

Else

Discard the word

End

End

Return sentence_unified_POS

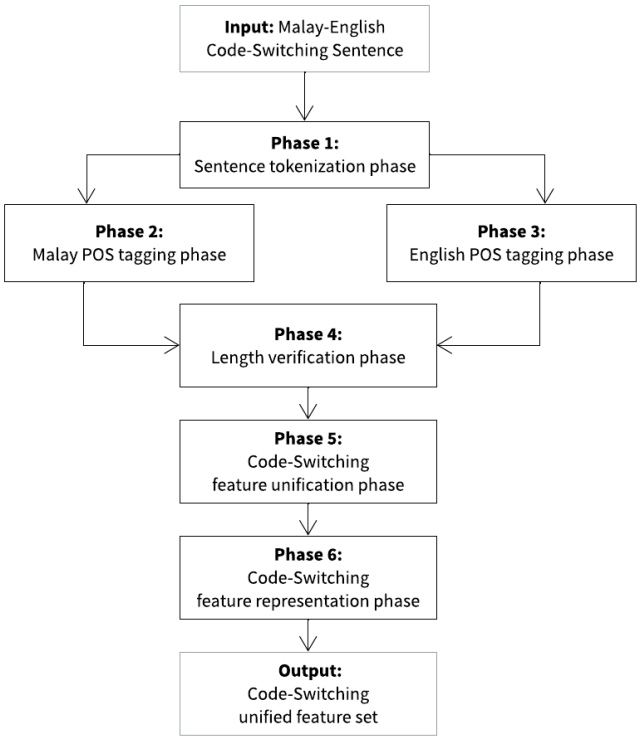

Fig. 4. The process to generate a unified POS feature set.

C. Stylistic Feature

The stylistic features are non-vocabulary words representing the author's emotional state in a text. The stylistic features that influence the code-switching text are exclamation marks, emoticons, exaggeration, intensifiers and interjections. An emoticon is a representation of the facial expression, mood and emotion of an author that is created by combining numerous keyboard strokes. An emoticon strengthens the message conveyed in the text [16]. It is observed that the widespread usage of emoticons in electronic documents has indicated the development of emotion [17]. An emoticon is also used to clarify the intent of the electronic message, such as sarcasm and criticism [16]. The findings of these studies have shown that emoticons are used in a subjective textual document. Table IV shows examples of emoticons and their underlying meaning.

Creative spelling with exaggeration of characters is another stylistic feature considered in the code-switching text. The exaggeration of characters is defined as the repetition of characters in a word that occurs more than two times. The repetition of characters in the text indicates the expression of information intensifying.

TABLE IV. REPRESENTATION OF EMOTICON AND ITS UNDERLYING MEANING

Emoticon	Meaning
:) or :-) or (^_^)	An emoticon that represents a smiling face that indicates the author is happy or pleased.
:(or :-(	An emoticon that represents a frown face that indicated the author is sad or unhappy.
;) or ;-)	An emoticon that represents a winking, indicating being flirtatious.

An interjection is used as a form of human expression of feeling in an electronic document. Interjection represents an immediate feeling by not literally describing it [18]. Some interjections are language-independent, such as "wow", "haha", and "hmm". Whilst some interjections are language-dependent, such as "jeng", "gap", and "seh", which are commonly used by Malay speakers, were found in the code-switching sentences. Table V shows some examples of interjections and their meaning.

TABLE V. EXAMPLE OF INTERJECTIONS AND ITS MEANING

Interjection	Meaning
wow	Represents astonishment
haha	Represent regular laughter
hmm	Represent thinking or hesitation
jeng	Represent surprise. Usually appear multiple times such as jeng…jeng..jeng
ngap	Represent act of eating

IV. BUILDING MALAY-ENGLISH SUBJECTIVE DATA SET

The data used in this research was harvested from 45,964 Malay-English blog posts. The blog is chosen as the source of data for numerous reasons: 1) The content is rich with objective and subjective information 2) The content was written using a mixture of Malay and English 3) The blog is still relevant 4) The blog is publicly available 5) The casual writing style.

Preparing the data set begins with extracting the content of the blog post. For that purpose, a Python program using the BeautifulSoup module was developed. The blog content was separated into individual sentences, and the distribution of Malay and English words was computed using the procedure shown in Algorithm 2.

Algorithm 2: Compute_MS_EN_distribution

For every sentence in the database , sn

$t_{malay} = 0$, $t_{english} = 0$, $t_{shared} = 0$, $t_{ood} = 0$

Separate the sentences into individual word, *s={w1, w2, w3, ... ,wn}*

For every word in the sentence, wn

If ((wn ∈ mswn) ∩ (wn ∉ enwn))

t_{malay} += 1

else if ((wn ∉ mswn) ∩ (wn ∈ enwn))

$t_{english}$ += 1

else if ((wn ∈ mswn) ∩ (wn ∈ enwn))

t_{shared} += 1

else

t_{ood} += 1

End

Record the language distribution for this sentence, sdn = { t_{malay}, $t_{english}$, t_{shared}, t_{ood}}

End

End

Return record of language distribution

After that, the Malay-English code-switching sentences were extracted using the procedure shown in Algorithm 3. A Malay-English code-switching sentence should contain at least one Malay and one English functional word. This research defined a functional word as a word that is not categorized as either a stop word or the name of an entity.

Algorithm 3: selecting_MS_EN_sentences

Get language distribution for every sentence from the database

For each language distribution of a sentence

Get Malay words distribution

Get English words distribution

If ((Total Malay words distribution >= 1) and (Total English words distribution >= 1))

Label the sentence as MS-EN-CS

Else if ((Malay words distribution >= 1) and (English words distribution == 0))

Label the sentence as MS

Else if ((Malay words distribution == 0) and (English words distribution >= 1))

Label the sentence as EN

End

End

The procedure has extracted 93,796 Malay-English sentences. Sentences containing 3 to 25 words are selected to be labelled as subjective or objective. Mohammad (2016) and Belisário et al. (2020) used a similar number of words per sentence [19], [20]. Sentences with less than three words were considered uninformative. Sentences containing more than 25 words contain overwhelming information that increases the complication of the annotation task. 56,207 sentences were selected after discarding the unwanted sentences. These sentences were labelled by two annotators using the annotation scheme described in Table VI.

The annotation process has produced 35,067 Malay-English code-switching sentences - 25,164 were subjective, while 9,903 were objective sentences. These sentences will be used as a dataset in the experiment section. The result is shown in Table VII.

TABLE VI. ANNOTATION SCHEME FOR MALAY-ENGLISH CODE-SWITCHING CORPUS

Label	Language			Description of sentence	
	Malay	***English***	***Malay-English***	***Fact***	***Opinion***
EN-OPI		√			√
EN-FAC		√		√	
MS-OPI	√				√
MS-FAC	√			√	
CS-EN-OPI			√		√
CS-MS-OPI			√		√
CS-FAC			√	√	

TABLE VII. DISTRIBUTION OF ANNOTATED SENTENCES

Language / Labels	Subjective	Objective
Malay-English	25, 164	9,903
Code-Switching	25, 164	9,903
English	3, 957	1, 197

The dataset contains 9,903 objective sentences and 25,164 subjective sentences. The number of subjective sentences is two and a half times greater than the number of objective sentences, making the data imbalanced. However, a machine learning algorithm works best with a balanced dataset.

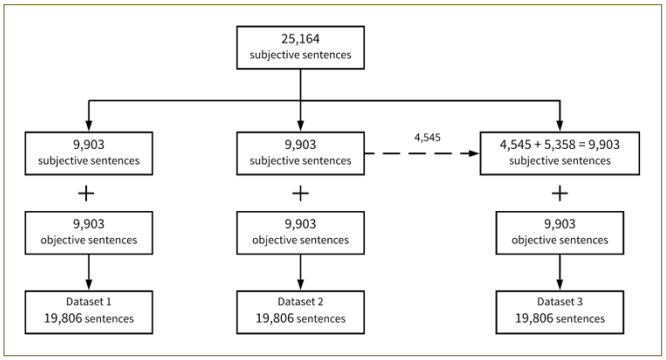

Fig. 5. Distribution of dataset.

This research divided the subjective sentences into three parts to create an equal number of sentences to objective sentences, as shown in Fig. 5. Krawczyk (2016) used a similar approach [21]. This research named these parts as Dataset 1, Dataset 2 and Dataset 3. With this distribution, all datasets have an equal number of subjective and objective sentences.

V. RESULT

Nine experiments were designed and carried out to evaluate the proposed feature sets' ability to analyse the subjectivity in Malay-English code-switching text. The experiments were carried out using two different machine learning classifiers, Naïve-Bayes and Support Vector Machine (SVM). Two experiments were carried out exclusively for English and Malay sentences using the initial feature sets as baseline experiments. The baseline experiments are established as comparisons for the proposed feature sets. Other experiments were conducted to identify the optimal feature sets and machine learning classifiers that function well with subjectivity classification for Malay-English code-switching text at the sentence level.

A. Baseline Feature Sets Performance Results

The results from the baseline experiments will be compared to those of the other experiments. The baseline experiments are performed separately, considering only English and Malay initial feature sets. These feature sets are trained and tested using two different classifiers and three different datasets. The accuracy performance of the experiments is shown in Fig. 6. The results were obtained after performing 10-fold cross-validation for both classifiers.

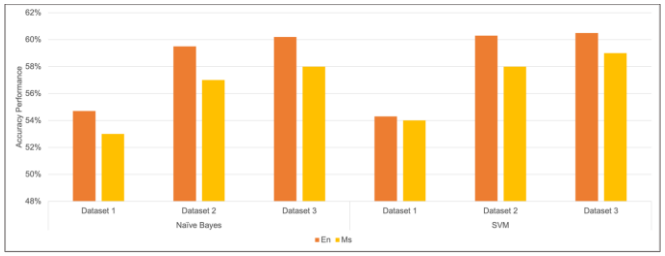

Fig. 6. Accuracy performances of baseline initial features models using different classifiers across multiple datasets.

Fig. 6 shows the baseline feature set performed at 55.00% accuracy for the English (En) feature set and 53.00% accuracy for the Malay (Ms) feature using Dataset 1 using the Naïve-Bayes classifier. The performance of the baseline feature sets using Dataset 2 is 59.00% accuracy for the English (En) feature set and 57.00% accuracy for the Malay feature set, using the same classifier. The baseline feature sets yielded 60.00% accuracy and 58.00% accuracy, respectively, for the English initial feature set and Malay initial feature set, using Dataset 3 and the same classifier. Dataset 3 has the highest accuracy performance among the three datasets using the Naive-Bayes classifier.

The accuracy of the baseline initial feature sets was lower in Dataset 1 using the SVM classifier, which is 54.00% for both, as shown in Fig. 9. The accuracy of the models increased significantly using Dataset 2 using the same classifier, where the English initial feature set achieved 60.00% accuracy, and the Malay initial feature set achieved 58.00% accuracy. The accuracy performance of the English initial feature set showed a slight improvement using Dataset 3. However, the accuracy performance for the Malay feature set increased by 1.00% to 59.00% using Dataset 3 and the SVM classifier.

In general, the bar chart in Fig. 9 showed increments of accuracy performance from Dataset 1 to Dataset 3 in both classifiers. Comparatively, the accuracy performances of the initial feature sets between the datasets and classifiers show a consistent increment pattern. Therefore, the baseline results can be compared with the proposed feature set models.

B. Embedded Code-Switching Feature Set Performance Results

The same setting of experiments that were carried out for the baseline experiment is used to evaluate the embedded code-switching feature set. The result of the experiment is shown in Fig. 7. The accuracy results show the embedded feature sets perform at equal accuracy, 54.00% for Dataset 1, using both classifiers, Naive-Bayes and SVM classifier. The result also shows that Dataset 3 outperforms other datasets for both classifiers.

There is an increment pattern of accuracy performance between the three datasets for both classifiers. There is a significant accuracy increment between Dataset 1 and Dataset 2 using both classifiers. Datasets 1 and 2 performed at 54.00% and 57.00% using the embedded feature set and Naive Bayes classifier. The accuracy increased by 3.00%. The accuracy performance increased by 5.00% using the SVM classifier between Datasets 1 and 2. Datasets 1 and 2 performed at 54.00% and 59.00%, respectively, with the SVM classifier. The accuracy increased by 1.00% between Dataset 2 and 3 using both classifiers. It is also noted that there are noteworthy accuracy differences between Dataset 1 and 3 for both classifiers. The differences in accuracy performance between Datasets 1 and 3 are 4.00% differences using the Naive Bayes classifier and 5.00% using the SVM classifier.

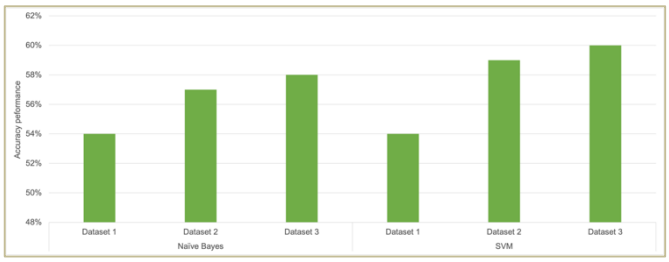

Fig. 7. Result of accuracy performance for subjectivity classification on Malay-English code-switching text using embedded feature set.

The significant differences in accuracy performance between the datasets are due to the presence of the subjectivity features in the dataset. The presence of the Malay and English subjectivity features is more significant in Datasets 2 and 3 compared to Dataset 1. The differences reveal the SVM classifier works better to identify the subjectivity presence in Malay-English code-switching sentences using the embedded feature-set, with 60.00% accuracy, where Datasets 3 outperformed others. Therefore, using both classifiers, the embedded code-switching feature set can be used to distinguish

subjective and objective Malay-English code-switching sentences.

The accuracy performance results of the embedded feature set were compared with the initial feature sets, which are the English and Malay initial feature sets. The comparison is shown in Fig. 8. The accuracy performance of the English initial feature set (En) was superior in all datasets and classifiers. The superiority of accuracy performances indicates the English initial feature sets are stable and robust. However, the accuracy performance of the embedded feature set is better in comparison to the Malay initial feature set (Ms) in three experiment settings, which are Dataset 1 with Naive Bayes classifier and Datasets 2 and 3 using the SVM classifier. The accuracy performance of the feature set is on par with the other three experiment settings. This is clearly shown using Datasets 2 and 3 with the Naive-Bayes classifier and Dataset 1 using the SVM classifier.

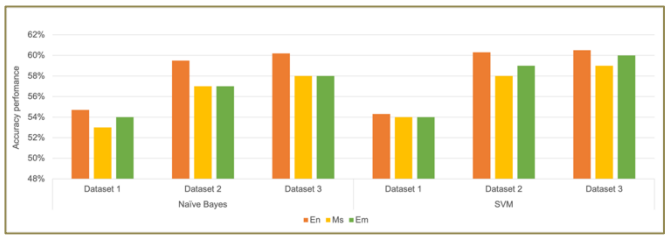

Fig. 8. Comparison of embedded feature sets with baseline initials feature sets.

The accuracy performance of the embedded feature set (Em) is either superior or on par with the Malay initial feature set (Ms). It is concluded that the Em feature set performed as good as the Ms feature set. The Em feature set had given a competitive advantage to the Ms initial feature set, given the appearance of multiple languages in a single sentence. It is also obvious that the Em feature set improved the subjectivity classification performance on code-switching text instead of using the Malay initial feature alone for the classification.

C. *Unified Code-Switching POS Feature Set Performance Results*

The unified code-switching POS feature set fused Malay and English using the algorithm described in Algorithm 1. The unified process determined the POS of a word based on its language. The unified feature set was trained and tested using a similar experimental setting as the embedded code-switching feature sets and the initial feature sets. The accuracy performances from the classification of this feature set are captured and shown in Fig. 9.

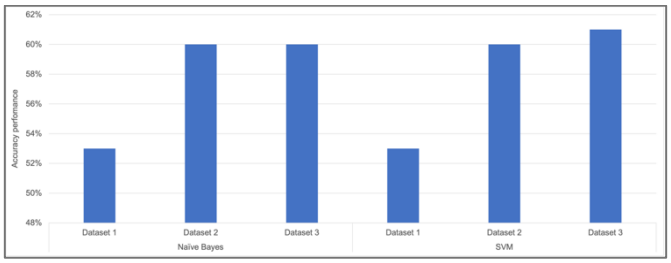

Fig. 9. Results of accuracy performance for the subjectivity of classification on Malay-English code-switching text using the unified feature set.

The bar chart in Fig. 9 shows the unified feature set performed at 52.00% accuracy using Dataset 1 and both classifiers. The accuracy performance of the unified feature set is on the same level at 60.00% for Datasets 2 and 3 using the Naive Bayes classifier and Data Set 2 using the SVM classifier. There is a slight increment in Dataset 3 using the SVM classifier, in which the feature set performed at 61.00% accuracy. The consistent performance of around 60.00 to 61.00% shows the feature set is usable to differentiate the subjective and objective sentences from a Malay-English code-switching dataset.

The bar chart in Fig. 10 compiled the accuracy results from the baseline experiments (En and Ms) and unified feature set (Un) using the Naïve-Bayes and SVM classifiers. In general, the unified feature set outperformed the English initial feature set using Dataset 2 with the Naive Bayes classifier at 60.00% and Dataset 3 with the SVM classifier at 61.00%. The unified feature set performed on par accuracy in comparison to the English initial feature set using Datasets 2 and 3 with the Naive-Bayes and the SVM classifier. The performances are 60.00%. The outperformance and on-par accuracy result shows the unified feature set is as good as the English initial feature set, given the fact that code-switching is a relatively new challenge to text processing and subjective analysis. However, the unified feature set did not share the same excellence of accuracy performance result using Dataset 1 for both classifiers. The unified feature set performed at 53.00% accuracy using Naive Bayes and SVM classifiers on the same dataset, less 2.00% and 1.00% for each classifier. Even though the same level of excellence is not shared by using Dataset 1 and the same classifiers, the unified feature set can still distinguish the subjective from the objective sentences of Malay-English code-switching text.

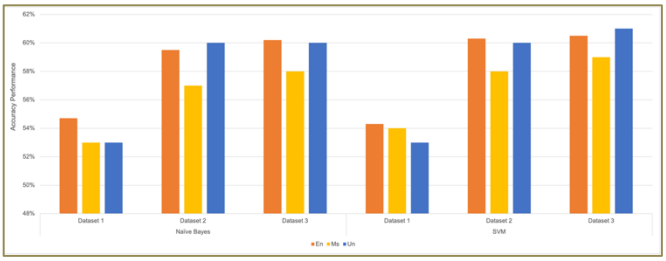

Fig. 10. Comparison of accuracy performance for the unified feature set with baseline initial feature sets.

The unified feature set outperformed the Malay initial feature set in four experiments. The accuracy performance of the unified feature set is 60.00% using Dataset 2 with Naive Bayes classifier, which is 3.00% higher than the Malay initial feature set. The same accuracy performance is seen as consistent for the feature set using Dataset 3 and the same classifier. However, it is 2.00% higher than the Malay initial feature set. The same pattern of differences is also apparent for Datasets 2 and 3 using the SVM classifier compared to the Malay initial feature set. The unified feature set performed at the same level of accuracy performance, 55.00%, using Dataset 1 with the Naive Bayes classifier. However, the feature set did not share the same accuracy performance as the SVM classifier using the same data set. The feature set performed at 53.00%, less than 1.00% from the Malay initial feature set. The significant differences in the accuracy performance between the unified Malay-English code-switching and the Malay initial feature sets across datasets and

multiple classifiers show the unification of feature sets for two different languages is necessary to distinguish the code-switching sentences into objective and subjective classes.

D. Stylistic Feature Sets Performance Results

Stylistic features significantly influence subjectivity analysis [22][23]. Six experiments were conducted to investigate the ability of stylistic features to distinguish subjective and objective sentences from Malay-English code-switching datasets. The experiments were conducted using the same settings as the embedded and unified feature sets. The results are shown in Fig. 11.

The bar chart in Fig. 11 shows a consistent accuracy performance achieved by the stylistic feature set using Datasets 3 with the Naive Bayes classifier and Datasets 1, 2 and 3 with the SVM classifier. The stylistic feature set performed at 55% accuracy. It is worth mentioning that the differences in the stylistic performance from these datasets and Datasets 1 and 2 using the Naive Bayes classifier are 2.00% and 1.00%, respectively. The insignificant accuracy performance differences across datasets and classifiers show that the stylistic feature set performance is nearly consistent. Therefore, these findings support the influence of stylistic feature sets in code-switching text.

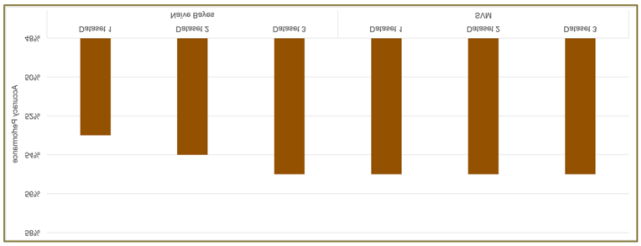

Fig. 11. Results of accuracy performance for the subjectivity of classification on Malay-English code-switching text using the unified feature set.

The accuracy performance of the stylistic feature set (St) is compared to the initial feature sets for English (En) and Malay (Ms). The accuracy performances are shown in Fig. 12. The bar chart in Fig. 11 shows the English and Malay initial feature sets have outperformed the stylistic feature sets using Datasets 2 and 3 for both classifiers. The stylistic feature set gives the same performance as the Malay initial feature set using Dataset 1 and the Naive Bayes classifier, but the English initial feature set still gives superior performance. However, the stylistic feature set can surpass the English and Malay initial feature sets using Dataset 1 and the SVM classifier.

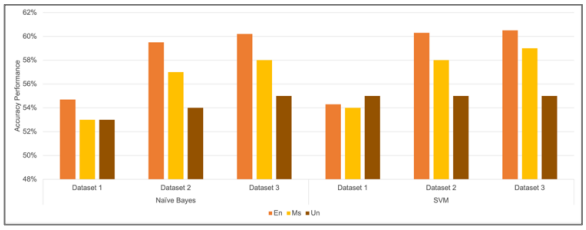

Fig. 12. Comparison of accuracy performance for a stylistic model with baseline initial feature sets.

E. Combinations of Feature Sets Performance Results

The experiments in the previous section were executed independently of each other. The proposed feature sets were then combined to determine the possibility of improved accuracy performances. Four more experiments were conducted for this purpose. The combinations are listed as follows:

1) The combination of an English (En) initial feature set and a stylistic (St) feature set is known as En + St.

2) The combination of the Malay (Ms) initial feature set and stylistic (St) feature set is known as Ms + St.

3) The combination of embedded (Em) feature set and stylistic (St) feature set is called Em + St.

4) The combination of a unified (Un) feature set and a stylistic (St) feature set is known as Un + St.

The experiments were carried out using a similar experimental setup as the previous experiments – using three datasets and two classifiers, the Naive Bayes and the SVM. The accuracy performances from each dataset were captured and averaged. The averaged accuracy performances are illustrated using bar charts in Fig. 13 and Fig. 14.

The average accuracy performances for the combination of proposed features using the Naïve Bayes classifier are shown using a bar chart in Fig. 13. The bar chart includes the averaged accuracy performance result from Malay and English feature sets. The result shows the combination of the unified (Un) and stylistic (St) feature sets outperformed other feature sets at 59.00% accuracy. The combination of embedded and stylistic feature sets (Em + St) shows an on-par performance with the English initial feature set. The results show that considering features from languages other than English in code-switching text and stylistic features improves the accuracy performance. Thus, the combination of feature sets is practical to classify the Malay-English code-switching sentences into subjective and objective classes. The bar chart also shows that the stylistic feature set alone performed at the lowest accuracy, 54.00%, in comparison to other feature sets. The performance reveals using only stylistic features to classify the subjectivity of code-switching sentences is insufficient.

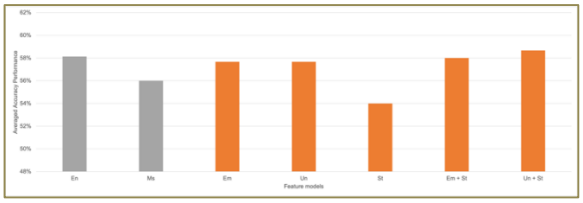

Fig. 13. Comparison of averaged accuracy performances for combined feature sets using the Naïve-Bayes classifier.

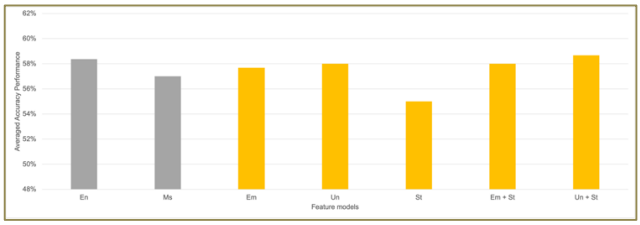

Fig. 14. Comparison of averaged accuracy performances for combined feature sets using the SVM classifier.

The results of averaged initial feature sets (Malay and English) proposed feature sets (embedded, unified and stylistic) combination feature sets that were classified using the SVM classifier are shown in a bar chart of Fig. 14. The bar chart shows the averaged accuracy performance using the combination of unified and stylistic (Un + St) feature set surpassed other feature sets. The Un + St feature set performed at 59.00% accuracy. The combination of embedded and stylistic (Em + St) feature sets performed at 58.00% accuracy, slightly lower than the English (En) initial feature set and on par as a unified feature set. The bar chart shows a slight accuracy performance improvement when the embedded and unified feature sets are combined with the stylistic feature set. The performance of both feature sets increased by 1.00% with the combination of stylistic feature sets. The bar chart also shows the stylistic feature set alone gave the lowest performance, that is, 55.00% accuracy. Therefore, a stylistic feature set should not be used as the only feature set to distinguish subjective and objective sentences for Malay-English code-switching datasets. In addition to that, stylistic feature sets have proven their influence on subjectivity classification for Malay-English code-switching text.

VI. Discussion

The results of this study provide a comprehensive evaluation of the subjectivity classification in Malay-English code-switching text using enhanced feature sets and machine learning classifiers. Several key insights emerged from the result analysis.

A. Performance of Baseline Feature Sets

The baseline feature sets focused exclusively on English and Malay feature sets. It provides a starting point to evaluate the effectiveness of the enhanced feature sets. The English feature set has outperformed the Malay feature sets across all datasets. This finding aligns with previous research that suggests English-based models tend to perform better in text classification tasks due to the availability of more extensive language resources and tools.

B. Embedded Code-Switching Feature Set

A key contribution of this study is how the embedded code-switching feature sets successfully captured subjectivity in code-switched text. These feature sets performed better than the baseline Malay sets, especially when using Dataset 3 with both Naive Bayes and SVM classifiers. This shows how important it is to include both languages in the feature sets to handle code-switching effectively. Additionally, the improvement in accuracy from Dataset 1 to Dataset 3 suggests that having a more diverse dataset could help models better recognise subjectivity in code-switched sentences.

C. Unified Code-Switching POS Feature Set

The unified feature set, which combined Malay and English parts of speech (POS), delivered competitive results, particularly with Dataset 3 when using the SVM classifier. This approach outperformed the baseline English feature set, highlighting the importance of considering the syntactic structures of both languages in a unified way for tasks involving code-switching. Interestingly, the unified feature set exhibited limitations in Dataset 1, indicating that while this method is advantageous, its performance depends on the characteristics of the dataset and the quality of POS tagging.

D. Impact of Stylistic Feature Set

Stylistic features, which include linguistic cues such as punctuation, sentence structure, and word choice, were also explored for their role in subjectivity classification. While the stylistic feature sets performed consistently across all datasets, they were outperformed by the initial English and Malay feature sets. This suggests that stylistic features alone may not be sufficient for classifying subjectivity in code-switched text. However, when combined with other feature sets, such as the unified and embedded feature sets, stylistic features contributed to incremental improvements in accuracy.

E. Combination of Feature Sets

The experiments that combined multiple feature sets revealed that unifying linguistic and stylistic features yield the best performance. Specifically, the combination of the unified and stylistic feature sets resulted in the highest accuracy across datasets, particularly with Dataset 3. This finding suggests that a hybrid approach—leveraging multiple linguistic levels (e.g., syntax, style, language features)—is most effective for handling the complexity of Malay-English code-switching text. These findings reinforce the idea that no single feature set can comprehensively capture the nuances of subjectivity in code-switching. Instead, a combination of feature sets is required to achieve optimal performance.

VII. Limitation and Future Research

There are several limitations to this study. First, the datasets used in the experiments, while diverse, may not fully capture the complexity of real-world code-switching scenarios. Future research could explore larger and more diverse datasets that include different levels of code-switching. Additionally, this study focused on sentence-level classification; however, subjectivity may vary within a single sentence. Thus, future work could explore word-level or phrase-level classification to gain a more fine-grained understanding of subjectivity in code-switching.

Moreover, the study was limited to Malay-English code-switching. As such, future research could apply the proposed methods to other language pairs where code-switching is common, such as Spanish-English or Arabic-French. Investigating whether the same feature sets and classifiers work across different language combinations could provide more generalizable insights into the phenomenon of code-switching in subjectivity classification.

VIII. Conclusion

The results have several implications for future research and practical applications. First, the importance of language-specific features in code-switched text is evident, highlighting the need for models that can process multiple languages simultaneously. This is particularly relevant in multilingual societies like Malaysia, where code-switching is prevalent in everyday communication. The findings also suggest that existing models and datasets can be improved by focusing on the interaction between syntactic and stylistic features in both languages.

Three feature sets were designed to distinguish the code-switching text into subjective and objective classes. The first feature set includes the Malay feature into the English feature – known as the embedded feature set. The second feature set combined and unified the Malay and English feature set, which is known as a unified feature set. The last feature set is non-vocabulary, known as a stylistic feature set. Part-of-speech is used as a feature of the code-switching text. The 1 and 0 are used to represent the presence and absence of the features. The feature sets are more interpretable where the contribution of a specific can be directly seen to the model decisions.

The experiments were carried out using three datasets and two machine learning classifiers – the Naive Bayes and SVM to verify the proposed feature sets. The results from the proposed feature sets were compared with the initial feature sets. The experiment shows the unified feature performed as good as the English initial feature set. Therefore, unifying the Malay and English feature sets is necessary to distinguish the subjective and objective sentences in the Malay-English code-switching dataset. An improvement in accuracy performance is achieved when the unified feature set is combined with the stylistic feature set. The improvement reveals this method is computationally lighter which requires fewer data to perform well.

ACKNOWLEDGMENT

The authors would like to thank Social Analytics and Intelligence Lab (SAIL) and the Applied Intelligent Computing (APIC) research group, the Center of Advanced Computing Technology (C-ACT), and Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for their incredible support for this research.

REFERENCES

[1] P. Muysken, Bilingual Speech: A Typology of Code-Mixing. Cambridge University Press, 2000.

[2] T. Huang, Z. H. Deng, G. Shen, and X. Chen, 'A Window-Based Self-Attention approach for sentence encoding', Neurocomputing, vol. 375, pp. 25–31, Jan. 2020, doi: 10.1016/j.neucom.2019.09.024.

[3] L. B. Belisário, L. G. Ferreira, and T. A. S. Pardo, 'Evaluating Richer Features and Varied Machine Learning Models for Subjectivity Classification of Book Review Sentences in Portuguese', Information (Switzerland), vol. 11, no. 9, Sep. 2020, doi: 10.3390/INFO11090437.

[4] A. Al Hamoud, A. Hoenig, and K. Roy, 'Sentence Subjectivity Analysis of Political Debate and Ideological Debate Dataset using LSTM and BiLSTM with Attention and GRU Models', Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, pp. 7974–7987, Nov. 2022, doi: 10.1016/j.jksuci.2022.07.014.

[5] A. K. Joshi, 'Processing of Sentences with Intra-Sentential Code-Switching', in Proceedings of the 9th Conference on Computational Linguistic, 1982, pp. 145–150. doi: 10.1017/cbo9780511597855.006.

[6] J. Chi, B. Lu, J. Eisner, P. Bell, P. Jyothi, and A. M. Ali, 'Unsupervised Code-switched Text Generation from Parallel Text', in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, International Speech Communication Association, 2023, pp. 1419–1423. doi: 10.21437/Interspeech.2023-1050.

[7] I. Hamed, N. Habash, and N. T. Vu, 'Data Augmentation Techniques for Machine Translation of Code-Switched Texts: A Comparative Study'. [Online]. Available: http://arzen.camel-lab.com/

[8] K. Hu et al., 'Improving Multilingual and Code-Switching ASR Using Large Language Model Generated Text', in 2023 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ASRU57964.2023.10389644.

[9] D. Gupta, A. Lamba, A. Ekbal, and P. Bhattacharyya, 'Opinion Mining in a Code-Mixed Environment: A Case Study with Government Portals', Proc. of the 13th Intl. Conference on Natural Language Processing, pp. 249–258, 2016, [Online]. Available: http://ltrc.iiit.ac.in/icon2016/proceedings/icon2016/pdf/W16-6331.pdf

[10] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, 'Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora', no. 2011, pp. 2–8, 2015, doi: 10.18653/v1/w15-2902.

[11] A. Jamatia, S. D. Swamy, B. Gambäck, A. Das, and S. Debbarma, 'Deep Learning Based Sentiment Analysis in a Code-Mixed English-Hindi and English-Bengali Social Media Corpus', International Journal on Artificial Intelligence Tools, vol. 29, no. 05, pp. 20–35, Jun. 2020, doi: 10.1142/S0218213020500141.

[12] V. Hatzivassiloglou and J. M. Wiebe, 'Effects of Adjective Orientation and Gradability on Sentence Subjectivity', in COLING '00: Proceedings of the 18th conference on Computational linguistics, 2000, pp. 299–305. doi: 10.3115/990820.990864.

[13] S. Ibrahim, Kamus Dwibahasa Bahasa Inggeris-Bahasa Melayu, Edisi Kedu. Dewan Bahasa dan Pustaka, 2019.

[14] J. M. Hawkins, Kamus Dwibahasa Oxford Fajar Inggeris-Melayu, Melayu-Inggeris. Fajar Bakti, 2006.

[15] T. Solorio and Y. Liu, 'Learning to Predict Code-Switching Points', in EMNLP 2008 - 2008 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference: A Meeting of SIGDAT, a Special Interest Group of the ACL, 2008, pp. 973–981. doi: 10.3115/1613715.1613841.

[16] M. A. Ullah, S. M. Marium, S. A. Begum, and N. S. Dipa, 'An algorithm and method for sentiment analysis using the text and emoticon', ICT Express, vol. 6, no. 4, pp. 357–360, 2020, doi: https://doi.org/10.1016/j.icte.2020.07.003.

[17] B. Jung, H. Kim, and S. H. (Shawn) Lee, 'The impact of belongingness and graphic-based emoticon usage motives on emoticon purchase intentions for MIM: an analysis of Korean KakaoTalk users', Online Information Review, vol. 46, no. 2, pp. 391–411, Jan. 2022, doi: 10.1108/OIR-02-2020-0036.

[18] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, 'Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations', IEEE/ACM Trans Audio Speech Lang Process, vol. 29, pp. 1675–1686, 2021, doi: 10.1109/TASLP.2021.3076364.

[19] L. B. Belisário, L. G. Ferreira, and T. A. S. Pardo, 'Evaluating Richer Features and Varied Machine Learning Models for Subjectivity Classification of Book Review Sentences in Portuguese', Information, vol. 11, no. 9, pp. 1–14, 2020, doi: 10.3390/INFO11090437.

[20] S. M. Mohammad, 'Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text', Emotion Measurement, pp. 201–237, 2016, doi: 10.1016/B978-0-08-100508-8.00009-6.

[21] B. Krawczyk, 'Learning from imbalanced data: open challenges and future directions', 2016. doi: 10.1007/s13748-016-0094-0.

[22] F. Bravo-Marquez, M. Mendoza, and B. Poblete, 'Meta-Level Sentiment Models for Big Social Data Analysis', Knowl Based Syst, vol. 69, no. 1, pp. 86–99, 2014, doi: 10.1016/j.knosys.2014.05.016.

[23] A. S. Altheneyan and M. E. B. Menai, 'Naïve Bayes Classifiers for Authorship Attribution of Arabic Texts', Journal of King Saud University - Computer and Information Sciences, vol. 26, no. 4, pp. 473–484, 2014, doi: 10.1016/j.jksuci.2014.06.006.